# MultiModal-GPT：与人对话的视觉和语言模型

**Tao Gong**[1*] **Chengqi Lyu**[1*] **Shilong Zhang**[2,1*] **Yudong Wang**[1,3*] **Miao Zheng**[1*]

**Qian Zhao**[1*] **Kuikun Liu**[1*] **Wenwei Zhang**[1*] **Ping Luo**[2,1] **Kai Chen**[1✉]

*equal contribution, in random order

[1]Shanghai AI Laboratory [2]The University of Hong Kong

[3]School of Electrical and Information Engineering, Tianjin University

{gongtao, lvchengqi, zhangshilong, wangyudong, zhengmiao}@pjlab.org.cn

{zhaoqian, liukuikun, zhangwenwei, chenkai}@pjlab.org.cn

## Abstract

我们提出了一种名为 MultiModal-GPT 的视觉和语言模型, 用于与人类进行多轮对话的任务。MultiModal-GPT 能够遵循各种指令来完成一些任务，例如生成详细的图像字幕，统计指定物体的数量和回答用户提出的一些普通问题。该模型是 OpenFlamingo 高效地微调而来的，其中还包括在将 Low-rank Adapter（LoRA）方法引入到 gated-cross-attention 和 self-attention 这两个语言模型的组件中。我们的方法涉及构建包含视觉和语言数据的指令模板，用于多模态指令调整，使模型能够理解和遵守人类指令。我们观察到，训练数据的质量对于对话有效性的表现至关重要，因为有限的数据集和短小的回答都可能会导致模型对任何指令都只能生成简短的回复。为了进一步提高 MultiModal-GPT 的对话能力，我们采用仅基于语言的指导来进行联合训练，并结合视觉语言指导的数据。采用相同的指令模板对两种类型的数据进行处理，可以显著提高对话性能。我们的实验展示了 MultiModal-GPT 在与人类保持连续对话方面的熟练表现。代码和演示可在https://github.com/open-mmlab/Multimodal-GPT找到。

## 1 Introduction

人类通过多种渠道与世界进行交互，包括视觉和语言，每种渠道在表达和传达世界某些概念方面都具有独特的优势，从而有助于更好地理解世界。人工智能研究的一个中心目标是创建一个通用的助手，能够有效地遵循与人类意图相一致的多模式视觉和语言指令，以完成各种实际任务。

Preprint. Under review.

最近，GPT-4 [11] 在与人类进行多模式对话方面展示了出色的能力。尽管已经观察到 GPT-4 [11] 的杰出表现，但支撑其优异表现的机制仍然不为人知。诸如 Mini-GPT4 [17] 和 LLaVA [8] 的研究已经通过编码方式将视觉表示和 LLM 的输入进行空间对齐，随后利用 LLM 中的原生的 self-attention 来处理视觉信息，试图在视觉模型也复制这种性能。然而，将这样的模型与详细的或时空视觉信息相结合可能会带来计算上的负担，因为图像 token 的数量可能很大。此外，这两个模型都采用 vicuna [2]，这是一个从 ChatGPT 的用户生成对话中数据中微调 llama [16] 从而完成优化的开源聊天机器人，不过在他们的研究中则省略了对语言指令调整阶段的描述。

为了解决这些挑战，我们的工作建立在 OpenFlamingo 框架 [1] 基础上，这是一个多模式预训练模型，它部署一个 perceiver resampler 来有效地从视觉编码器中提取视觉信息，并同时使用 gate-cross-attention 来进行图像和文本的交互。该模型已在大量的图像-文本对数据集上进行了预训练，展示了强大的少样本视觉理解能力。但遗憾的是，它缺乏进行零样本多轮图像-文本对话的能力。因此，我们的目标是使用全面的图像和文本指令数据微调 OpenFlamingo，使模型能够进行更符合人类喜好的对话。利用 OpenFlamingo 的基础优势，我们希望缩小模型现有能力和多模式对话中更准确、类似人类交互的期望结果之间的性能差距。我们将我们的多模式聊天机器人称为 MultiModal-GPT。

