
mPLUG-Owl : 模块化使大型语言模型具有多模态性

叶庆浩*,^{徐海洋}□, 徐国海*, 叶家波, 严明,[†]周益阳, 王俊阳, 胡安文, 史鹏程, 石亚亚, 李晨亮, 徐元红, 陈和红, 田俊峰, 齐倩, 张吉, 黄飞
大摩学院, 阿里巴巴集团
{yeqinghao.yqh, shuofeng.xhy, guohai.xgh, ym119608}@alibaba-inc.com

摘要

大型语言模型 (LLMs) 在各种开放式任务中表现出令人印象深刻的零点能力，而最近的研究也探索了使用LLMs进行多模式生成。在这项研究中，我们介绍了mPLUG-Owl，这是一种新的训练模式，通过对基础LLM、视觉知识模块和视觉抽象模块的模块化学习，使LLM具备多模式能力。这种方法可以支持多种模态，并通过模态协作促进各种单模态和多模态的能力。mPLUG-Owl的训练模式包括两阶段的图像和文本对齐方法，它在LLM的协助下学习视觉知识，同时保持甚至提高LLM的生成能力。在第一阶段，视觉知识模块和抽象器模块与冷冻的LLM模块一起训练，以对齐图像和文本。在第二阶段，通过冻结视觉知识模块，使用仅有的语言和多模式监督数据集来联合微调LLM和抽象器模块的低等级适应 (LoRA) 模块。我们精心建立了一个与视觉相关的指令评价集OwlEval。实验结果表明，我们的模型优于现有的多模式模型，证明了mPLUG-Owl令人印象深刻的指令和视觉理解能力、多轮对话能力和知识推理能力。此外，我们还观察到一些意想不到的令人兴奋的能力，如多图像关联和场景文本理解，这使得我们有可能将其用于更难的实际场景，如纯视觉的文档理解。我们的代码、预训练的模型、指令调整的模型和评估集都可以在<https://github.com/X-PLUG/mPLUG-Owl>。在线演示可在<https://www.modelscope.cn/studios/damo/mPLUG-Owl>。

1 简介

大型语言模型 (LLMs)，如GPT-3 [Brown et al.,2020], BLOOM [Scao et al.,2022], LLaMA [Touvron et al.,2023]经历了快速发展，使通用人工智能成为可能，它在各种语言应用上展示了令人印象深刻的零距离能力。然而，除了GPT-4[OpenAI,2023]，目前的通用人工智能不能支持不同的输入模式，也不能发展令人印象深刻的多模态能力。

尽管GPT-4[OpenAI,2023]表现出了显著的多模态能力，但其非凡能力背后的方法仍然是个谜。最近，研究人员一直在扩展LLMs

*平等贡献

†通讯作者

预印本。正在审查中。

来理解两种不同范式的视觉输入：系统协作和端到端训练模型。然而，系统协作方法，包括Visual ChatGPT[Wu等人,2023]、MM-REACT[Yang等人,2023]和HuggingGPT[Shen等人,2023]，都是为了促进各种视觉模型或工具的协调，用文字描述表达视觉信息。然而，这些方法可能无法理解特定的多模态指令，因为它们缺乏与不同模态的一致性。此外，这些方法可能会遇到与推理效率和成本有关的挑战。端到端模型，如BLIP-2 [Li et al.,2023]，LLaVA [Liu et al.,2023]，和MiniGPT-4 [Zhu et al.,2023a]旨在使用统一的模型来支持不同的模式。然而，这些模型有一些局限性，因为它们采用冻结的视觉模型，由于参数数量有限，可能会导致不充分的对齐。此外，由于缺少单模态和多模态的指导，它们不能释放各种能力。

在本文中，我们提出了mPLUG-Owl，其创新的模块化训练范式适用于大型多模态语言模型，可以同时支持多种模态，这是从模块化的概念中获得的灵感[Xu et al.,2023b,Li et al.,2022,Xu et al.,2021,Ye et al.,2022]。

我们的方法利用预先训练好的LLM、视觉知识模块和连接的视觉抽象模块的力量来实现图像和文本之间的有效对齐，并利用一个两阶段的训练方案来激发令人印象深刻的单模态和多模态的能力。我们的方法甚至通过模态之间的协作，增强了LLM的强大生成能力。在第一步，我们将图像和文本对齐，利用文本-图像对获得全面的视觉知识，这是通过用冻结的LLM模块训练视觉知识模块和抽象器模块完成的。随后，我们用纯语言和多模态指令对mPLUG-Owl进行微调，以解锁一系列单模态和多模态的能力。我们冻结了视觉知识模块，并在LLM和视觉抽象模块上联合训练低等级适应（LoRA）[Hu et al.,2022]。这种方法可以有效地整合文本和视觉信息，促进多功能和强大认知能力的发展。

我们在精心构建的视觉相关指令评估集OwlEval上的实验表明，mPLUG-Owl的性能优于现有的模型，如MiniGPT-4 [Zhu et al.,2023a] 和 LLaVA [Liu et al.,2023]。我们分别验证了mPLUG-Owl在指令理解、视觉理解、知识转移和多轮对话方面的卓越能力。我们进行了大量的消融研究，以显示我们训练范式的有效性。此外，我们还发现了一些未被发现的新兴能力，如多图像关联、多语言对话和场景文本理解。

我们的主要贡献可以强调如下：

- 我们提出了mPLUG-Owl，这是一种通过模块化的大型语言模型的新型训练范式。
- 我们精心构建了一个指令评估集，称为OwlEval，以评估不同模型在视觉相关任务中的能力。
- 实验结果表明，mPLUG-Owl在多模式指令理解和多轮对话方面表现出色，超越了现有模型的性能。

2 相关工作

2.1 大型语言模型

近来，大型语言模型（LLMs）因其在各种自然语言处理（NLP）任务中的卓越表现而获得了越来越多的关注。最初，BERT [Devlin et al.,2019]、GPT [Radford and Narasimhan,2018] 和T5 [Raffel et al.,2020]等转化器模型是以不同的预训练目标开发的。然而，GPT-3 [Brown et al.,2020]的出现，扩大了模型参数的数量和数据规模，展示了显著的零点泛化能力，使他

们在以前未见过的任务上有值得称赞的表现。因此，许多LLMs，如OPT[Zhang et al.,2022]、BLOOM[Scao et al.,2022]、PaLM[Chowdhery et al.,2022]和LLaMA[Touvron et al.,2023]被创造出来，使用LLMs的成功。此外，Ouyang等人[Ouyang et al.,2022]通过将人类的指导和反馈与GPT-3相一致，提出了InstructGPT。此外，它还被应用于Chat-

GPT[OpenAI,2022]，它通过响应广泛的多样化和复杂的查询和指令，促进与人类的对话互动。

2.2 多模式大型语言模型

尽管LLMs在自然语言处理中得到了成功的应用，但LLMs在感知其他模式（如视觉和音频）方面仍在努力。最近，研究人员一直在扩展语言模型以理解两种不同的范式的视觉输入：系统协作和端到端训练模型。系统协作方法，如Visual ChatGPT[Wu et al.,2023]、MM-REACT[Yang et al.,2023]和HuggingGPT[Shen et al.,2023]，利用各种视觉专家或工具用文字描述来表达视觉信息。随后，大型语言模型，如ChatGPT，可以作为代理，并被提示选择合适的专家和工具进行视觉理解。最后，LLMs将总结这些专家的输出，以回答用户的询问。另一方面，一些方法[Li et al.,2023,Alayrac et al.,2022,Liu et al.,2023]利用预先训练好的大型语言模型来建立多模式的统一模型。例如，Flamingo [Alayrac et al.,2022] 冻结了预先训练好的视觉编码器和大型语言模型，并将视觉和语言模态与门控交叉注意力融合在一起，显示出令人印象深刻的少数几个镜头的能力。此外，BLIP-2[Li et al.,2023]设计了Q-Former，将冻结的视觉编码器和大型语言模型的视觉特征与Flan-T5[Chung et al.,2022]和OPT[Zhang et al.,2022]对齐。此外，PaLM-E[Driess等人,2023]直接将来自传感器模式的特征输入PaLM[Chowdhery等人,2022]，它有5200亿个参数，有助于在真实世界的感知中获得强大的性能。此外，一些建立在开源基础模型LaMA[Touvron et al.,2023]上的power-erful指令调整的语言模型，如Alpaca[Taori et al.,2023]和Vicuna[Vicuna,2023]，表现出与ChatGPT[OpenAI,2022]和GPT-4[OpenAI,2023]相仿的性能。MiniGPT-4 [Zhu et al.,2023a] 和LLaVA [Liu et al.,2023] 将这些微调的模型与从冻结的视觉骨干中提取的视觉特征相一致。相比之下，mPLUG-Owl不仅将视觉和语言基础模型（例如CLIP和LLaMA）之间的表述在知识获取和立足于现实世界方面进行了调整，而且还能理解语言和多模式指令，展示了强大的零距离概括和多轮对话能力。

