**mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality**

mPLUG-Owl：模块化赋能多模态大语言模型

**摘要**

大语言模型（LLM）在各种开放式NLP的零样本能力表现出色，最近的研究也探索了LLM的多模态能力。在本研究中，研究人员提出一个基于模块化实现的多模态大模型mPLUG-Owl，由LLM、视觉模块和视觉摘要模块构成。mPLUG-Owl的训练采用对图像和文本的两阶段对齐方法。在第一阶段，使用冻结的LLM模块对视觉模块和视觉摘要模块进行训练，使图像和文本对齐。在第二阶段，冻结视觉模块和视觉摘要模块，使用文本和多模态SFT数据集以[LoRA](https://zhuanlan.zhihu.com/p/646791309)方式指令微调LLM。研究人员构建了一个与视觉相关的指令评估集OwlEval进行评估，实验结果表明，本文模型优于现有的多模态模型。此外，研究人员观察到一些意想不到的能力，如多图关联和场景文本理解，这使得该模型可以用于更复杂的真实场景，如视觉文档理解。代码已开源：<https://github.com/X-PLUG/mPLUG-Owl>

**引言**

大语言模型（LLM）在各种语言应用上显示了令人印象深刻的零样本能力。然而，除了GPT-4外，目前的一般LLM不能支持不同的输入模式。尽管GPT-4 表现出了非凡的多模态能力，但其背后的方法仍然是一个谜。最近，研究人员使用两种不同范式“系统协作、端到端训练模型”扩展LLM来理解视觉输入。

系统协作方法，包括Visual ChatGPT、MM-REACT和HuggingGPT，旨在促进各种视觉模型或工具的协调，通过文本描述来表达视觉信息。然而，这种方法可能无法理解特定的多模态指令，因为它们缺乏模态对齐。此外，这种方法可能会遇到推理效率和成本相关的问题。

端到端模型，如BLIP-2、LLaVA和MiniGPT-4，旨在使用统一的模型来支持多模态。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，由于参数数量有限而导致对齐不足。此外，由于缺少单模态和多模态指令，它们无法解锁各种能力。

受模块化概念的启发，研究人员提出了模块化训练范式的大型多模态语言模型mPLUG-Owl。该模型利用预先训练好的LLM、视觉模块和视觉摘要模块来实现图像和文本之间的有效对齐，并利用两阶段的训练方案来提升单模态和多模态能力。第一阶段，通过图像-文本对将视觉模块和视觉摘要模块向冻住的LLM对齐；第二阶段，冻住视觉模块和视觉摘要模块，使用文本和多模态SFT数据集以LoRA方式指令微调LLM。通过模态间的协作，我们方法能同时促进单模态和多模态任务效果。这种方法可以有效地整合文本和视觉信息，促进模型认知能力的发展。

为了评估方法的有效性，第一次针对视觉相关的指令理解提出一个全⾯的测试集 OwlEval，通过人工评测对比了已有模型。实验结果表明 mPLUG-Owl 展示出更优的多模态能力，尤其在多模态指令理解能力、多轮对话能力、知识推理能力等方面表现突出。除此之外，还涌现出一些意想不到的能力，比如多图关联、文字识别、文档理解等。

本研究主要贡献如下：

（1）提出了mPLUG-Owl，一种模块化的大语言模型训练范式。

（2）构建了指令评估集OwlEval，评估不同模型在视觉相关任务下的能力。

（3）实验结果表明，mPLUG-Owl在多模态理解和多回合对话方面表现优异，超过了现有的模型。

**相关工作**

**大语言模型**

近年来，大语言模型（LLM）因其在各种自然语言处理（NLP）任务中的特殊表现而获得越来越多的关注。最初，诸如BERT、GPT和T5等Transformer模型是用不同的预训练目标开发的。GPT-3的出现扩大了模型参数的数量和数据大小，表现出显著的零样本泛化能力。因此，许多模型如OPT、BLOOM、PaLM和LLaMA被创造出来，引领了LLM的成功。此外，欧阳等人通过将人工指导和反馈与GPT-3对齐，提出了InstructGPT。该方法已被应用于ChatGPT，它通过响应多样和复杂的指令，提升了与人类对话交互的能力。