为了在模型训练过程中同时适用于语言和视觉指令数据，我们使用统一的指令模板。我们首先使用视觉和语言数据构建指令模板来训练 MultiModal-GPT。我们发现，训练数据对于 MultiModal-GPT 的性能至关重要。某些数据集，例如 VQA v2.0 [3]、OKVQA [9]、GQA [5]、CLEVR [6] 和 NLVR [15] 数据集，会降低 MultiModal-GPT 的对话性能，因为这些数据集的响应限制在一个或两个单词（例如“是”或“否”）。因此，当这些数据集纳入训练过程中时，模型会倾向于生成仅包含一个或两个单词的答案。这种简洁不利于用户友好性。

为了进一步提高 MultiModal-GPT 与人类的交互能力，我们还收集语言数据并定义了一个统一的指令模板来共同训练 MultiModal-GPT。语言指令和视觉-语言指令的联合训练有效地提高了模型的性能。我们展示了各种演示来展示 MultiModal-GPT 与人类进行连续对话的能力。

总之，本研究致力于构建一个通用的、基于视觉和语言的对话助手，能够有效地遵循与人类意图相一致的多模式视觉和语言指令，并执行多样化的实际任务。我们通过利用 OpenFlamingo 框架和 MultiModal-GPT 模型，以及联合训练语言和视觉指令数据，显著提高了模型的对话能力，为实现更加准确、人类化的多模式对话交互奠定了基础。

## 2 统一的模板

我们提出了一种统一的模板，用于集成单模语言数据和多模式视觉语言数据，旨在以协同的方式有效地训练 MultiModal-GPT 模型。这种统一的方法旨在通过利用两种数据模态的互补优势，促进对基本概念的更深入理解，从而提高模型在各种任务中的性能。

### 2.1 语言指令模板

<BOS> Below is an instruction that describes a task. Write a response that appropriately completes the request
### Instruction: {instruction}
### Input: {input}
### Response: {response} <EOS>

表 1: 用于训练模型的语言数据输入序列。其中，instruction、input和response是源数据中的文本。只有response部分和EOStoken 将被计算损失。

我们使用 Dolly 15k 和 Alpaca GPT4 数据集 [12] 作为评估仅语言指令遵循能力的资源。这些数据集是专门设计用于提高语言模型在执行基于指令的任务中的性能。为了确保一致的指令遵循格式，我们利用表 1中呈现的提示模板来组织数据集输入。

### 2.2 视觉和语言指令模板

<BOS> Below is an instruction that describes a task. Write a response that appropriately completes the request
### Image: <image_token>
### Instruction: {question}
### Response: {response}<EOS>
### Instruction: {question}
### Response: {response} <EOS>

表 2: 用于训练模型的视觉和语言数据的输入序列。The {question} and {response} 是源数据中的文本。<image_token> 是表示图像存在的标记。请注意，如果数据集存在多轮对话，则会出现多轮对话。只有 {response} part and <EOS> token 将会被被计算损失.

我们在研究中使用了各种视觉和语言指令遵循数据集，包括 LLaVA [8]、Mini-GPT4 [17]、A-OKVQA [14]、COCO Caption [7] 和 OCR VQA [10]。这些数据集涵盖了广泛的应用和领域，从而有助于全面评估我们模型的性能。

为了以一致的指令遵循格式呈现文本，我们采用表 2中的提示作为模板来组织这些数据集。通过遵循标准化格式，我们确保模型更好地处理信息并做出相应回应。

需要注意的是，COCO Caption 数据集通常不包含指令内容，因为它主要由描述性标题组成。为了克服这个限制并纳入指令数据，我们使用 GPT-4 [11] 模型为 COCO Caption 数据集生成相关指令。这种综合指令的整合丰富了数据集，使我们的模型能够在处理和响应人类指令方面实现更强大的能力。

- Can you describe the image?
- Could you provide a description of the image?
- What do you see in this image?
- Share your thoughts on the content of the image.
- Please narrate what's happening in the picture.
- Can you give a brief explanation of the image?
- Describe the main elements and details present in the image.
- In your own words, what is depicted in the image?
- How would you describe the image's content in a caption?
- Can you suggest an insightful caption that highlights the underlying message of the image?

表 3: The list of instructions for image caption.