3 mPLUG-Owl

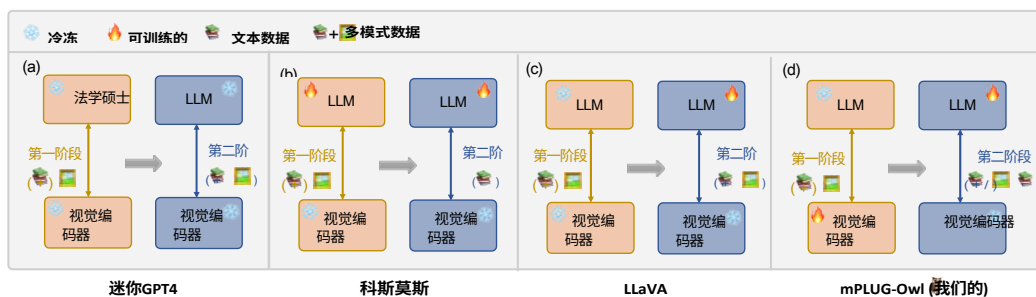


图1：不同训练范式之间的比较。所有这些都是以两阶段的方式进行训练。第1阶段代表预训练，第2阶段代表指令调整。

3.1 架构概述

如图1所示，主要存在三种类型的端到端多模态LLMs：1) 在预训练和指令调谐期间利用有限的参数与冻结的LLM和视觉模型，如MiniGPT4；2) 结合可训练的LLM和冻结的视觉模

型的模型，如Kosmos-1；3）在指令调谐期间涉及可训练的LLM和冻结的视觉模型的模型，如LLaVA。然而，这些模型表现出一定的局限性，因为它们依赖于冻结的视觉模型，由于参数数量有限，这可能导致不充分的对齐。此外，它们不能有效地刺激多样化的能力，因为它们缺乏单模态和多模态的指令。

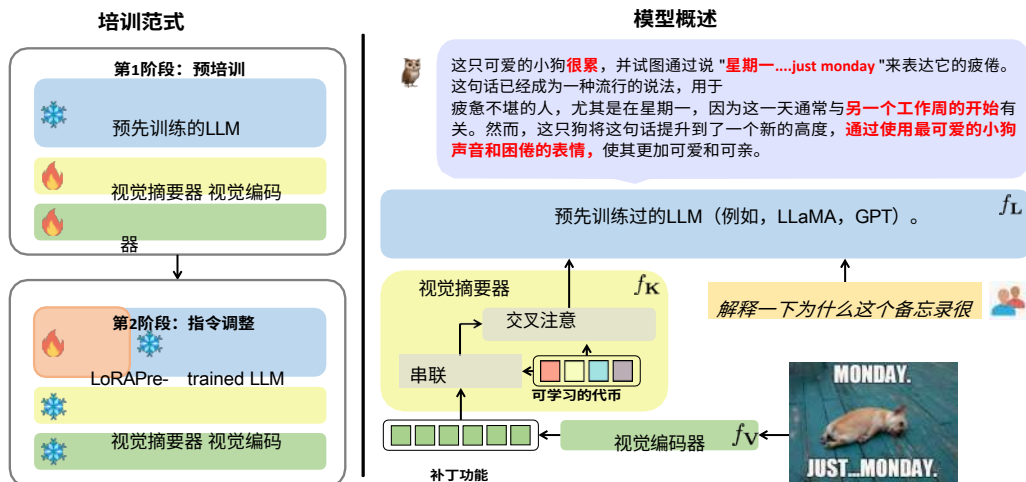


图2：我们的训练范式和模型概述。

为此，我们提出了mPLUG-Owl，一个多模态的语言模型，能够感知各种模态，同时考虑到视觉背景和信息，并产生相应的输出。具体来说，如图2所示，mPLUG-Owl由一个编码视觉知识的视觉基础模型 f_v 、一个语言基础模型 f_L 和一个视觉抽象模块 f_k 组成。我们首先从预先训练好的视觉基础模型 f_v 获得密集的图片表示。然而，这种密集的特征会使细粒度的图像信息支离破碎，并且在输入 f_L 时由于序列冗长而带来大量的计算。为了缓解这个问题，我们采用视觉抽象模块 f_k ，将视觉信息概括为几个可学习的标记，从而获得更高的语义的视觉表征，减少计算，如图2所示。视觉表征与文本查询相结合，并被送入语言模型以产生响应。

3.2 培训计划

多模态预训练 大型语言模型，如GPT-3 [Brown et al.,2020] 和 LLaMA [Touvron et al.,2023]，是在从互联网上收集的广泛而多样的数据上训练的，为它们提供了对世界的全面理解。这个庞大的知识库赋予了这些模型在一系列任务中的卓越能力。然而，在这些模型中对视觉信息的利用仍然没有得到充分的探索。以前的方法[Zhu et al.,2023a, Liu et al.,2023]采用了数量有限的额外参数来学习视觉数据和语言模型之间的一致性，限制了它们理解复杂视觉信息的能力。为了提高大规模语言模型感知视觉信息的能力，同时整合其内部能力，我们提出了一种新的训练范式，即在保持预先训练好的语言模型 f_L 处于冻结状态的同时，加入一个可训练的视觉骨干 f_v 和一个额外的视觉抽象器 f_k 。这种方法使模型能够有效地捕捉低级和高级语义的视觉信息，并在不影响其性能的情况下与预先训练的语言模型保持一致。

联合指令调整 在完成前一阶段后，该模型获得了保留大量知识的能力，并为人类查询提供合理的答案。尽管如此，它在产生连贯的语言反应方面仍然表现出挑战。正如GPT-3 [Brown等人,2020]中提出的，通过指令调整来完善模型对于准确辨别用户的意图至关重要。之前在多模态学习方面的尝试[Li et al.,2022,Xu et al.,2023b]已经证明，由于不同模态之间的协作，从单模态和多模态来源的联合学习可以带来显著的改善。在此基础上，我们提出了一种新的视觉-语言联合指令调整策略，以促进mPLUG-Owl和人类指令及意图之间的更好协调。具体来说，鉴于该模型可以通过视觉知识学习来理解图像中的视觉概念和知识，我们冻结整个模型并采用低等级适应（即LoRA [Hu

et al.,2022])，通过训练多个低等级矩阵来适应 f_L ，以实现与人类指令的高效对齐。对于每个数据记录，我们按照Vicuna[Vicuna,2023]的做法，将它们统一在一个对话片段中，并计算出响应的损失。在训练过程中，我们对纯文本指令数据和多模式指令数据进行了多批次的梯度积累，并对参数进行了更新。因此，通过对语言和多模态指令的联合训练，mPLUG-Owl可以更好地理解各种指令，并以更自然和可靠的输出进行响应。此外，我们的方法可以很容易地处理各种文本和多模式指令，而不需要像MiniGPT-4[Zhu et al.,2023a]和LLaVA[Liu et al.,2023]等方法那样对视觉和语言模型进行重新调整。

训练目标 使用语言建模任务来训练模型，这需要学习根据前面的语境来生成后续的标记。训练过程的主要目标是使标记的对数可能性最大化。值得注意的是，在计算训练损失时只考虑离散的标记，如文本标记。最重要的是，在联合指令调整阶段，由训练任务产生的多样化能力的出现，增强了mPLUG-Owl在下游应用中的性能。

4 实验

4.1 实验设置

模型设置。我们选择ViT-L/14 [Dosovitskiy et al.,2021]作为视觉基础模型 f_v ，它有24层，隐藏维度设置为1024，补丁大小设置为14。为了更快地收敛，ViT从CLIP ViT-L/14模型初始化，通过对比学习进行预训练。与LLaVA[Liu et al.,2023]和MiniGPT-4[Zhu et al.,2023a]不同，为了证明有效性和泛化能力，我们利用原始的LaMA-7B[Touvron et al.,2023]，而不是其指令调整后的变体，如Alpaca[Taori et al.,2023]和Vicuna[Vicuna,2023]。mPLUG-Owl的参数总数约为7.2B。关于超参数的更多细节可以在附录中找到。