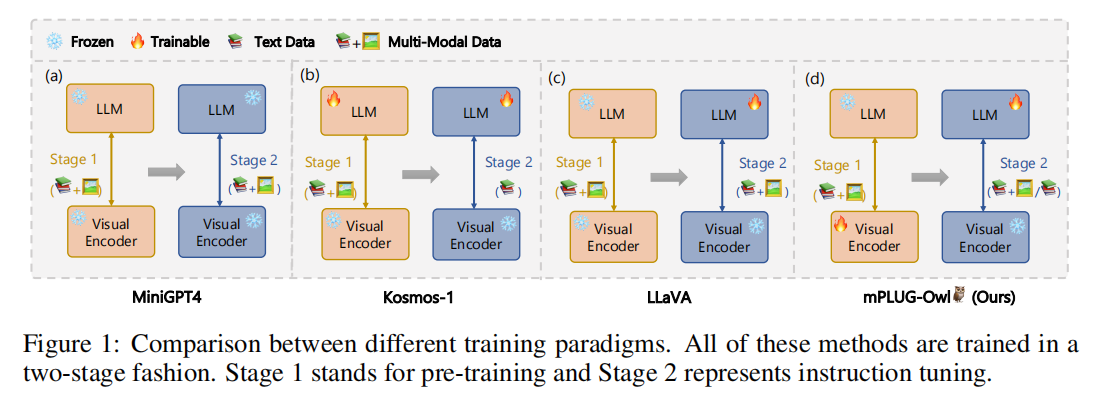
**多模态大语言模型**

尽管LLM在NLP中取得了成功的应用，但LLM感知其他模态仍有困难。最近，研究人员使用两种不同范式“系统协作、端到端训练模型”扩展LLM来理解视觉输入。系统协作方法用各种视觉工具通过文本描述来表达视觉信息。大语言模型（如ChatGPT）可以作为代理，被提示选择适当的工具来进行视觉理解。LLM将总结这些输出来回答用户的查询。端到端训练模型利用预先训练好的大语言模型来建立多模态的统一模型。例如，Flamingo冻结了预先训练过的视觉编码器和大语言模型，并将视觉和语言模式与门控交叉注意力融合起来，显示出令人印象深刻的少样本能力。BLIP-2设计了Q-Frorer，将冻结的视觉编码器输出的视觉特征和大语言模型对齐。此外，一些基于开源基础模型LLaMA的强大的指令微调语言模型，如Alpaca和Vicuna，表现出与ChatGPT和GPT-4相当的性能。MiniGPT-4和LLaVA将这些精细的模型与从冻结的视觉主干中提取的视觉特征对齐。相比之下，mPLUG-Owl不仅将视觉和语言基础模型（如CLIP和LLaMA）与现实世界相结合，还可以理解语言和多模态指令，表现出很强的零样本泛化和多回合对话能力。

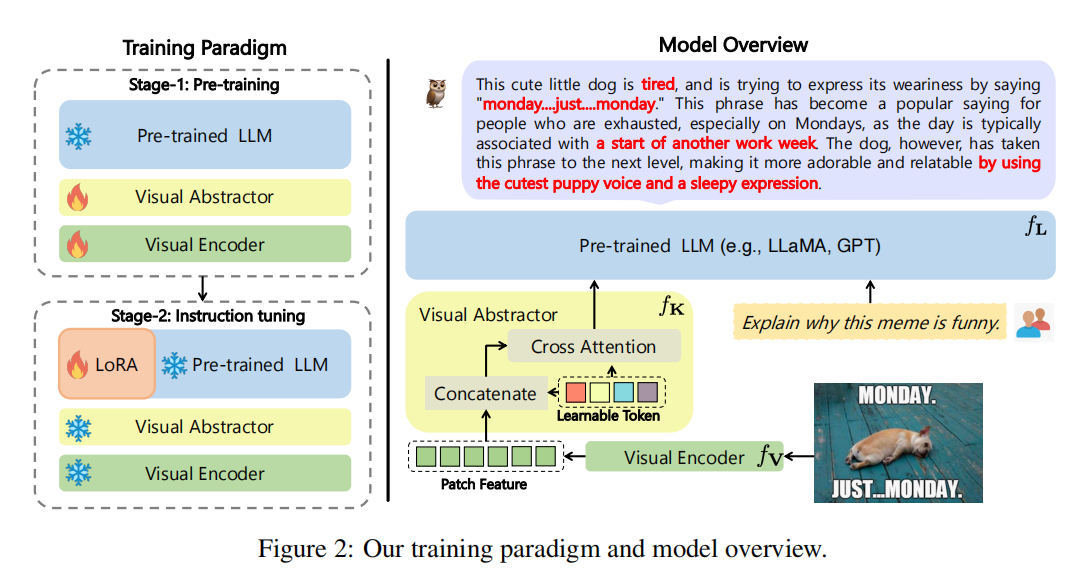
**mPLUG-Owl**

**整体架构**

如图1所示，主要存在三种类型的端到端多模态LLM：（1）在预训练和指令微调过程中冻结LLM和视觉编码器的有限参数模型，如MiniGPT4；（2）包含可训练的LLM和冻结的视觉编码器的模型，如Kosmos-1；（3）包含在指令微调中可训练LLM和冻结的视觉编码器的模型，如LLaVA。这些模型依赖于冻结的视觉编码器，可能由于参数的数量有限而导致对齐不足。



如图2所示，mPLUG-Owl由视觉基础模型fV、语言基础模型fL和视觉摘要模块fK组成。首先从预先训练好的视觉基础模型fV中获得密集的图像表示。这种密集的特征会分割细粒度的图像信息，并且由于较长的序列，会给LLM带来较大的计算量。为了缓解这个问题，研究人员使用视觉摘要模块fK来总结视觉信息，生成可学习的视觉Token，从而获得更高层的视觉表示并减少计算。生成的视觉Token与文本查询相结合，一同输入语言模型以生成响应。



**训练方式**

**多模态预训练**

先前的方法使用有限的额外参数来学习视觉数据和语言模型之间的对齐，限制了理解复杂视觉信息的能力。本文提出了一种新的训练范式，该范式包含了可训练的视觉主干fV和额外的视觉摘要器fK，同时将预训练的语言模型fL保持在冻结状态。这种方法使模型能够有效地捕获低层和高层语义视觉信息，并将其与预先训练的语言模型对齐，而不影响其性能。

**联合指令调优**

第二阶段中，视觉基础模型、视觉摘要模块和原始LLM的参数都被冻结，使用LoRA方法指令微调LLM。基于在mPLUG和mPLUG-2中不同模态混合训练对彼此有收益的发现，mPLUG-Owl在第二阶段的指令微调训练中也同时采用了纯文本的指令数据和多模态的指令数据。对于每个数据记录，按照Vicuna的方法将它们统一在一个对话片段中，并计算响应上的损失。

**训练目标**

该模型使用语言建模任务进行训练，该任务需要基于上下文生成后续文本。训练过程的主要目标是最大化token的对数似然函数。需要注意的是，在计算训练损失时，只考虑了离散的标记，比如文本标记。最重要的是，在联合指令调整阶段，由于训练任务不同，模型产生了不同能力的涌现，这提高了mPLUG-Owl在下游应用中的性能。