表 3展示了各种示例，说明为 COCO Caption 数据集生成的指令，展示了我们的方法在适应数据集以更好地符合我们的研究目标方面的有效性。

## 3 方法

### 3.1 架构

所提出的 MultiModal-GPT 基于 open-flamingo 模型 [1]。如图 1 所示，MultiModal-GPT 由来自 CLIP [13] 的视觉编码器、一个感知器重采样器以接收来自视觉编码器的空间特征，以及一个语言解码器 LLaMA [16] 组成。需要注意的是，为了将视觉特征编码为文本，语言解码器通过交叉注意力被条件化于来自感知器重采样器的空间特征。有关模型架构的更多细节，请参见 [1]。

### 3.2 联合训练

我们同时使用仅语言指令遵循数据和视觉和语言指令遵循数据来联合训练 MultiModal-GPT。如图 1 所示，我们冻结整个 open-flamingo 模型，并在语言解码器的自注意力、交叉注意力和 FFW 部分中添加 LoRA [4] 来微调 MultiModal-GPT。MultiModal-GPT 是通过预测文本的下一个标记进行训练的，只有输入序列中的response和<EOS>标记参与损失计算。

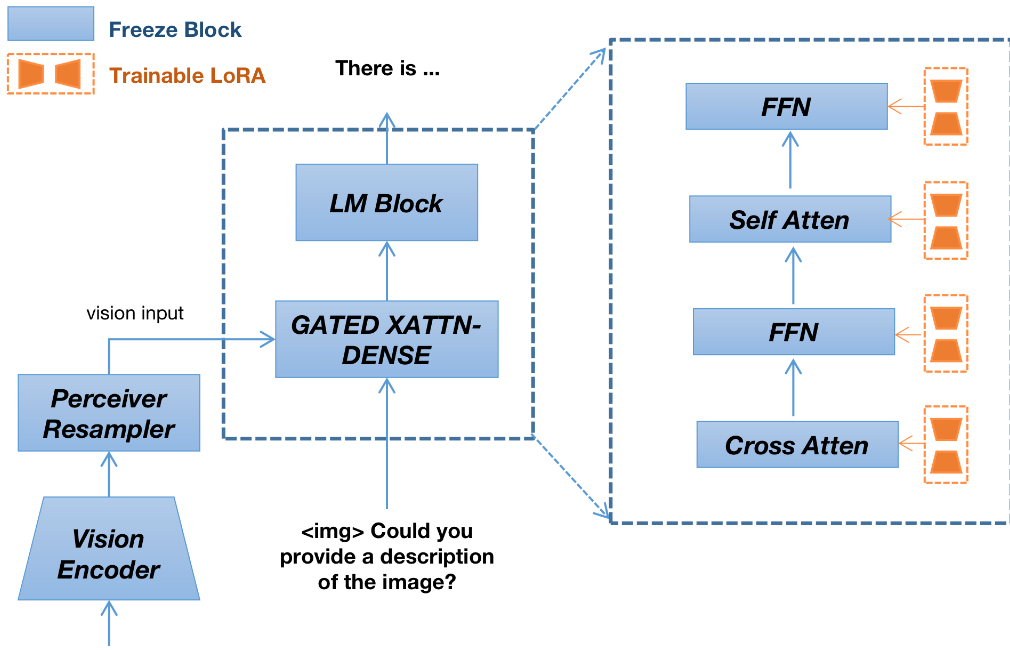


图 1: MultiModal-GPT 的总体框架如下。MultiModal-GPT 包括一个视觉编码器，一个 perceiver 重采样器，用于接收视觉编码器的空间特征，以及一个语言解码器。语言解码器通过交叉注意力来将来自 perceiver 重采样器的空间特征与文本进行编码，以便将视觉特征编码为文本。我们冻结整个 open-flamingo 模型，并将 LoRA 添加到语言解码器的自注意力部分、交叉注意力部分和 FFN 部分中，以微调 MultiModal-GPT。

# 4 实验

## 4.1 实验细节

我们使用多种语言数据和视觉与语言数据源来联合训练 MultiModal-GPT 模型以提高其性能。语言数据集包括 Dolly 15k 和 Alpaca GPT4 [12]，而视觉和语言数据集包括 LLaVA [8]、Mini-GPT4 [17]、A-OKVQA [14]、COCO Caption [7] 和 OCR VQA [10]。这种数据集的组合旨在为 MultiModal-GPT 模型提供多样化和丰富的训练环境。