数据和训练细节。在第一阶段，我们利用了几个数据集的图像-标题对，包括LAION-400M[Schuhmann等人,2021]、COYO-700M[Byeon等人,2022]、概念性标题[Sharma等人,2018]和MSCOCO[Chen等人,2015]。我们使用210万个令牌的批处理规模，训练mPLUG-Owl 50k步，对应约1040亿个令牌。我们采用AdamW优化器， $\beta=(0.9, 0.98)$ ，并设置学习率和权重衰减分别为0.0001和0.1。我们用2k个预热步骤来预热训练，然后用余弦时间表衰减学习率。输入的图像被随机地调整为 224×224 。此外，我们用SentencePiece[Kudo and Richardson,2018]标记器对文本输入进行标记。在第二阶段，我们从三个不同的来源收集纯文本指令数据：102k数据来自Alpaca [Taori et al.,2023]，90k来自Vicuna [Vicuna,2023]，以及50k来自Baize [Xu et al.,2023a]。此外，我们还利用了来自LLaVA数据集[Liu et al.,2023]的150k多模态指令数据。我们对mPLUG-Owl进行了2k步的训练，批次大小为256，学习率被设置为0.00002。

基线。我们将我们的mPLUG-Owl与端到端模型和系统协作方法进行了如下比较：

- *OpenFlamingo* [Zhu et al.,2023b] 是Flamingo [Alayrac et al., 2022] 模型的开源版本。我们使用OpenFlamingo-9B的发布代码³来运行零点生成。
- *BLIP-2* [Li et al.,2023]是通过引导学习从现成的冷冻预训练图像模型和大型语言模型中使用高效的预训练策略进行预训练。我们使用BLIP-2 ViT-G FlanT5的发布代码来进行零点生成。⁴来进行零点生成。
- *MiniGPT-4* [Zhu et al.,2023a]利用一个单一的投影层，将来自预先训练好的视觉编码

器的视觉信息与LLM对齐。具体来说，他们采用相同的视觉

³https://github.com/mlfoundations/open_flamingo

⁴<https://github.com/salesforce/LAVIS/tree/main/projects/blip2>

在BLIP-2中使用的编码器，一个ViT与他们预先训练的Q-Former相结合，以及Vicuna作为LLM。我们使用发布的演示⁵来进行图像指令的生成。

- *LLaVA* [Liu et al.,2023]应用一个单一的投影层，将图像特征从预先训练好的CLIP视觉编码器ViT-L/14转换成Vicuna的语言嵌入空间。我们使用他们发布的示范⁶来进行图像指令的生成。
- *MM-REACT*[Yang等人,2023]将ChatGPT/GPT-4与各种专门的视觉专家结合起来，实现多模态推理和行动。我们使用他们发布的演示⁷来获得回应。

4.2 量化分析

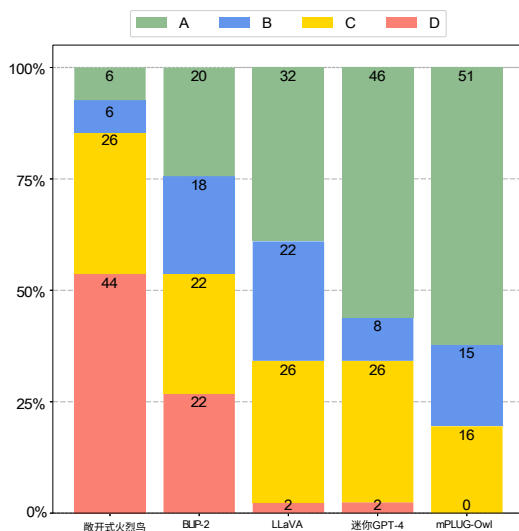


图3：mPLUG-Owl和基线在OwlEval上的人工评价指标对比。响应质量排名的顺序如下：A > B > C > D。

为了全面评估各种模型，我们通过收集基于50张图片的82个人工构建的问题，构建了一个视觉相关的评估集OwlEval，其中21张来自MiniGPT-4，13张来自MM-REACT，9张来自BLIP-2，3张来自GPT-4，4张由我们收集。部分图像有多轮问题，指的是多轮对话情况。这些问题体现了各种模型的能力，包括自然图像理解、图表和流程图理解、光学字符识别（OCR）、多模式创建、知识密集型QA和参考性交互QA。由于问题是开放式的，我们采用人工评估指标，按照Self-Instruct[Wang等人, 2022]中提出的评级方法，将模型的回答评为A、B、C或D。

我们对mPLUG-Owl和基线给出的82个回答进行了人工评分。比较结果显示在图3中。首先，mPLUG-Owl得到66个A和B，而最具竞争力的基线MiniGPT-4得到54个。其次，mPLUG-Owl没有得到任何D分数，超过了所有的模型。这些结果表明，mPLUG-Owl能够更好地理解指令和图像，从而在产生令人满意的响应方面具有更强的能力。为了公平比较，我们排除了那些MM-REACT未能做出预测的情况。结果分别显示在图15中，mPLUG-Owl仍然表现出卓越的性能。

为了分别考察单转和多转的对话能力，我们将82个问题重新组织成一个单转对话集和一个多转对话集。前者包含了50张图片中的第一个问题。后者包含了52个来自多轮对话案例的问题。如图4所示，mPLUG-Owl在单轮对话和多轮对话中都取得了出色的表现。

⁵<https://huggingface.co/spaces/Vision-CAIR/minigpt4>⁶

<https://lava.hliu.cc>

⁷<https://huggingface.co/spaces/microsoft-cognitive-service/mm-react>

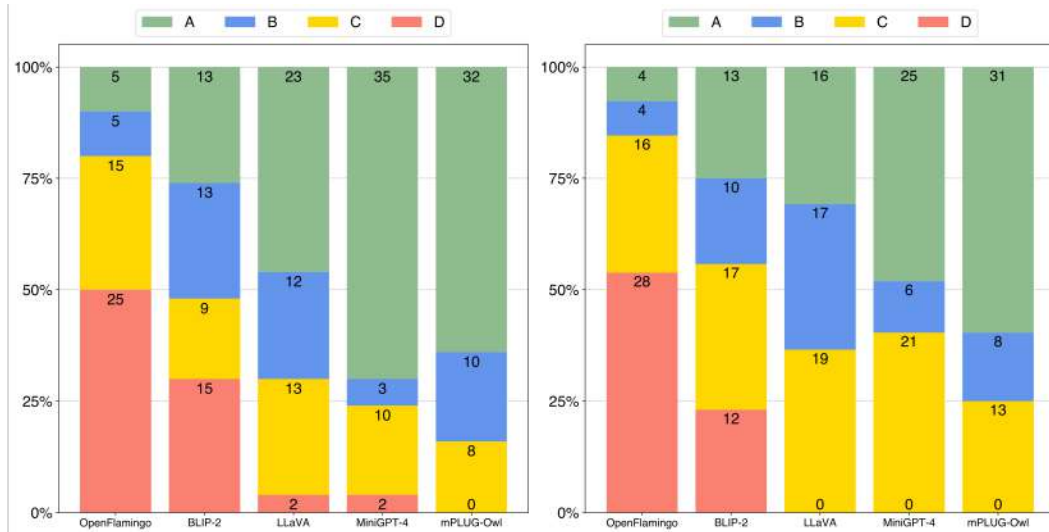


图4：mPLUG-Owl和基线在OwlEval上的50个单匝响应（左）和52个多匝响应（右）的比较结果，并有人工评价指标。

4.3 消融研究

我们取消了两阶段的训练计划和指令调整的数据模式。为了完成视觉相关的任务，我们定义了六个方面的能力，如表1所示。对于每个问题，我们手动标注所需的能力，并注释哪些能力反映在模型的反应中。表2显示了mPLUG-Owl的不同变体的能力准确性。

训练策略的消减。如表2所示，在没有联合指令调谐的情况下，模型对指令的理解能力不强，也不能将训练前的能力推广到其他任务中（r1 vs r5）。在单独进行指令调谐的情况下，虽然模型能够更好地理解指令，但由于缺乏视觉相关知识的预训练，模型在视觉知识相关的任务中无法取得良好的表现（r2 vs r5）。在多模态预训练和联合指令调整的情况下，该模型取得了最好的表现，证明了我们的两阶段训练方案的有效性。

指令数据消减。通过比较r3和r4，纯文本指令调整带来了更多的指令理解的提高，而多模式指令调整则实现了更好的知识边缘和推理能力。这是因为视觉问题的回答主要需要视觉和语言知识的配合，而这一点在纯文本指令调整中没有得到优化。此外，我们还验证了在指令调整过程中引入多模态数据可以进一步提高模型在纯文本任务中的表现，如表3（r5 vs r4）所示。总之，按照Vicuna[Vicuna,2023]的评估设置，对于每个问题，我们将每个模型的回答与ChatGPT的回答配对，并提示ChatGPT⁸来为这两个回答分别打分。表3显示了总分和以ChatGPT分数为参考的分数比。