**实验**

**模型设置**

视觉基础模型：ViT-L/14，24层，隐藏维度设置为1024，patch大小设置为14。为了更快收敛，采用CLIP的预训练模型参数。

基础LLM：LLaMA-7B。

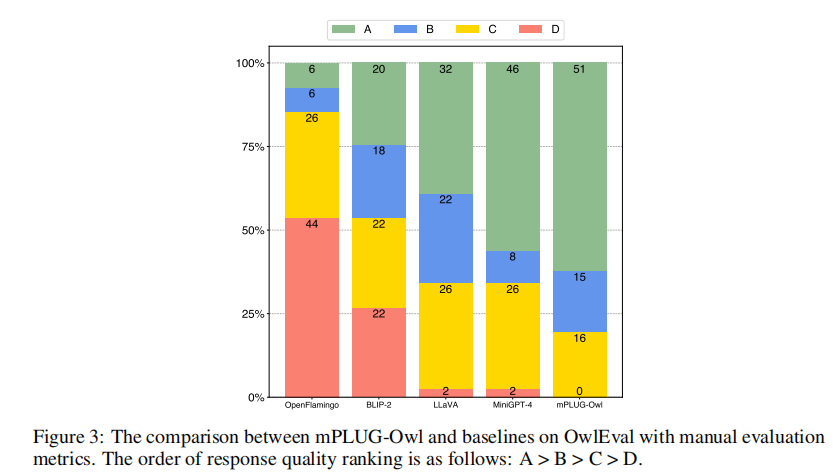
**数据和训练的详细信息**

第一阶段，采用LAION-400M，COYO-700M，CC以及MSCOCO训练视觉基础模块和视觉摘要模块。Batchsize为210万tokens，训练5万步，使用β =（0.9,0.98）的AdamW优化器，学习率0.0001，权重衰减0.1。用2k步来热身训练，然后用余弦调整策略来衰减学习率。输入图像被随机调整到224×224。此外，使用SentencePiece标记器对文本输入进行标记。

第二阶段，从三个不同的来源收集纯文本指令数据：102k from Alpaca+90k from Vicuna+50k from Baize。多模态数据来自LLaVA数据集的150k个多模态指令数据。训练2k步，batchsize为256，学习率设置为0.00002。

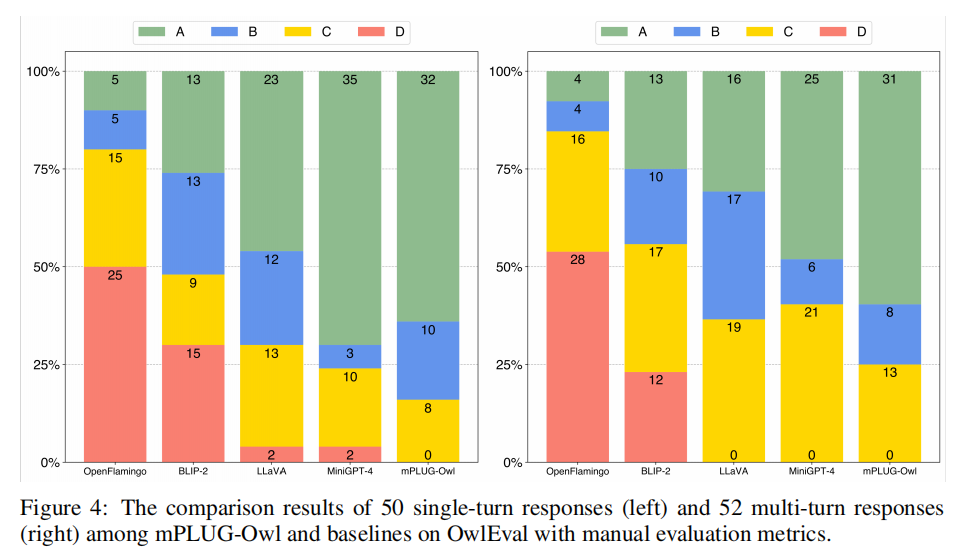
**定量分析**

为了比较不同模型的多模态能力，该工作构建一个多模态指令评测集OwlEval（图片和问题绝大部分都来自已有工作，避免了cherry pick问题）。由于目前并没有合适的评价指标，参考Self-Intruct对模型的回复进行人工评测，打分规则为：A=“正确且令人满意”；B=“有一些不完美，但可以接受”；C=“理解了指令但是回复存在明显错误”；D=“听不懂指令”。对比结果如图3所示，实验证明Owl在视觉相关的指令回复任务上优于已有的OpenFlamingo、BLIP-2、LLaVA、MiniGPT-4。



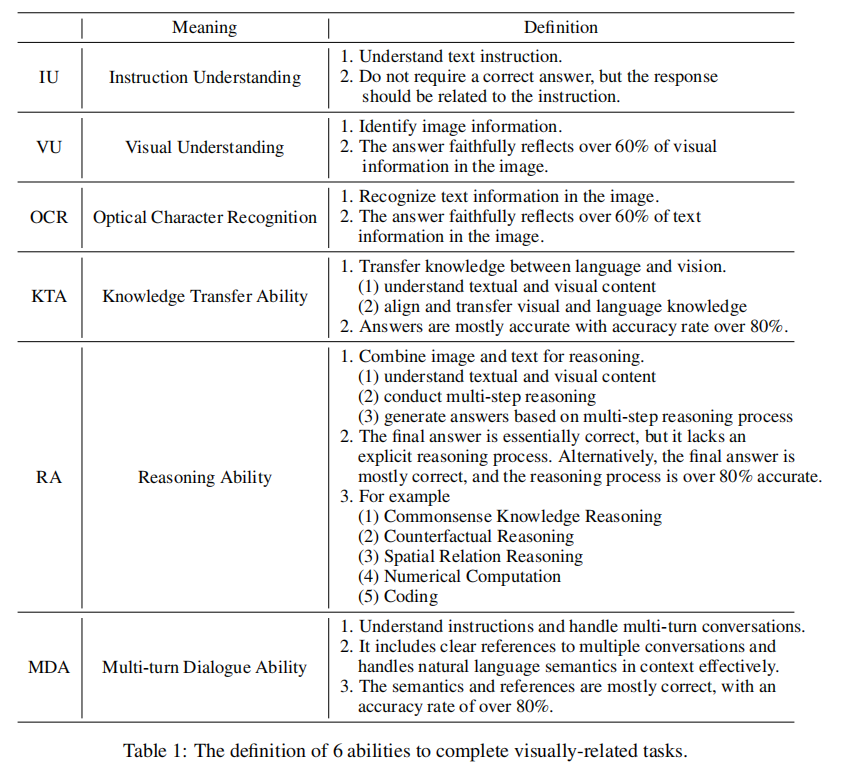
mPLUG-Owl得到66个A和B，而最具竞争力的基线MiniGPT-4得到54个。mPLUG-Owl没有得到任何D分数，表现优于所有的模型。这些结果表明，mPLUG-Owl能够更好地理解指令和图像。