为了有效地训练模型，我们将 Dolly 15k 和 Alpaca GPT4 数据集的整个文本语料库都纳入其中。同样地，我们包含了 LLaVA 和 Mini-GPT4 数据集中所有可用的图像-文本对，以确保对各种情境和场景进行充分的暴露。然而，相比 LLaVA 和 Mini-GPT4，A-OKVQA、COCO Caption 和 OCR VQA 数据集的质量较低。为了在仍然从额外的数据中受益的同时考虑到这种差异，我们在训练过程中包括了 A-OKVQA 数据集的随机抽样 5000 个图像-文本对，以及 COCO Caption 和 OCR VQA 数据集的每个 512 个图像-文本对。

为了训练模型，我们使用了 8 个 A100 GPU，并在一个 epoch 内完成了训练过程。每个 GPU 的批量大小对于视觉和语言指令遵循数据和仅语言指令遵循数据均设置为 1。我

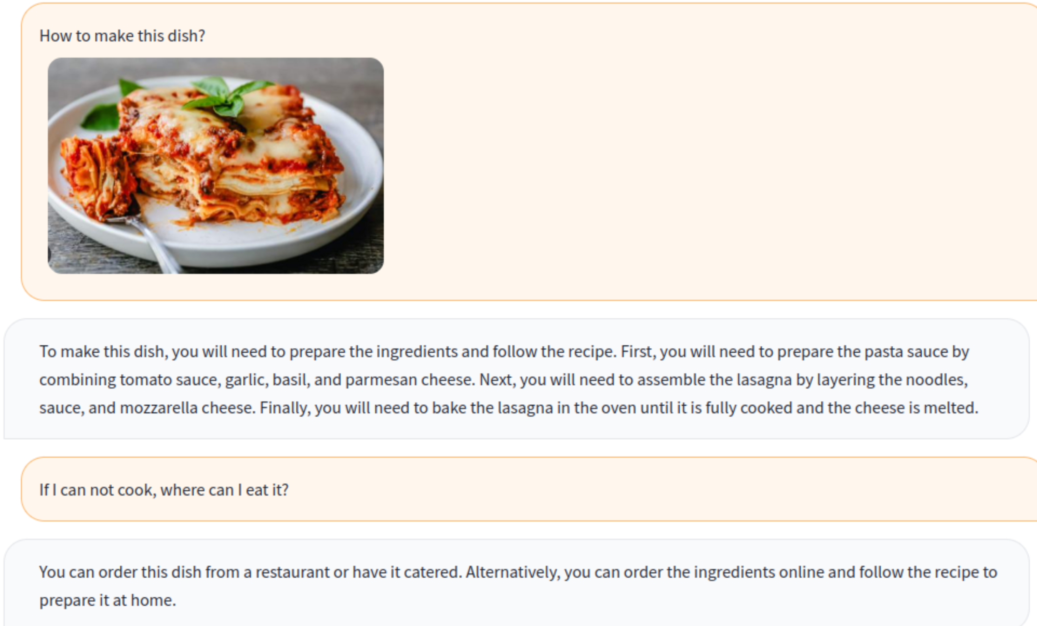


图 2: MultiModal-GPT 可以提供烤宽面条的食谱，并告诉用户去哪里吃。

们使用梯度累积，并在 16 个迭代后更新 LoRA 的参数。每个迭代包含一个视觉-语言对和一个仅语言指令数据。因此，聚合批量大小为 256。我们采用学习率为 1e-5，并使用余弦学习率调度器来调整训练过程中的学习率。

### 4.2 数据质量至关重要

为了增强视觉和语言数据的多样性，我们在初步实验中纳入了多个数据集。然而，我们发现某些数据集的质量并不理想，其中包括 VQA v2.0 [3]、OKVQA [9]、GQA [5]、CLEVR [6] 和 NLVR [15]。这主要是因为这些数据集中的 response 只限于一个或两个词（例如是/否）。因此，当这些数据集被纳入训练过程时，模型往往会倾向于生成仅包含一个或两个词的答案，这对用户友好性不利。因此，我们决定在最终版本的研究中排除这些数据集。

### 4.3 演示

我们提供了多个演示，展示了 MultiModal-GPT 与人类交流的能力。如图 2 所示，当用户在初始对话中询问 MultiModal-GPT 如何制作千层面时，它可以成功地识别出这道菜并提供一个详细的配方。在接下来的对话中，当用户询问可能的餐厅时，MultiModal-GPT 提供了合理的建议。这展示了该模型维持连续对话的能力。

此外，图 4 展示了 MultiModal-GPT 的计数能力，图 6 展示了它的 OCR 能力。MultiModal-GPT 还擅长回答有关用户旅行计划的一般性问题，如图 7 所示。更多演示可以在 https://github.com/open-mmlab/Multimodal-GPT 上找到。

Who is Elon Musk?