4.4 定性分析

在本节中，我们展示了我们的评估集OwlEval的定性结果。

知识密集型QA 如图5所示，该指令希望模型能够识别图像中的电影人物。MM-REACT无法对该指令做出有效的反应，而MiniGPT-4理解该指令，但未能回答电影人物。相比之下，mPLUG-Owl回答了图像中五个人物中的四个。这表明mPLUG-Owl对图像中的知识有更好的理解。

⁸由于无法获得GPT-4，我们使用ChatGPT作为次优评分器。

	意义	定义
IU	教学理解	1. 理解文本指导。 2. 不要求有正确的答案，但回答应与指令有关。
VU	视觉理解	1. 识别图像信息。 2. 答案忠实地反映了图像中60%以上的视觉信息。
OCR	光学字符识别	1. 识别图像中的文字信息。 2. 答案忠实地反映了图像中60%以上的文字信息。
KTA	知识转移能力	1. 在语言和视觉之间转移知识。 (1) 了解文字和视觉内容 (2) 调整和转移视觉和语言知识 2. 答案大多是准确的，准确率超过80%。

RA	推理能力	1. 将图像和文字结合起来进行推理。 (1) 了解文字和视觉内容 (2) 进行多步骤推理 (3) 根据多步骤推理过程产生答案 2. 最后的答案基本上是正确的，但缺乏明确的推理过程。吃了或者，最终答案基本正确，推理过程的准确率超过80%。 3. 比如说 (1) 常识性知识推理 (2) 反事实推理 (3) 空间关系推理 (4) 数值计算
MDA	多轮对话能力	1. 理解指示并处理多轮对话 2. 它包括对多个对话的明确引用，并有效处理上下文中的自然语言语义。

表1: 完成视觉相关任务的6种能力和参考文献大多正确，准确率超过80%。

	多式联运 预培训	纯文本 指示	多式联运 指示	能力					
				IU	VU	OCR	KTA	RA	MDA
r1	C			58.5 _(-41.5)	38.1 _(-57.1)	13.3 _(-43.4)	16.7 _(-70.8)	17.1 _(-62.9)	40.0 _(-55.0)
r2		C	C	93.9 _(-6.1)	47.6 _(-47.6)	23.3 _(-33.4)	29.2 _(-58.3)	14.3 _(-65.7)	45.0 _(-50.0)
r3	C	C		93.0 _(-7.0)	73.0 _(-22.2)	40.0 _(-16.7)	41.7 _(-45.8)	48.6 _(-31.4)	80.0 _(-15.0)
r4	C		C	86.6 _(-13.4)	68.3 _(-26.9)	40.0 _(-16.7)	50.0 _(-37.5)	60.0 _(-20.0)	75.0 _(-20.0)
r5	C	C	C	100.0	95.2	56.7	87.5	80.0	95.0
MiniGPT-4 [Zhu et al., 2023a]				97.6	81.0	40.0	83.3	65.7	75.0

表2: 消融结果。每个数值代表相关能力正确反映在模型反应中的问题的比例。IU: 指令理解, VU: 视觉理解, OCR: 光学字符识别, KTA: 知识转移能力, RA: 再认能力, MDA: 多轮对话能力。

多轮对话 图6中的指令要求模型根据参考信息来识别图像的内容。基准模型在面对与空间方向、人类行为和问题中的目标属性相关的指代表达时经常犯错，而mPLUG-Owl则提供了最准确的回答。这种能力源于mPLUG-Owl对图像的精细理解，使其能够根据指令中的指代信息找到图像的相应部分。

推理 图7显示了一个指令，要求模型根据视觉形成给出一个预测，并解释原因。

模型	调谐策略	模型得分	聊天记录得分	比率
羊驼-7B	全程	573	708	80.93%
维库纳-7B	全程	612	684	89.47%
mPLUG-Owl w/o多模态调谐 (r4)	贷款利率	587	682	86.07%
mPLUG-Owl (r5)	贷款利率	600	692	86.71%

表3: 由Chat-GPT评估的Vicuna[Vicuna,2023]的80个纯文本问题的表现。



图5: 知识密集型QA的比较。



图6: 多转弯对话的比较。

阵容和战术的各个方面，并利用它们来推理结果。虽然MiniGPT-4的表现也不错，但其推理的说服力略逊于mPLUG-Owl。

笑话理解 图8中的案例来自于GPT-4[OpenAI,2023]，它要求模型理解并解释一个与视觉有关的笑话。GPT-4不仅按照指示逐个面板进行分析，而且几乎完美地理解了充电方法的幽默感。这主要是由于我们的训练数据中视觉信息的限制。更多关于笑话理解的案例显示在图9中。

MM-REACT	MiniGPT-4	mPLUG-Owl
 <p>Score: C</p> <p>Analyze the tactical characteristics and lineups of both sides and predict who will win.</p> <p>I do not have enough information about the teams. This question requires further context. Please provide more information about the teams, their recent performance, or the context of the competition, so I can help you with your prediction.</p>	 <p>Score: A</p> <p>Analyze the tactical characteristics and lineups of both sides and predict who will win.</p> <p>Argentina and France are two of the strongest teams in the World Cup, with both sides having the potential to win the tournament. Argentina has a strong attacking lineup, with Lionel Messi leading the charge up front and Sergio Agüero providing backup. The midfield is also strong, with Javier Mascherano and Lucas Riglia providing stability and creativity. The defense is led by Nicolás Otamendi and Marcos Rojo, with Gabriel Mercado providing depth. France, on the other hand, has a strong team overall, with a solid defense led by Samuel Umtiti and Raphaël Varane. The midfield is also strong, with N'Golo Kanté and Paul Pogba providing stability and creativity. The attack is led by Kylian Mbappé and Olivier Giroud, with Antoine Griezmann providing depth. In terms of tactics, both sides play a similar style, with a focus on possession and quick transitions. Argentina tends to play a more direct style, while France is more patient and looks to build from the back. Overall, it's difficult to predict who will win this match, as both sides have strong players and tactics. However, France may have a slight advantage due to their overall team strength and ability to control the game. I would give France a slight edge and predict a 2-1 victory.</p>	 <p>Score: A</p> <p>Analyze the tactical characteristics and lineups of both sides and predict who will win.</p> <p>Based on the tactical characteristics and lineups of both Argentina and France, the game is expected to be a tight, competitive encounter. Argentina's 4-3-3 formation with Lionel Messi as the focal point of their attack is well-known for its ability to create chances through the use of swift passing and individual skill. Their lineup includes two experienced strikers (Sergio Agüero and Paulo Dybala) who are known for their goal-scoring capabilities, as well as an attacking midfielder in Nicolás Pope, who is capable of providing creative support and linking the midfield and front line. France's 4-4-2 formation is also known for its ability to create goals through its attacking midfielders (Olivier Giroud, Paul Pogba, and N'Golo Kanté) and forwards (Antoine Griezmann and Kylian Mbappé). This lineup is built around a strong defensive unit led by Raphaël Varane, who is a world-renowned center back and one of the best defenders in the world. Both teams have the potential to create and score goals through their attacking prowess. However, the match is expected to be a tactical battle between Argentina's attacking flair and France's defensive solidity. Argentina's attacking midfielders and forwards may find it difficult to break down France's strong defensive unit, while France's attacking lineup may find it difficult to create chances against Argentina's solid backline. Overall, it is difficult to predict a winner based on the tactical characteristics and lineups, as both sides have the potential to create and score goals. The result may ultimately be determined by the players' individual performances and the ability of each team to adapt to the changing game situations.</p>

图7：推理QA的比较。




MM-REACT	GPT-4	mPLUG-Owl
 <p>Score: C</p> <p>What is funny about this image? Describe it panel by panel.</p> <p>To summarize, this image shows a close-up of a cell phone with a cable plugged into it, a finger, and another device. The image contains various objects related to mobile phones, gadgets, and electronic devices.</p>	 <p>Score: A</p> <p>What is funny about this image? Describe it panel by panel.</p> <p>The image shows a package for a "Lightning Cable" adapter with three panels. Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port. Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it. Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end. The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.</p>	 <p>Score: B</p> <p>What is funny about this image? Describe it panel by panel.</p> <p>This image is funny because it shows a person holding a cell phone and a USB cord, with a small battery and a USB charger plug. The person is using the USB charger to charge the phone using the battery, which is an unusual and comical way to charge a cell phone. It is humorous because it shows the person's resourcefulness in finding ways to charge their phone without the usual charging cable or plug-in. Additionally, the person's hand holding the phone and the battery is positioned in a way that makes it appear as if they are trying to keep the battery and the phone together, adding to the comical element. Overall, the image is funny due to the unusual charging method, the positioning of the hand holding the phone and the battery, and the small battery and USB charger being combined.</p>

