# mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration

## 背景

像GPT-3、LLaMA、GPT-4这样的大语言模型（LLM），由于它们在文本理解和泛化上出色的泛化能力，受到大家广泛的关注。为了促进视觉语言任务的应用，GPT-4V在多个任务中展示了令人印象深刻的多模态能力，比如描述、问答任务。这让研究者开始关注于视觉语言领域中的潜在关联，导致了一系列多模态大预言模型的出现。这使得大语言模型具备了理解并处理视觉问题的能力。

但是，目前的方法仅仅关注于提升处理多模态任务的能力，忽略了单模态性能的保证。最近的工作使用了跨模态的对齐模块来将视觉编码器输出的视觉特征映射到预训练的LLM上，从而通过使用已有的语言能力来处理多模态任务。另外，一些研究者在多模态指令微调的时候也顺便把LLM微调了。这虽然提升了处理多模态任务的能力，但是这削弱了其单独处理文本任务的能力。

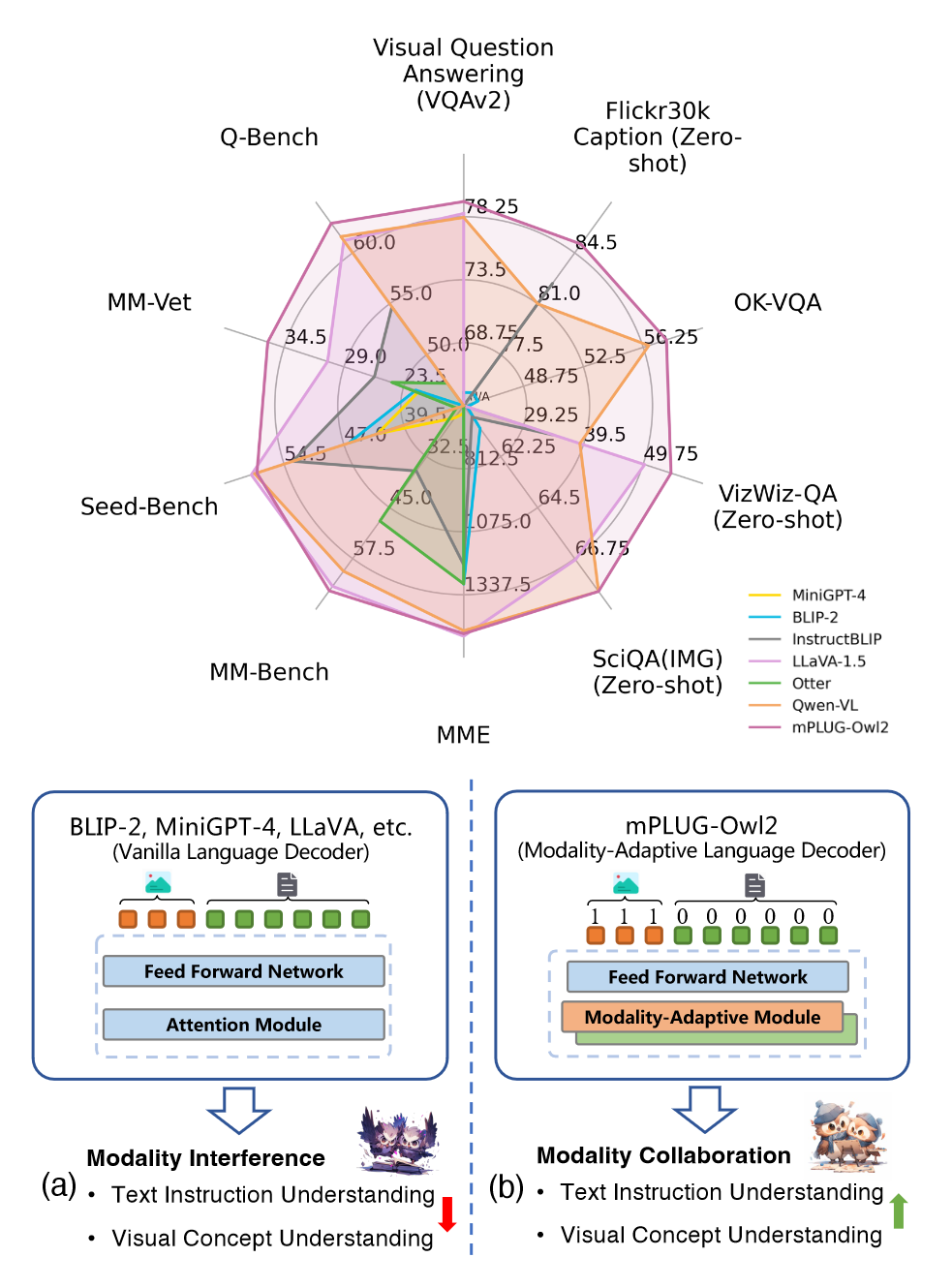


Fig1

目前的主要挑战是：运用一个单独的模块来平衡模态联合（多模态）和模态推理（单模态）的增益，保证模态能在多模态的指导数据集上相互交互。

## mPLUG-Owl2

### 架构介绍

本文将引入mPLUG-Owl2，一个全能的多模态大模型，它有效利用了模态融合来同时提升文本和多模态任务。

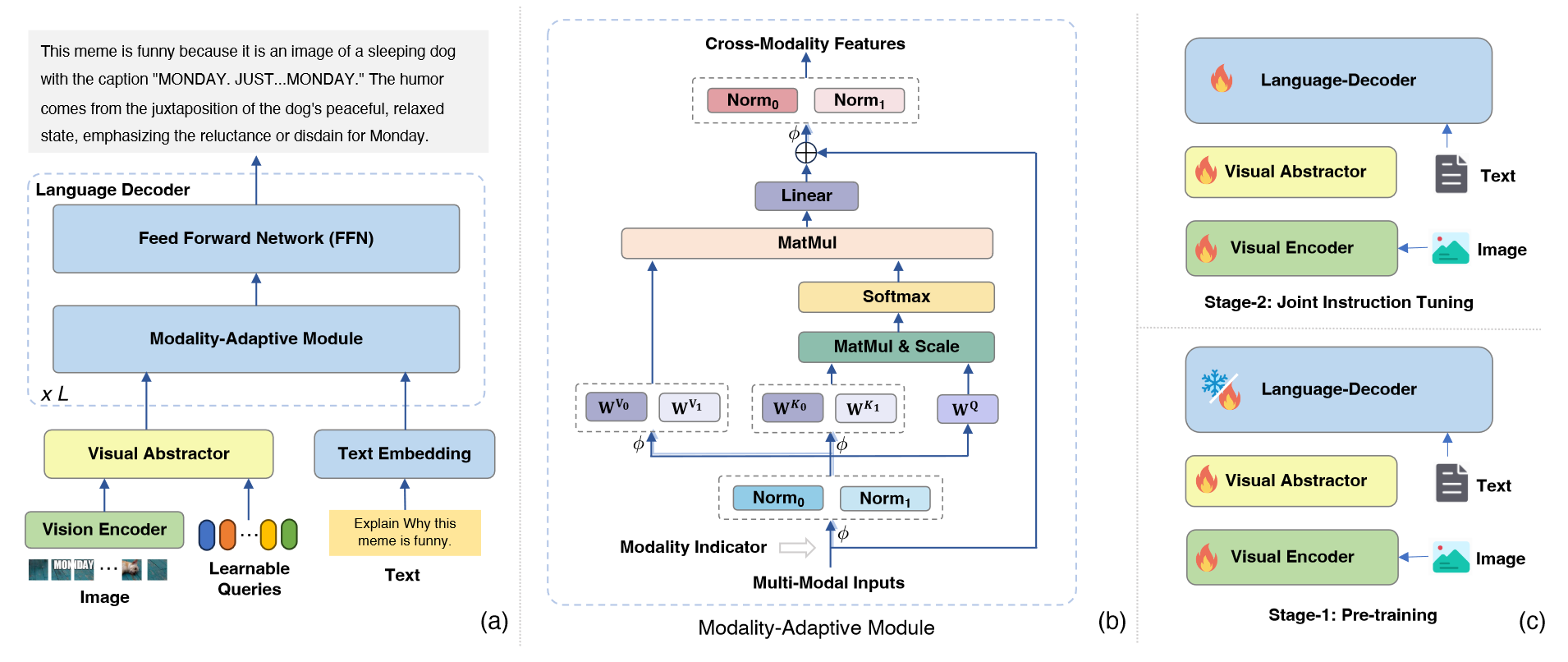


Fig2

mPLUG-Owl2使用一个模块化的网络结构，包括了一个视觉编码器、一个视觉提炼器、一个文本嵌入层、一个语言解码器。具体来说，本文使用ViT-L/14作为视觉编码器，将LLaMA-2-7B作为语言解码器。