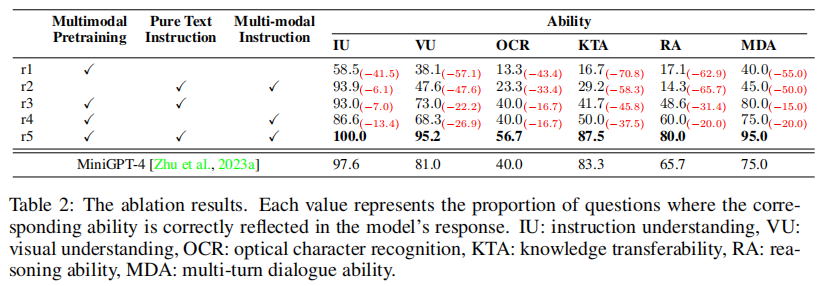
为了分别检查单回合和多回合对话功能，研究人员将82个问题重新组成一个单回合对话集和一个多回合对话集。前者包含了从50张图片中提取的第一个问题。后者包含了52个来自多回合对话案例的问题。如图4所示，mPLUG-Owl在单回合和多回合对话中都取得了出色的性能。



**消融实验**

消除了两阶段的训练方案和指令调优的数据模式。定义了六个能力维度来进行评估，如表1所示。对于每个问题，研究人员手动标记所需的能力，并注释哪些能力反映在模型的响应中。表2显示了mPLUG-Owl的不同变体的能力准确性。



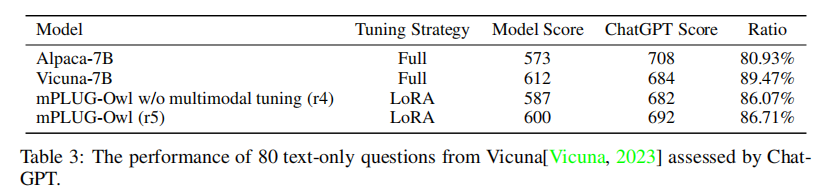


**训练策略消融**

如表2所示，在没有联合指令调整的情况下，该模型不能善于理解指令，也不能将训练前的能力推广到其他任务（r1 vs r5）。仅使用指令调优，虽然模型可以更好地理解指令，但由于缺乏视觉相关知识的预训练（r2 vs r5），该模型无法在视觉相关任务中取得良好的性能。通过多模态预训练和联合指令调整，该模型取得了最好的性能，并证明了本文的两阶段训练方案的有效性。

**指令数据消融**

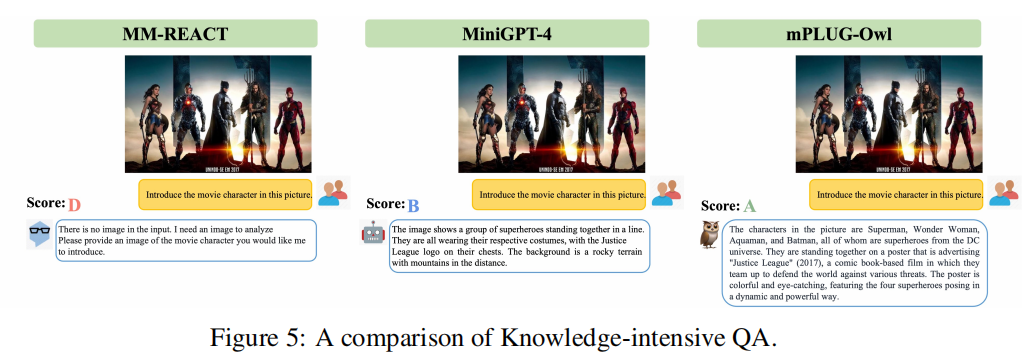
通过比较r3和r4，纯文本指令调优可以进一步提高指令的理解能力，而多模态指令调优可以获得更好的知识和推理能力。这是因为视觉问题的回答主要需要视觉和语言知识的对齐，而这在纯文本指令调优过程中并没有得到优化。此外，还验证了在指令调优过程中引入多模态数据可以进一步提高模型在纯文本任务上的性能，如表3（r5 vs r4）所示。具体地说，对于每个问题，将每个模型的回答与ChatGPT给出的回答配对，并提示ChatGPT分别为这两个回答给出两个分数。表3为以ChatGPT评分为参考的评分。