图8：笑话理解的比较。






 <p>Score: A</p> <p>解释一下为什么这个备忘录很</p> <p>这个备忘录的幽默感来自于它对词语的俏皮使用，以及对一只可爱的、疲惫的法斗犬的视觉描述。它可能是在经历了一天之后躺在地上。短语“只是...星期一”，这句话通常用来表达对新的一周开始的疲惫或沮丧的感觉，而这只狗对疲惫的表达完美地抓住了这种感觉。</p>	 <p>Score: A</p> <p>这幅画有什么不寻常之处？</p> <p>这个场景的不寻常之处在于，一名男子在繁忙的纽约市街道上购买衣服，而背景是一辆大型黄色出租车驶过。通常情况下，人们不会在这样的城市环境中买衣服，因为在更私密或更安静的地方买衣服更常见。街头场景和一个在公共场合买衣服的男人相结合，可以创造一个幽默的对比。这就是为什么它被认为是有趣的。</p>	 <p>Score: A</p> <p>描述一下这个形象</p> <p>图片展示了一只坐在椅子上、身穿蓝色运动衫的猫。周围是一堆饼干。这只猫似乎正在享受这些饼干，因为它的两只前爪放在这些饼干上。椅子上放着一张桌子，桌子上散落着多个饼干。可能用来盛放饼干的。这只猫出现在一个不寻常的环境中，增加了画面的幽默感。这只猫似乎穿着整齐，这使它看起来对这种情况和环境更加好奇。饼干、舒适的椅子和各种样子的饼干在房间里创造了一种舒适和放松的气氛。使这个场景看起来像是一个有趣和愉快的方式来度过一个</p>
 <p>Score: A</p> <p>一个人为什么会觉得这个画面很有趣？</p> <p>该图片展示了一个有趣的场景，一只松鼠和他在拿着数码相机拍照。松鼠以其顽皮的行为而闻名，因此看到一只松鼠使用相机来拍摄他们的滑稽动作会感到很有趣。此外，这个场景以艺术风格呈现，增加了滑稽的气氛，使其更有趣。看起来严肃的松鼠和画面中的轻松性质并列，创造了一种幽默和俏皮的感，吸引了观众，引</p>	 <p>Score: B</p> <p>这幅画有什么有趣的地方？请一描述。</p> <p>这张图片很有趣，因为它显示一个人拿着一部手机和一根USB线。上面有一个小电池和一个USB充电器插头。这个人正在用USB充电器给手机充电，这是一种不寻常、滑稽的手机充电方式。它之所以幽默，是因为它显示了这个人机智，找到了不寻常的充电线或插头为手机充电的方法。此外，这个人拿着手机和电池的手的位置，让人觉得他们似乎想把电池和手机放在一起，增加了喜剧元素。总的来说，由于不寻常的充电方法，拿着手机和电池的手的位置，以及小电池和USB充电器结合在一起，这幅图片很有趣。</p>	<p>Score: A</p> <p>解释一下为什么这个图像很有趣</p> <p>这张图片很有趣，因为它在一个不寻常、出乎意料和幽默的环境中描绘了一只被驯养的动物。在这个场景中，一只可爱的猫坐在一件蓝色运动衫上，周围是一大堆饼干和几个椅子。这只猫穿着运动衫，使它的处境更加不寻常。这种类型的场景通常与人类的行为有关，例如，当一个孩子穿上他们的万圣节服装参加一个意外的活动。这只猫出现在这个场景中，使其成为一个有趣、有趣、轻</p>

图9：通过mPLUG-Owl理解笑话的更多案例。

5 讨论和局限性

在这一节中，我们展示了mPLUG-Owl的一些尚未完全开发的能力，并讨论其局限性。本节中的部分案例（没有分数）不在OwlEval中。

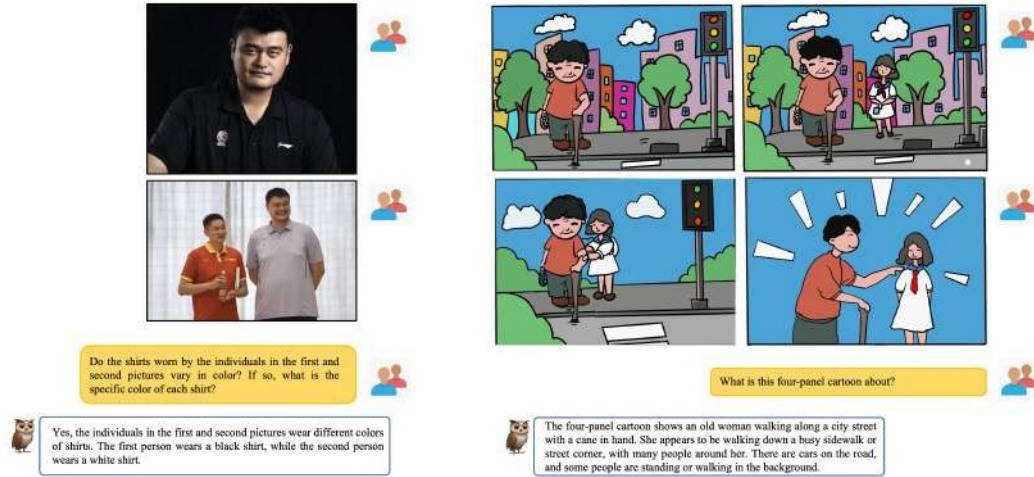


图10：多图像相关案例。

多图像相关性 在图10中，mPLUG-Owl显示了一种新兴的但并不强大的跨多图像的视觉相关性能力。在左边的案例中，该模型可以在两张图片中识别出一个相同的人，并正确地分辨出布的颜色差异。但是在左边的案例中，该模型未能将4张图片联系起来，并产生了一些文字幻觉。

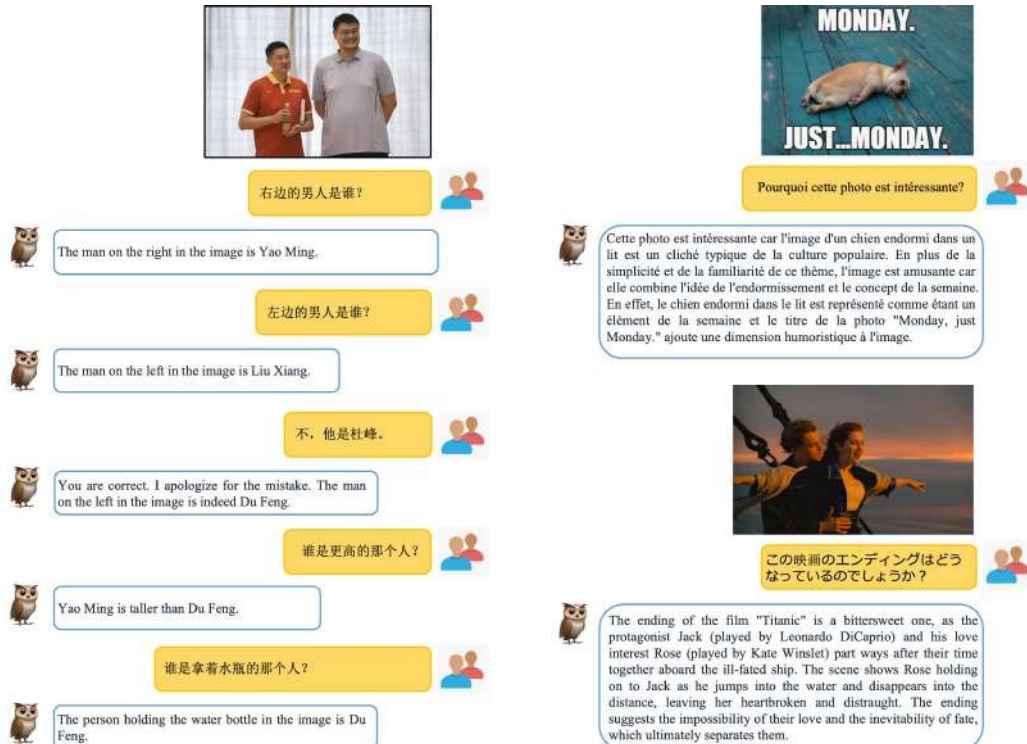


图11：多语言理解的实例提示，分别展示了中文、法文和日文的多语言能力。

多语言对话 除了英语，我们还进一步测试了该模型的多语言能力。如图11所示，虽然在我们的两阶段训练中没有多语言数据，但mPLUG-Owl对中文、法文和日文的多语言理解能力很强。我们主要将这种能力归功于LLaMa[Touvron等人,2023]中的原始文本知识。然而，由于缺乏多语言训练，mPLUG-Owl可能无法对相应的语言做出反应。