Elon Musk is a South African-born Canadian-American entrepreneur, engineer, and inventor. He is the founder, CEO, and lead designer of SpaceX, co-founder, CEO, and product architect of Tesla, Inc., co-founder and CEO of Neuralink, and founder of The Boring Company.

Is this a photo of him?

Yes, this is a photo of Elon Musk.

How many companies does he have?

Elon Musk has three companies: SpaceX, Tesla, and The Boring Company.

图 3: MultiModal-GPT 可以识别 Elon Musk 的图像，并回答有关 Elon Musk 的问题。

How many people in this image?

There are four people in this image.

What is their gender

The people in the image are all women.

What are they doing

The women are wearing sunglasses and posing for a photo.

图 4: The MultiModal-GPT knows there are 4 women in the image, and knows what they are doing.


图 5: MultiModal-GPT 知道图像中有 4 名女性，并且知道她们在做什么。

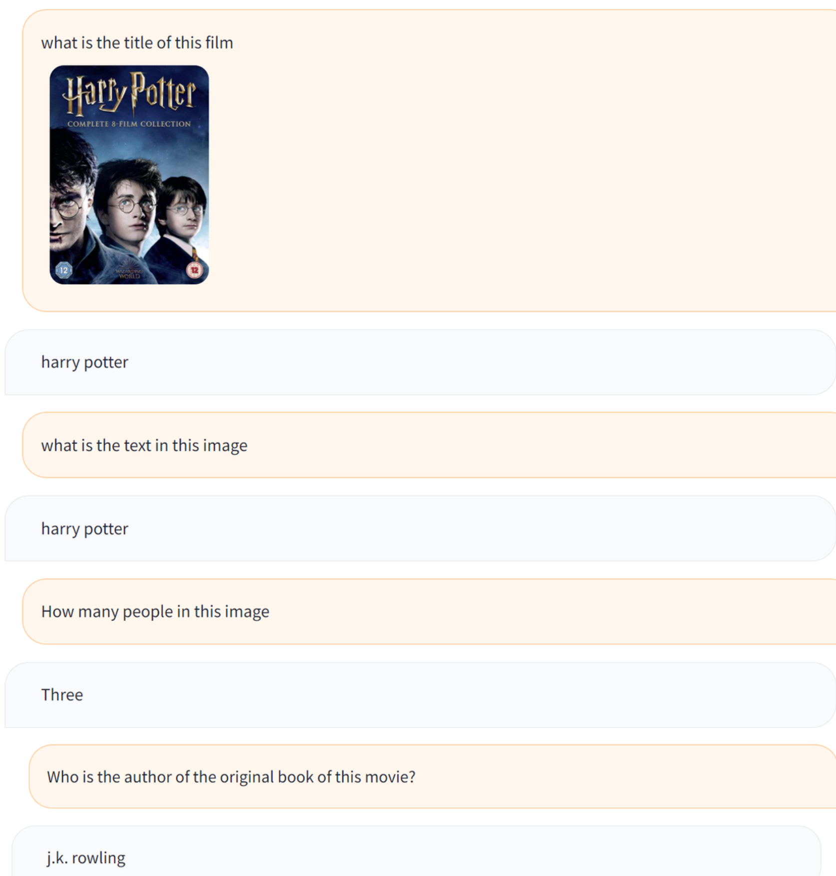


图 6: MultiModal-GPT 可以识别图片中的人物，知道书的作者。

What can I do if I travel there?

If you travel to the Hawaiian island of Kauai, you can enjoy a variety of activities such as swimming, snorkeling, surfing, hiking, and sightseeing. You can explore the island's natural beauty by visiting the Waimea Canyon, also known as the "Grand Canyon of the Pacific", and the Na Pali Coast, which boasts breathtaking cliffs and waterfalls. You can also explore the island's cultural heritage by visiting the National Tropical Botanical Garden and the Kauai Museum. Additionally, the island is home to several resorts and hotels, where you can stay and enjoy amenities such as swimming pools, spas, and restaurants.

How much will it cost?