**定性分析**

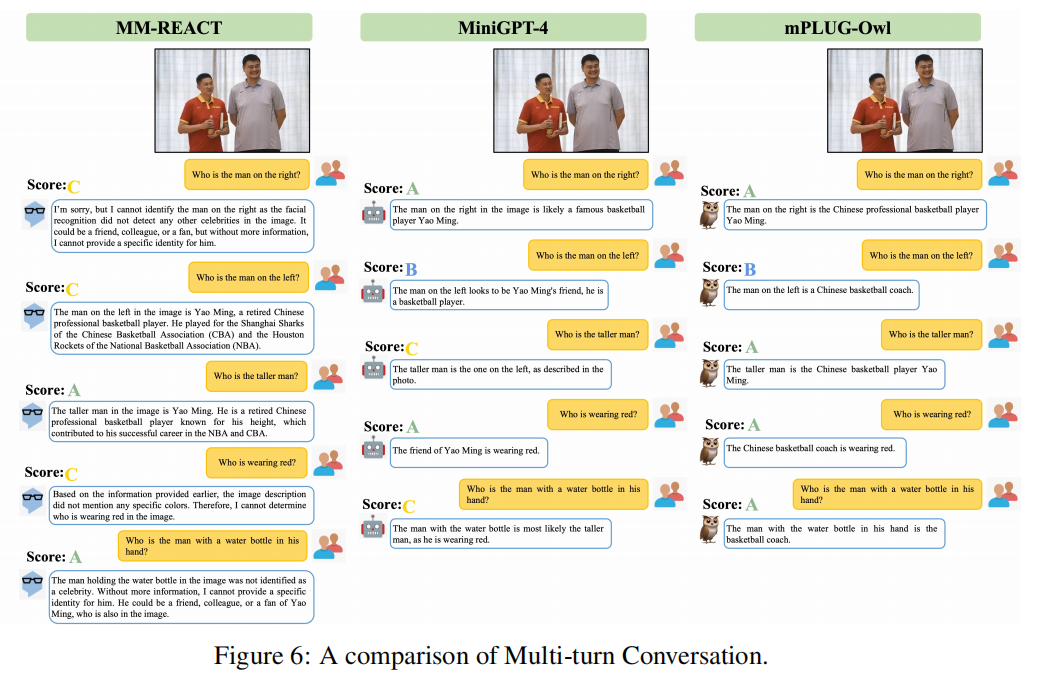
**知识密集型问答**

如图5所示，该指令要求模型能够识别图像中的电影角色。MM-REACT不能对指令提供有效的响应，而MiniGPT-4可以理解指令，但没有回答电影角色。相比之下，mPLUG-Owl回答了图像中五个角色中的四个。这说明mPLUG-Owl对图像中的知识有了更好的理解。



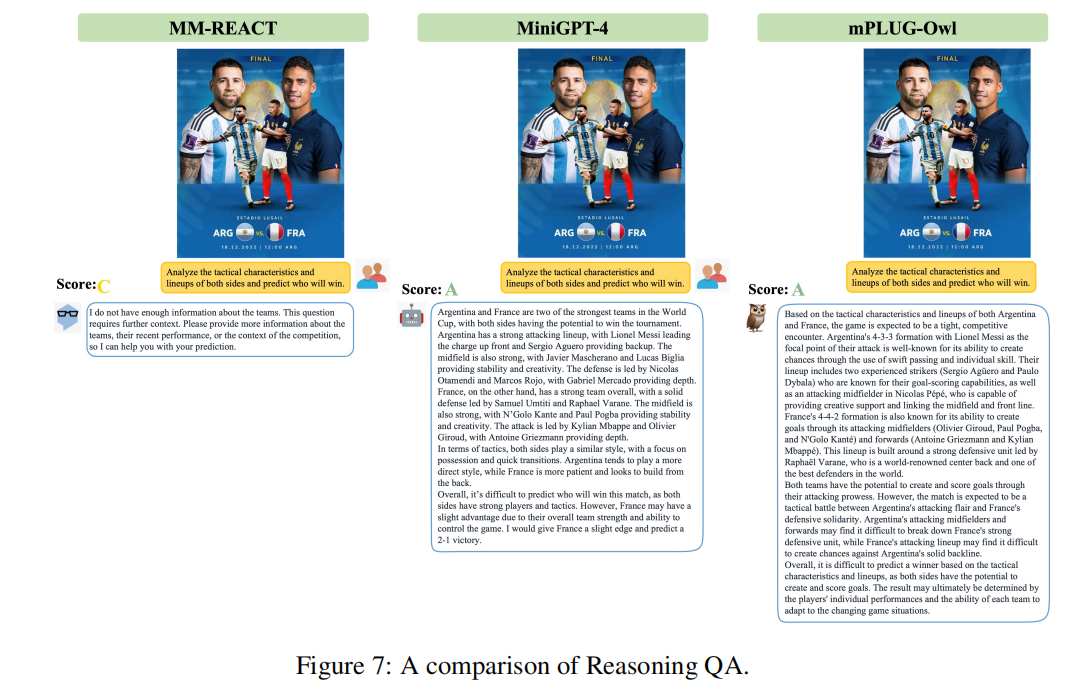
**多回合对话**

图6中的指令要求模型根据参考信息来识别图像的内容。基线模型在面对与空间取向、人类行为和目标属性相关的参考表达式时往往会出错，而mPLUG-Owl提供了最准确的回答。该功能源于mPLUG-Owl对图像的细粒度理解，允许它根据指令中的参考信息来定位图像的相应部分。