场景文本理解 在图16中，mPLUG-Owl展示了其在一些简单场景中的OCR能力，但我们可以看到，该模型对图像中的数字的感知仍然有限。然而，对于复杂场景的OCR，如图17-18所示，mPLUG-Owl的表现比较一般，主要是因为对图像中数字的感知比较弱，影响了后续的推理计算。

纯视觉的文档理解 虽然我们没有使用任何文档注释数据进行训练，但该模型显示了一些文本识别和文档理解能力。因此，我们深入研究了我们的模型的文档理解和功能的结合。如图12所示，我们探索了电影评论写作、代码生成、代码解释、聊天总结和应用程序指导。该模型在(a)和(b)中显示了不错的性能，但仍有一些错误。同时，在(d)、(e)和(f)中，它无法提供可用的回答。

(f).因此，我们的模型在文件理解和下游应用方面的潜力还有待进一步探索。



图12：关于各种文件理解和应用的例子。

开放式创作 mPLUG-Owl在创作诗歌、歌词、广告和其他基于图像的作品方面表现良好。

它在某些情况下的表现如图13-14所示。然而，对于更多功能性和实用性的创作，还需要进一步探索。



你能为这个画面写一首美丽的诗给我吗？



Score: A



当一天接近尾声时，这个年轻人和他的金毛猎犬分享了一段陪伴的时光。太阳开始落山，将它的光芒投射在城市景观上，创造了一个如画的场景。他们坐在山顶上，被大自然的美景所包围。这个男人把他的宠物紧紧地抱在怀里，分享爱和感情的纽带。他们的联系超越了语言，他们的共同经历创造了一种和平的感觉。当这个男人俯身向他的爱犬时，他感受到了它的皮毛的温暖和它给予的无条件的爱的回报。这个男人和他的狗之间的特殊时刻捕捉到了伴侣关系、信任和共度时光的喜悦的本质。随着夕阳西下，他们之间的联系越来越紧密。



帮我写个广告吧。



Score: A



广告为独特而醒目的鹦鹉台灯，这盏灯的设计类似于一只逼真而充满活力的鹦鹉。这盏灯有一个逼真的身体、头和尾巴，由耐用的金属制成，有一个看起来很真实的羽毛细节，为任何空间增添了一丝奇思妙想和个性。这盏灯还配备了一个节能的LED灯，提供明亮而均匀的照明，没有传统灯泡的热量 and 眩光。这款实用而又具有视觉冲击力的台灯是任何生活空间、办公室、卧室，甚至是户外区域的完美补充。它可以作为一个装饰灯，一个夜灯，或者



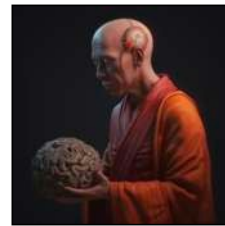
你能根据这幅图给我写一首押韵很好的说唱大师的歌吗？



Score: A



当然，根据这一形象，一首押韵很好的说唱大师的歌曲可以这样唱：
我戴上了太阳镜，穿上了绿夹克，
我已经准备好用我的老式动作来震撼舞池。我手里拿着我的助听器，还有我的助听器，因为我是聚会的祖母，我已经准备好说唱了。当我在摇晃时，我会向你展示我的技能、
我的嘻哈舞蹈动作是如此流畅和滑稽。
老太太们对我在地板上的动作印象深刻，
我的孙子不相信他的祖母会为了更多的东西而这样做。我正在向年轻人展示它是如何做到的，
我是周围最酷的奶奶。



我为我的游戏设计了一个人物，如图所示。你能给我写一个他的背景故事吗？



Score: A



画面中的古人是一位肤色灰白的僧人，穿着传统的僧侣装，一只手拿着一个大脑袋。他站得很高，很骄傲，有一种自信的气质。这位僧人可能来自遥远的土地或遥远的时代，他在那里修炼古老的知识和智慧。他可能一生都在研究和理解大脑的神秘特性及其在人类思维中的作用。他对大脑的知识和理解可能来自于广泛的冥想和沉思，也可能来自于对古代文本和文物的研究。他对知识的追求可能来自于解开意识和人类心理的神秘的愿望，或者仅仅是对人脑及其能力的好奇心。

图13：开放式的创作案例。

6 总结

我们提出了mPLUG-Owl，一种新的训练模式，可以提高大型语言模型（LLM）的多模态能力。我们的方法由模块化的学习基础LLM、视觉知识模块和视觉抽象模块组成，这些模块可以支持多种模式，并通过模式协作促进不同的单模式和多模式能力。我们采用了一种两阶段的图像和文本对齐方法，在LLM的协助下学习视觉知识，同时保持甚至提高LLM的生成能力。实验结果表明，mPLUG-Owl的能力令人印象深刻，表明其在多模态生成方面的各种应用潜力。



Write a guide to travel here.



The Potala Palace in Tibet is an architectural masterpiece that stands as the most iconic and recognizable landmark in the region. Located in the Lhasa Valley, it is a UNESCO World Heritage Site and a symbol of Tibetan cultural heritage.

The Potala Palace was originally constructed as the winter residence of the Dalai Lama, the spiritual leader of Tibetan Buddhism, in the 7th century, and it has since served as a residence for the ruling dynasties of Tibet. The palace is made up of several buildings, including the White Palace, which was the residence of the Dalai Lama himself, and the Red Palace, which was the administrative center for the ruling government.

The Potala Palace is renowned for its elaborate architecture and artwork. The palace's exterior is covered with thousands of intricate carvings, murals and statues, and its interior is adorned with precious Buddhist relics, paintings, and murals.

The palace also boasts stunning views of the surrounding landscape, making it an ideal destination for visitors looking to take in the beauty of the Tibetan countryside and experience a piece of its unique history and culture.

To visit the Potala Palace, one must obtain a travel visa to Tibet, as the region is currently only open to foreign visitors on a limited basis. Once in Tibet, tourists will have to obtain a travel permit from the Tibetan Government to visit the palace.



Write a long tweet to recommend this movie to your friends.



I'd recommend Avatar: The Last Airbender for friends who enjoy action-packed movies with stunning visuals and a unique storyline. The movie takes place in a world of elemental magic and follows the protagonists' struggle to defend their people against the oppressive forces that seek to exploit their resources. The film's diverse cast, including the protagonist, who is a young Avatar with the ability to control all four elements, and a group of friends and allies who support him, make for a dynamic and entertaining watch. It's a great movie for fans of fantasy, adventure, and martial arts.



Write a newscast based on the picture.



The Los Angeles Times is reporting that O.J. Simpson, the infamous American former football player and actor, has been paroled from the Nevada State Prison after spending nearly nine years incarcerated for a 2007 armed robbery.

Simpson was granted parole on Thursday by Nevada's Parole Board, which cited his positive behavior and good conduct during his time in prison. Simpson's release comes as President Donald Trump recently announced a new list of potential pardons, which included Simpson.

In the wake of the news, Simpson's longtime friend, Al Sharpton, issued a statement expressing support for the parole decision, saying that Simpson has been "rehabilitated" and deserves a second chance in life. Sharpton added that Simpson's release is a victory for "all Americans who believe in redemption and second chances."

However, not everyone is celebrating Simpson's release. Many people, including some of the victims of his crimes, have voiced opposition to his release, saying that it undermines the rule of law and sends a negative message to society.

As a result of his parole, Simpson will have several restrictions placed on his freedom, including a curfew, a ban on alcohol consumption, and the requirement to wear an electronic monitoring bracelet. Simpson will also be required to stay in Nevada for the first year of his parole and will not be allowed to return to California, where the murders of Nicole Brown Simpson and Ron Goldman took place in 1994. Simpson will also be required to participate in a mental health program and complete community service work.

Overall, the release of O.J. Simpson has sparked controversy and divided public opinion. While some believe that Simpson has been rehabilitated and deserves a second chance, others believe that his release sends the wrong message about the importance of accountability for crimes committed.