The cost of traveling to Kauai can vary depending on your choice of accommodation, transportation, and activities. You can find affordable accommodation options, such as hostels and Airbnb rentals, as well as more luxurious options, such as hotels and resorts. Transportation costs can vary depending on your mode of transportation. For example, if you choose to rent a car, you will have to pay for gas and parking fees. If you prefer to use public transportation, you will have to pay for bus tickets. The cost of activities, such as swimming, snorkeling, surfing, hiking, and sightseeing, can also vary depending on your preferences and the availability of equipment.

What else should I prepare for this trip?

To prepare for your trip to Kauai, you should research the island's attractions, activities, and transportation options. You should also check the weather forecast to ensure that you are prepared for any potential weather conditions. Additionally, you may want to bring sunscreen, a swimsuit, and waterproof clothing, depending on your planned activities and the weather.

图 7: MultiModal-GPT 可以回答有关用户旅行的一般问题。

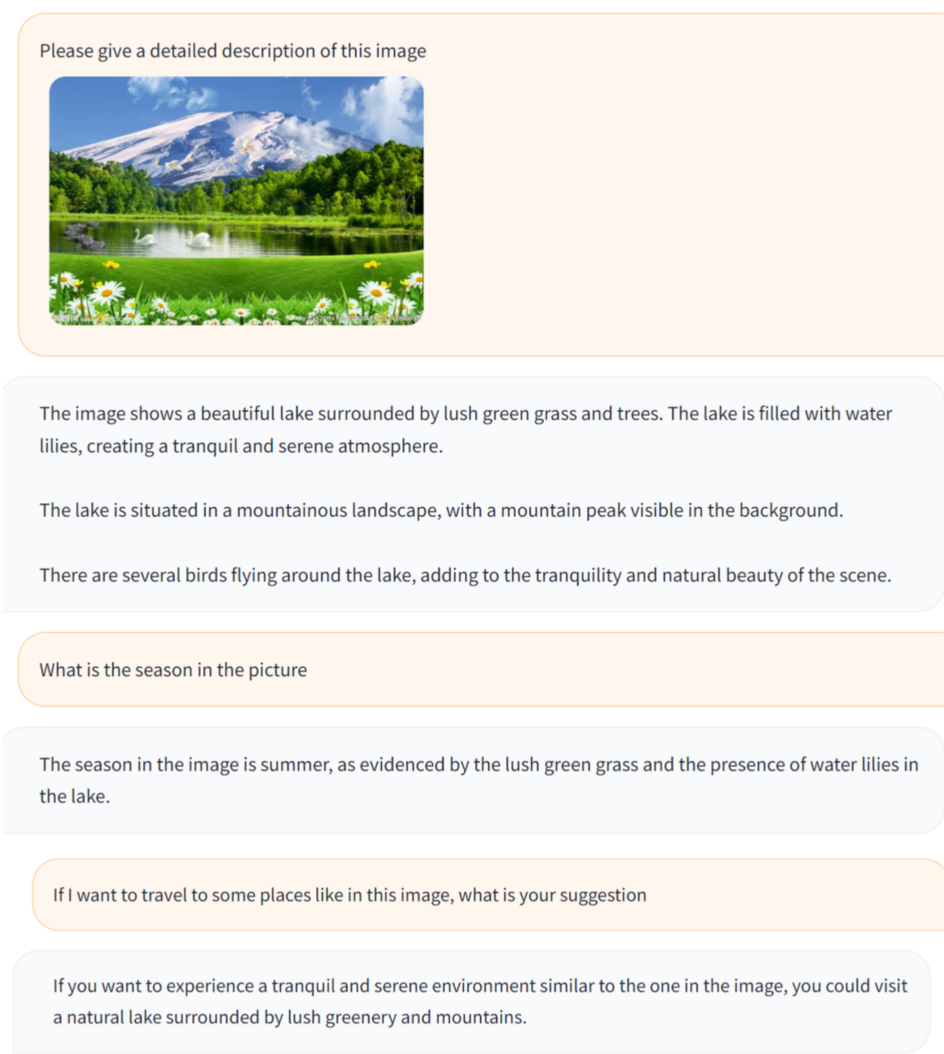


图 8: MultiModal-GPT 可以为图像生成详细的描述，并具有推理图像中季节的能力。

## 参考文献


[1] 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.

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[6] 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.

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[9] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[10] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[11] OpenAI. Gpt-4 technical report. 2023.

[12] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022.

[15] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[16] 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.

[17] 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.