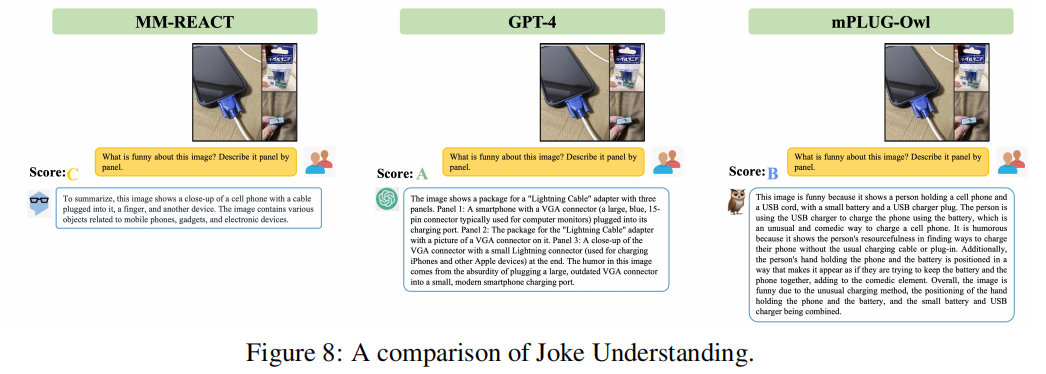
**推理**

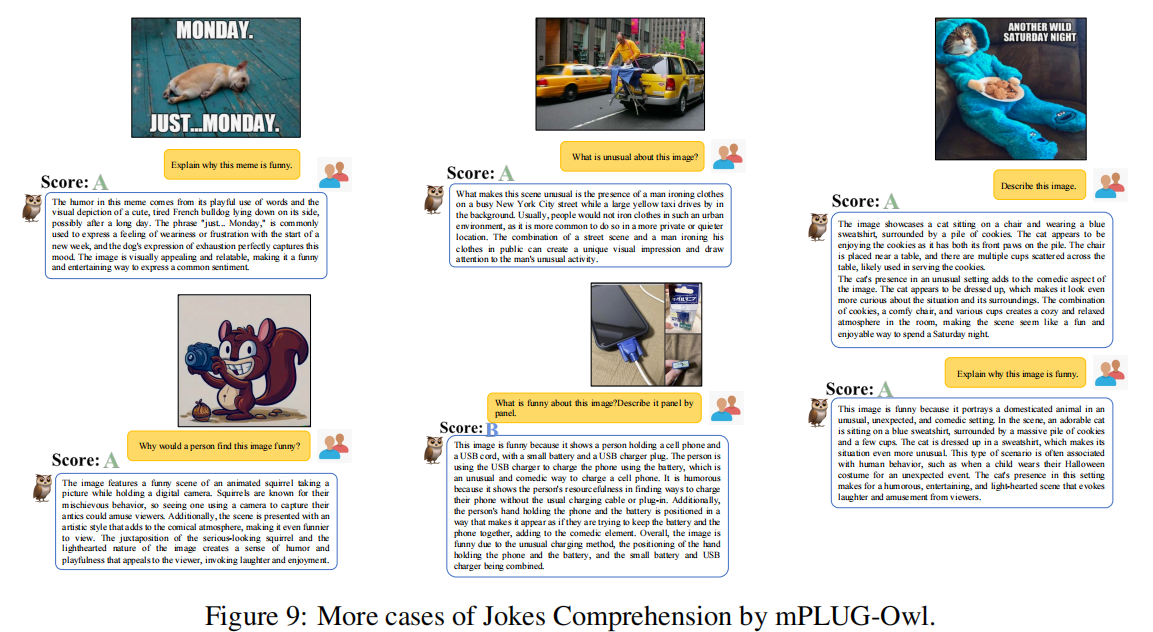
图7显示了一个指令，要求模型基于视觉信息给出预测并解释原因。mPLUG-Owl从阵容和战术方面分析了两个队的特点，并使用它们来解释结果。虽然MiniGPT-4也表现良好，但它在推理方面的说服力略低于mPLUG-Owl。



**笑话理解**

图8中的案例来自于GPT-4，它要求模型来理解和解释一个与视觉相关的笑话。GPT-4不仅进行了循序渐进的说明，而且几乎完全理解充电方法的幽默。猫头鹰也理解这种不寻常的幽默，但它错误地将“VGA”转换为“USB”。这主要是由于训练数据中的视觉信息的限制。图9显示了更多关于笑话理解的案例。