图14：文案写作案例。

参考文献

- J.Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M.Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J.Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O.Vinyals, A. Zisserman, and K. Simonyan.Flamingo: 一个用于少量学习的视觉语言模型。 *CoRR*, abs/2204.14198, 2022.
- T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J.Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei.语言模型是少数学习者。 In *NeurIPS*, 2020.
- M.Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim. Coyo-700m: 图像-文本对数据集。 <https://github.com/kakaobrain/coyo-dataset>, 2022年。
- X.Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick.微软COCO字幕: 数据收集和评估服务器。 *CoRR*, abs/1504.00325, 2015.
- A.Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C.Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y.Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M.Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X.Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A.Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A.Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O.Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. 棕榈: 用路径扩展语言建模。 *CoRR*, abs/2204.02311, 2022.
- H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S.Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A.Roberts, D. Zhou, Q. V. Le, and J. Wei.缩放指令调整的语言模型。 *CoRR*, abs/2210.11416, 2022.
- J.Devlin, M. Chang, K. Lee, and K. Toutanova.BERT: 用于语言理解的深度双向变换器的预训练。在 *NAACL-HLT (1)* 中, 第4171-4186页。 Computational Linguistics协会, 2019年。
- A.Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M.Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby.一张图片值16x16个字: 规模化的图像识别的变形。 In *ICLR.OpenReview.net*, 2021.
- D.Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q.Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhouke, K.Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence. Palm-e: 一个具身的多模态语言模型。 *CoRR*, abs/2303.03378, 2023.
- E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: 大型语言模型的低秩适应。 In *ICLR.OpenReview.net*, 2022.
- T.Kudo and J. Richardson.Sentencepiece: 一个简单且独立于语言的子词标记器和去标记器, 用于神经文本处理。在 *EMNLP (演示)* 中, 第66-71页。计算语言学协会, 2018。
- C.Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F.Huang, J. Zhou, and L. Si. mplug: 通过跨模态跳跃连接进行有效和高效的视觉语言学习

。在*EMNLP*中，第7241-7259页。计算语言学协会，2022年。

J. Li, D. Li, S. Savarese, and S. C. H. Hoi. BLIP-2: 用冻结的图像编码器和大型语言模型对语言-图像进行引导性预训练。 *CoRR*, abs/2301.12597, 2023.

- H.Liu, C. Li, Q. Wu, and Y. J. Lee.视觉指令的调整.*CoRR*, abs/2304.08485, 2023.
- OpenAI.Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022.
- OpenAI.GPT-4技术报告。 *CoRR*, abs/2303.08774, 2023.
- L.Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K.Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P.F. Christiano, J. Leike, and R. Lowe.训练语言模型以遵循人类反馈的指令。 *CoRR*, abs/2203.02155, 2022.
- A.Radford and K. Narasimhan.通过生成性预训练提高语言理解。 2018.
- C.Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu.用一个统一的文本到文本转换器探索转移学习的极限。 *J. Mach.Learn.Res.*, 21:140:1-140:67, 2020.
- T.L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M.Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N.Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I.Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J.Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A.Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, et al.Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model.*CoRR*, abs/2211.05100, 2022.
- C.Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki.LAION-400M: 经过剪辑过滤的4亿个图像-文本对的开放数据集。 *CoRR*, abs/2111.02114, 2021.
- P.Sharma, N. Ding, S. Goodman, and R. Soricut.概念性标题: 一个经过清理的、超链接的、用于自动图像标题的图像alt-text数据集。在*ACL (1)*, 第2556-2565页。计算语言学协会, 2018。
- Y.Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang.Hugginggpt: 用chatgpt和它的朋友们在huggingface中解决AI任务。 *CoRR*, abs/2303.17580, 2023.
- R.Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto.Stanford alpaca: 一个指令跟随的骆驼模型。 https://github.com/tatsu-lab/stanford_alpaca, 2023.
- H.Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E.Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample.Llama: 开放和高效的基础语言模型。 *CoRR*, abs/2302.13971, 2023.
- 维库纳。维库纳: <https://github.com/lm-sys/FastChat>, 2023年, 一个开放的聊天机器人打动了gpt-4。
- Y.Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi.Self-instruct: 将语言模型与自我生成的指令对齐。 *CoRR*, abs/2212.10560, 2022. doi: 10.48550/arXiv.2212.10560.URL <https://doi.org/10.48550/arXiv.2212.10560>.
- C.Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan.Visual chatgpt: 用可视化基础模型进行交谈、绘图和编辑。 *CoRR*, abs/2303.04671, 2023.

C.Xu, D. Guo, N. Duan, and J. J. McAuley. Baize: 一个开源的聊天模型, 在自我聊天数据上进行参数有效调整。 *CoRR*, abs/2304.01196, 2023a.

H.Xu, M. Yan, C. Li, B. Bi, S. Huang, W. Xiao, and F. Huang. E2E-VLP: 通过视觉学习增强端到端的视觉-语言预训练。在 *ACL/IJCNLP (1)* 中, 第 503-513 页。 Association for Computational Linguistics, 2021.

- H.Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S.Huang, F. Huang, and J. Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *CoRR*, abs/2302.00402, 2023b.
- Z.Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang. MM-REACT: 提示多模态推理和行动的聊天记录。 *CoRR*, abs/2303.11381, 2023.
- Q.Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang. Hitea: 分层时间感知的视频语言预训练。 *CoRR*, abs/2212.14546, 2022. doi: 10.48550/arXiv.2212.14546. URL <https://doi.org/10.48550/arXiv.2212.14546>.
- S.Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, 和 L.Zettlemoyer. OPT: 开放的预训练转化器语言模型。 *CoRR*, abs/2205.01068, 2022。
- D.Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: 用先进的大型语言模型加强视觉-语言理解, 2023a.
- W.Zhu, J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, 和 Y. Choi. 多模态C4: 一个开放的、10亿规模的图像与文本交错的语料库。 *CoRR*, abs/2304.06939, 2023b.

A 训练超参数

我们在表4中报告了视觉知识学习的详细模型训练超参数，在表5中报告了视觉-语言联合指令的调整。

超参数	
培训步骤	50,000
热身步骤	375
最大长度	512
图像-标题对的批量大小	4,096
优化器	邓小平
学习率	2e-4
学习率衰减	余弦
亚当 ϵ	1e-6
亚当 β	(0.9, 0.98)
重量衰减	0.01

表4：多模态预训练阶段的训练超参数。

超参数	
培训步骤	2,000
热身步骤	50
最大长度	1,024
文本指令数据的批量大小	128
多模态指令数据的批量大小	128
优化器	邓小平
学习率	2e-5
学习率衰减	余弦
AAA	1e-6
AdamW β	(0.9, 0.999)
重量衰减	0.0001

表5：视觉-语言联合指令调整阶段的训练超参数。

B 与MM-REACT比较

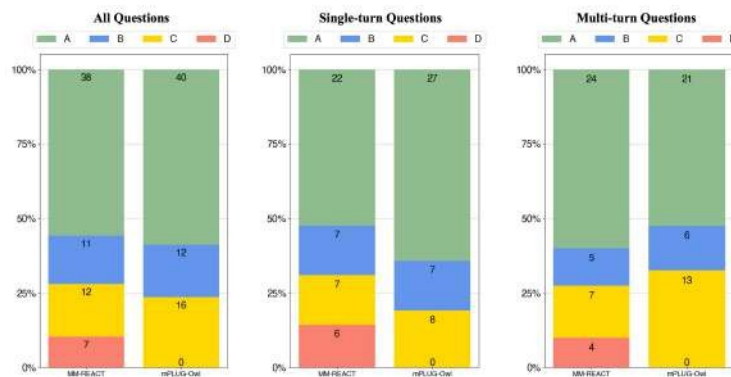
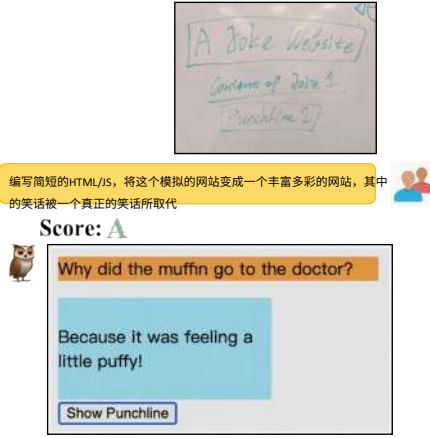



图15：排除了MM-REACT产生的不成功的案例的比较结果。



编写简短的HTML/JS，将这个模拟的网站变成一个丰富多彩的网站，其中的笑话被一个真正的笑话所取代

Score: A



这个平面图上有多少间卧室?

Score: A

在这个平面图中，有两间卧室。

主卧室的尺寸是多少?

Score: C

12' x 10'

主卧室是否有浴室?

Score: A

是的，主卧室里有一个浴室。

第二间卧室的尺寸是多少?

Score: C

11' x 8'

阳台是否比厨房大?

Score: A

是的，阳台比厨房还大。

在这个平面图中，厨房里有哪些电器?

Score: A

炉灶、冰箱、微波炉、洗碗机

当你从入口进入时，你会看到什么? 它的尺寸是多少?

Score: B

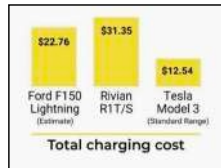
当你从入口进入时，你会看到厨房，里面有一个大冰箱。入口处的门洞有10英尺长，可以看出这个空间的大小和布局。这是一个新房子的平面图，厨房和入口门洞的尺寸显示出一个更大、更宽敞的生活区。

你能根据我感兴趣的方面，用几句话概括一下这个平面图吗?