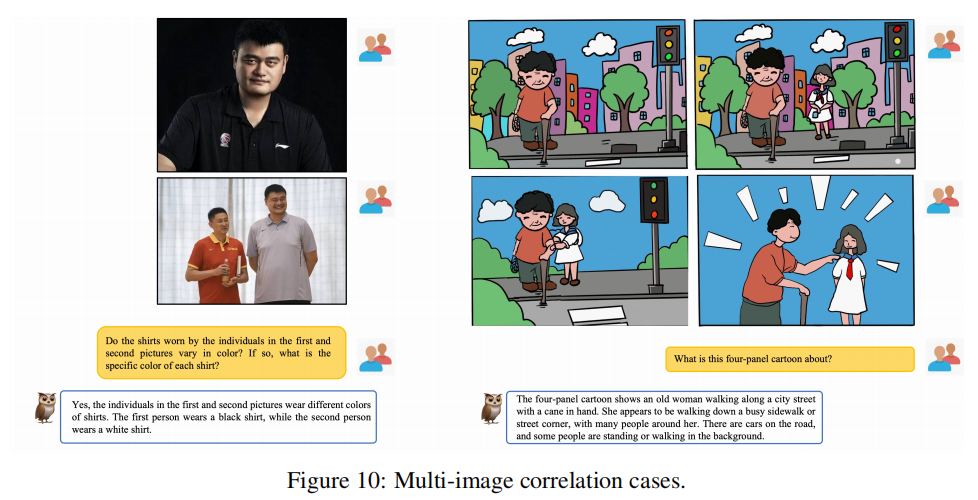


**结果讨论和局限性**

本节展示了一些尚未完全开发的mPLUG-Owl的能力，并讨论了其局限性。

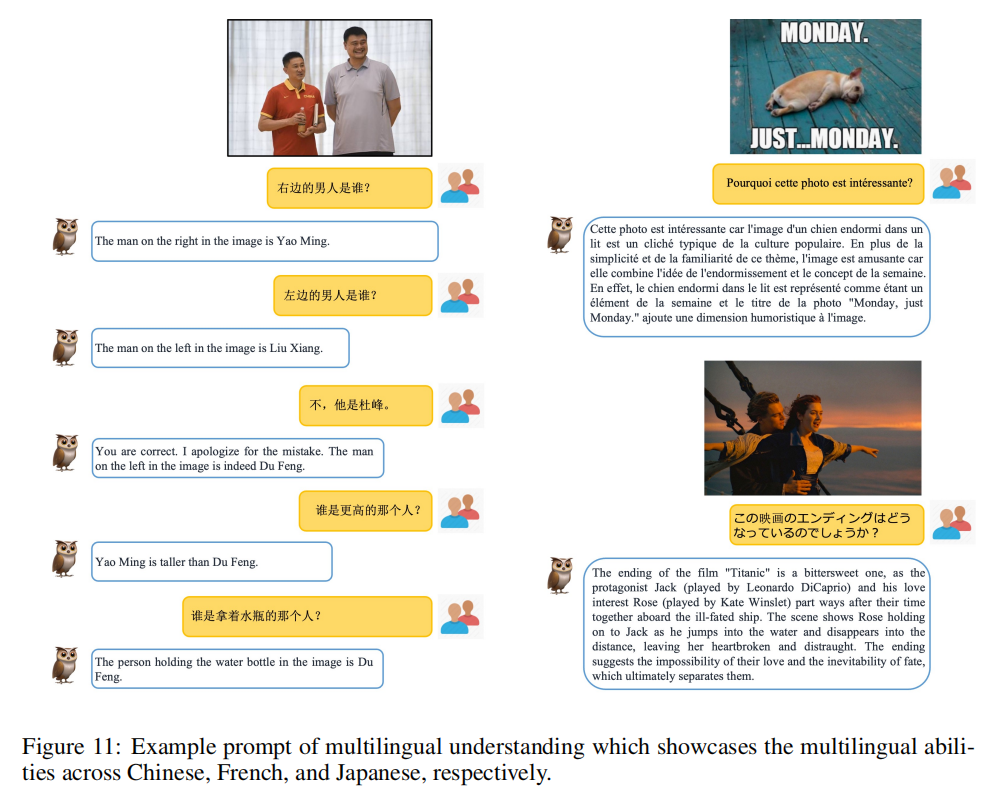
**多图关联**

在图10中，mPLUG-Owl在多个图像上显示了一种涌现的但不强的视觉关联能力。在左边的例子中，模型可以在两幅图像中识别出相同的人，并正确判断布料颜色的差异。但在右边的例子中，该模型没有将4张图像联系起来，并产生了一些文本幻觉。



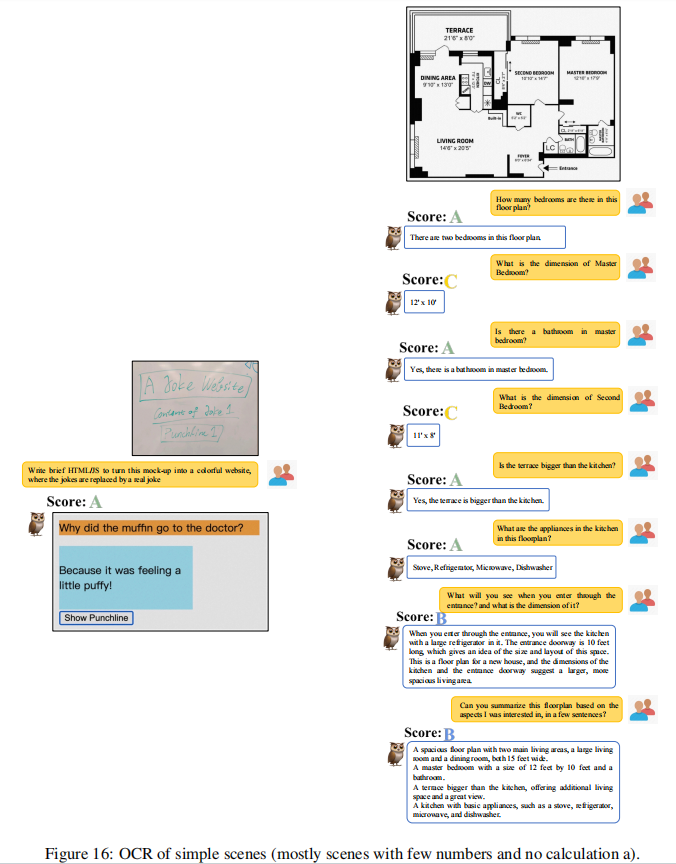
**多语言对话**

除了英语之外，还进一步测试了该模型的多语言能力。如图11所示，虽然在本文两阶段训练中没有多语言数据，但mPLUGOwl显示了对汉语、法语和日语的多语言理解能力。作者将这种能力归因于LLaMa中的原始文本知识。然而，由于缺乏多语言训练，mPLUG-Owl可能无法响应相应的语言。



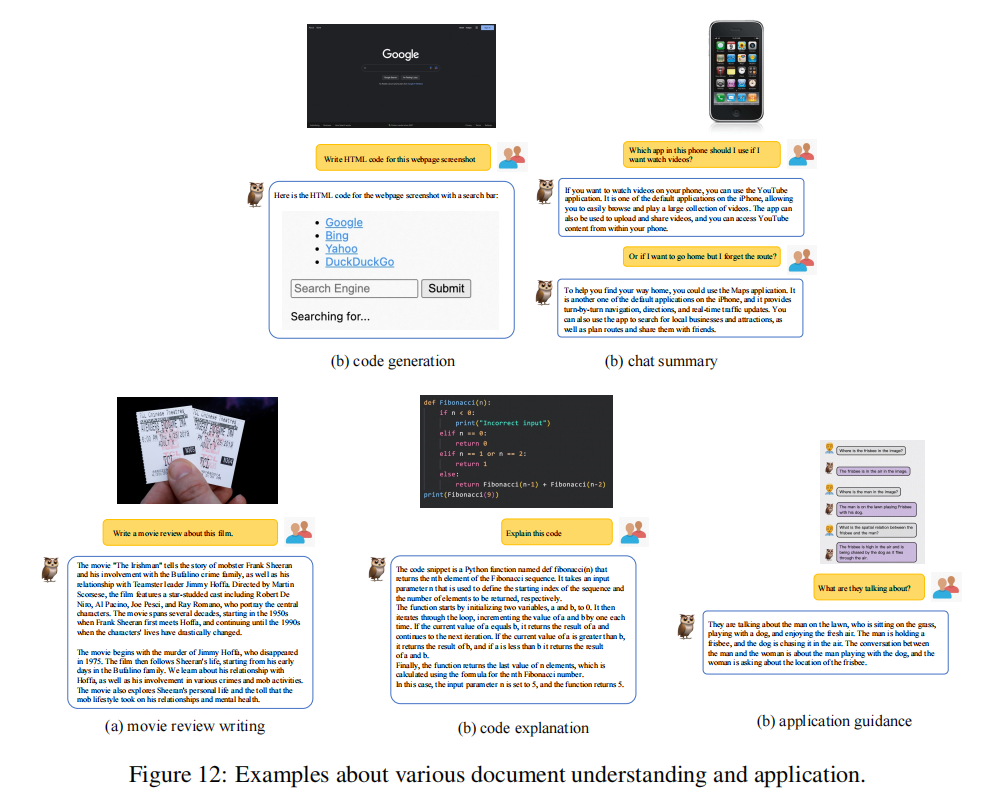
**场景文本理解**

在图16中，mPLUG-Owl在一些简单的场景中展示了它的OCR能力，但模型对图像中数字的感知仍然有限。但是对于复杂场景的OCR，如[图17-18](./论文原文.pdf)所示，mPLUGOwl的性能更为普遍，主要是因为对图像中数字的感知较弱，影响了后续的推理计算。



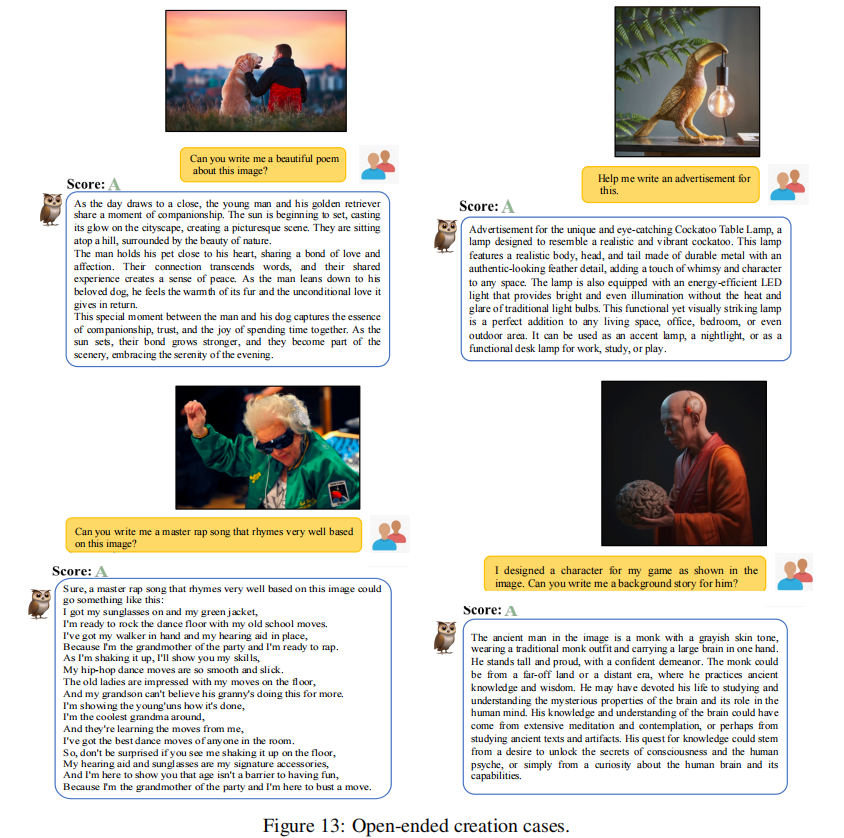
**视觉文档理解**

虽然该工作没有使用任何文档注释数据进行训练，但mPLUG-Owl显示出了一些文本识别和文档理解能力。对此，研究人员进行了更详细的研究。如图12所示，研究人员探讨了电影评论的编写、代码生成、代码解释、聊天摘要和应用程序指导。mPLUG-Owl在(a)和(b)中表现出了良好的性能，但仍存在一些错误。同时，它无法在(d)、(e)和(f)中提供可用的响应。因此，可以进一步探索该模型在文档理解和下游应用方面的潜力。



**开放式创作**

mPLUG-Owl在诗歌、歌词、广告和其他基于图像的作品的创作方面表现出色。它在某些情况下的性能如图13所示。需要进一步的探索来实现更功能性和实用的创作。



**总结**

作者提出了mPLUG-Owl，它增强了LLM的多模态能力。该模型包括基础LLM、一个视觉知识模块和一个视觉摘要模块，这些模块可以支持多种模态，并通过模态协作提升单模态和多模态能力。采用了一种两阶段的训练方法来对齐图像和文本，该方法在LLM的帮助下学习视觉知识，同时保持甚至提高了LLM的生成能力。