Score: B

宽敞的楼层规划，有两个主要的生活区，一个大客厅和一个餐厅，都有15英尺宽。
一间主卧室，面积为12英尺x10英尺，还有一间浴室。
一个比厨房还大的露台，提供额外的生活空间和良好的视野。

图16: 简单场景的OCR (主要是数字少且没有计算a的场景)。



不包括福特F150闪电, 平均总充电成本

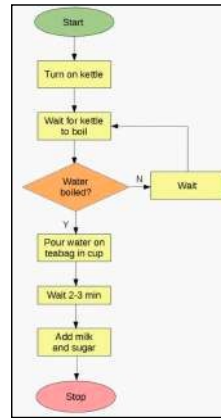
Score: C

包括福特F150 "闪电"在内的三辆电动汽车的平均总充电成本, 可以通过将图片中提供的每辆汽车的每千瓦时电费乘以驾驶里程数来计算。例如, 福特F150闪电每英里消耗12千瓦时, 雪佛兰Bolt每英里消耗35千瓦时, 而特斯拉Model S每英里消耗36千瓦时。那么, 总的充电成本将是: (福特F150闪电的每千瓦时1.02美元) x (福特F150闪电的12千瓦时/英里) x (福特F150闪电的200英里) = \$24.24。对于雪佛兰博尔特, 它将是 (雪佛兰博尔特每千瓦时1.49美元) x (雪佛兰博尔特35千瓦时/英里) x (雪佛兰博尔特200英里) = \$1138.40美元, 对于特斯拉Model S, 它将是 (特斯拉Model S每千瓦时0.94美元) x (特斯拉Model S 36千瓦时/英里) x (特斯拉Model S 200英里) = \$703.20美元。

现在, 电动汽车 (包括福特F150闪电) 的平均充电成本比汽油汽车 (包括福特F150 (普通驾驶室)) 的平均加油成本要便宜多少?

Score: C

计算出汽油车的平均总加油费用: 福特F150普通驾驶室: 平均加油费用为每年1240美元
福特F150普通驾驶室 (EcoBoost): 平均每年的加油费用为1324美元
福特F150 SuperCrew (4x2): 平均加油费用为每年1,360美元 福特F150 SuperCrew (4x4): 平均加油成本为每年1,396美元 福特F150 SuperCrew EcoBoost (4x2): 平均燃料成本为每年1,408美元
计算电动汽车的平均总充电费用: 福特F150闪电: 平均充电费用为5,400美元 Rivian R1T: 平均充电费用为2,490美元
特斯拉Model 3: 平均充电费用为2,490美元
雪佛兰博尔特: 平均充电成本为2,490美元
日产聆风: 平均充电成本为1484美元
计算两个平均总成本之间的差异:
汽油车: 平均总加油费用=37,560美元 (1,240美元+1,324美元+1,360美元+1,396美元+1,408美元)。
电动汽车: 平均总充电费用=47,340美元 (5,400美元+2,490美元+2,490美元+1,484美元)



你能读懂这个流程图中的文字并

Score: B

该流程图显示了准备饮料的说明, 特别是将橙汁和牛奶的混合物放在搅拌机中。该过程包括四个步骤-切橙子, 将其放入搅拌机, 加入牛奶, 然后打开搅拌机。最后一步指示将搅拌好的饮料盛在杯子里。这个简单明了的流程图为遵循指示的人提供了明确的指导, 使其能够轻松地制作橙子和牛奶饮料。

图17: 复杂场景的OCR (a)。



图片中是否有文字？如果有，你能读懂它并向我解释吗？

Score: C

是的，图像中存在文字。文字内容是：“以前..我对自己的长相感觉不好，测试文字，我是不是怀孕了？现在...”。我对自己的容貌感觉不好，测试一下，我有骗局吗？”这条信息是一个笑话，说的是一个人对自己外表的不安全感，以及他们对接受怀孕测试的压力。

这有什么好笑的？

Score: C

这个笑话在几个方面很有趣。首先，它用对比来突出这个人在过去和现在对自己外表的感受不同。过去的自己是不确定的，而现在的自己似乎更确定。这就产生了幽默感，因为读者可能会对自己的外表产生类似的不确定感。其次，该信息还触及了人们在进行怀孕测试时常感到的压力。许多人在做这种测试时可能会感到焦虑或担心，特别是在他们没有试图怀孕的情况下。这两个元素的并列，即这个人的外表不安全感 and 怀孕测试的焦虑，增加了图片的幽默感，并为许多人创造了一个可笑的场景。

Department	Name	Salary
Administration	Emily	\$2,000
Finance	Nick	\$1,800
Marketing	John	\$2,600
Marketing	Gloria	\$1,800
Administration	Jane	\$2,500
Finance	Max	\$2,000
Administration	Tyler	\$3,000
Finance	Wilkins	\$2,500
Marketing	Miles	\$2,800
Administration	Sabine	\$1,900

威尔金斯的收入是多少？

Score: C

威尔金斯的年薪为11万美元。

薪水最高的人是来自哪个部门？

Score: C

工资最高的人是在销售部门。

这个表是报告月薪的，你能不能计算出每个员工的年薪，并汇总到一个表中，按部门分组？

Score: B

是的，我可以根据所提供的月薪信息，提供一个汇总每个员工年薪的表格。该表可以按以下方式排列：姓名|部门|月薪|年薪|

姓名	部门	月薪	年薪
Wilkins	销售	11万美元	132万美元
Smith	销售	9万美元	108万美元
琼斯	销售	68,000美元	816,000美元
布朗	销售	836,000美元	10,032,000美元
史密斯	工程	75,000美元	900,000美元
史密斯	工程	103,000美元	1,236,000美元
史密斯	工程	60,000美元	720,000美元
史密斯	工程	45,000美元	540,000美元
史密斯	工程	60,000美元	720,000美元

通过安排这个表格中的数据，我们可以看到每个员工的年薪和他们各自的部门的清晰分类。我们还可以看到哪些部门的平均年薪最高，哪些部门的平均年薪最低。总之，四个员工坐在显示他们月薪和年薪的桌子前，以及他们的部门的图片，提供了一个快速和翔实的公司薪酬分布概况。

鉴于行政、财务和营销的年度奖金分别为10%、12%和15%。请给我提供Emily、Max和Gloria的奖金。

Score: C

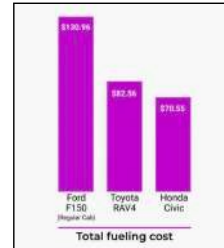
根据所提供的信息，艾米丽、麦克斯和格洛莉亚的奖金分别为10%、12%和15%。这个奖金的计算方法是 $0.10 \times (100-500)/100 = 10\%$, $12\% \times (100-750)/100 = 12\%$, 和 $15\% \times (100-1000)/100 = \text{其基本工资的} 15\%$ 。



总金额是多少？

Score: C

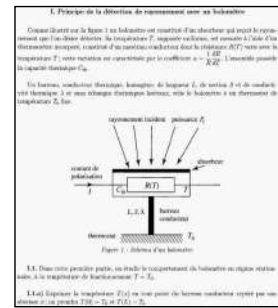
图片中的总金额为33.83美元。



不包括普通驾驶室，平均总加油费用是多少？

Score: C

图片中车辆的平均总加油费用为10,000美元。这是通过将每辆车的总加油费用相加，然后取平均值来计算的。每辆车的每加仑成本各不相同，福特F-150的成本为每加仑2.35美元，福特F-450的成本为每加仑2.44美元，福特Ranger的成本为每加仑2.46美元，福特C-MAX混合动力车的成本为每加仑2.58美元。当这些费用加在一起得到总数时，所有四种车辆的平均加油费用为每加仑2.50美元，40,000加仑的费用为10,000美元。



回答问题1.a. 逐步思考。

Score: C

为了计算加热器的总热损失，我们首先需要确定其热导率。这可以通过将加热器材料的热导率除以其厚度来完成。加热器金属的导热性能通常以W/m·K为单位给出。为了用这个单位计算加热器的导热性，我们需要用这个值除以厚度（毫米）。加热器的金属通常由钢制成，其厚度为5毫米。因此，其热导率计算为150 W/m·K。接下来，我们需要找到从加热器到环境的热流，这是由热电阻代表的。我们知道，热电阻的电阻为R=100欧姆，热阻为Rt=0.01欧姆·C。为了计算加热器的总热损失，我们将加热器的输入功率，即P=1千瓦，乘以热电阻和热阻。结果是10.0W的热量从加热器通过Rt和热电阻损失到环境中。最后，我们可以用热传导的热损失公式来计算加热器的总热损失：加热器的总热损失=0.02 W/m·K * 5 mm * 2 * 1 kW = 10 W。

图18: 复杂场景的OCR (b)。