* 视觉编码器使用的分辨率处理输入的图谱，并生成大小为的token序列。
* 上述视觉token特征与文本token嵌入合并，喂入语言解码器中。
* 语言解码器被用作一个处理不同模态的通用接口。这样，视觉和语言模态被映射进一个跨模态共享的语义空间上，同时模态特定特征也有所保留。

但是，随着图像分辨率的提升，编码后的视觉token序列会成指数地增长。另外，图像中充分冗余导致了计算上的浪费和不小的噪声。为了解决这个问题，本文提出一个配备有一组固定可学习的请求的**视觉提炼器**来从图像中提取更高的语义特征。

具体来说，本文喂入提取的视觉token序列和固定个的可学习请求进入视觉提炼器。在视觉提炼器中，表示这视觉patch的数量，代表隐藏维度。视觉提炼器包含了一系列视觉提炼层。在第层中，压缩的视觉表示由如下公式进行计算：

其中，表示自注意力函数、和为可学习的参数、为SwiGLU激活函数。本文将来初始化提炼。（注意一下，注意力函数的输出和第一个参数大小一样）

此外，为了增强细粒度感知能力，本文将正弦位置嵌入和视觉特征、进行整合，从而保留文职信息，相关论文论证了这一处理的有效性。

通过上述方法，语言解码器的计算量从降为了，如果就会很大程度减少计算负载，这对那些图片数量很多、文本长度很短的场景很友好。一旦压缩后的视觉特征被获取了，它会与文本token特征合并，通过语言解码器来生成预测。

### 模态适应模块

先前的方法将视觉特征映射到语言语义空间上会导致颗粒度的不匹配（视觉特征含有丰富的语义信息、文本特征只含有零散的语义信息），从而限制了模型的表现。

mPLUG-Owl2引入一个模态适应模块来保留特定模态特征（MAM），通过将视觉特征和语言特征映射到共享的语义空间上，从而分离视觉语言表示。同时，也保留了各个模态的独特特性。

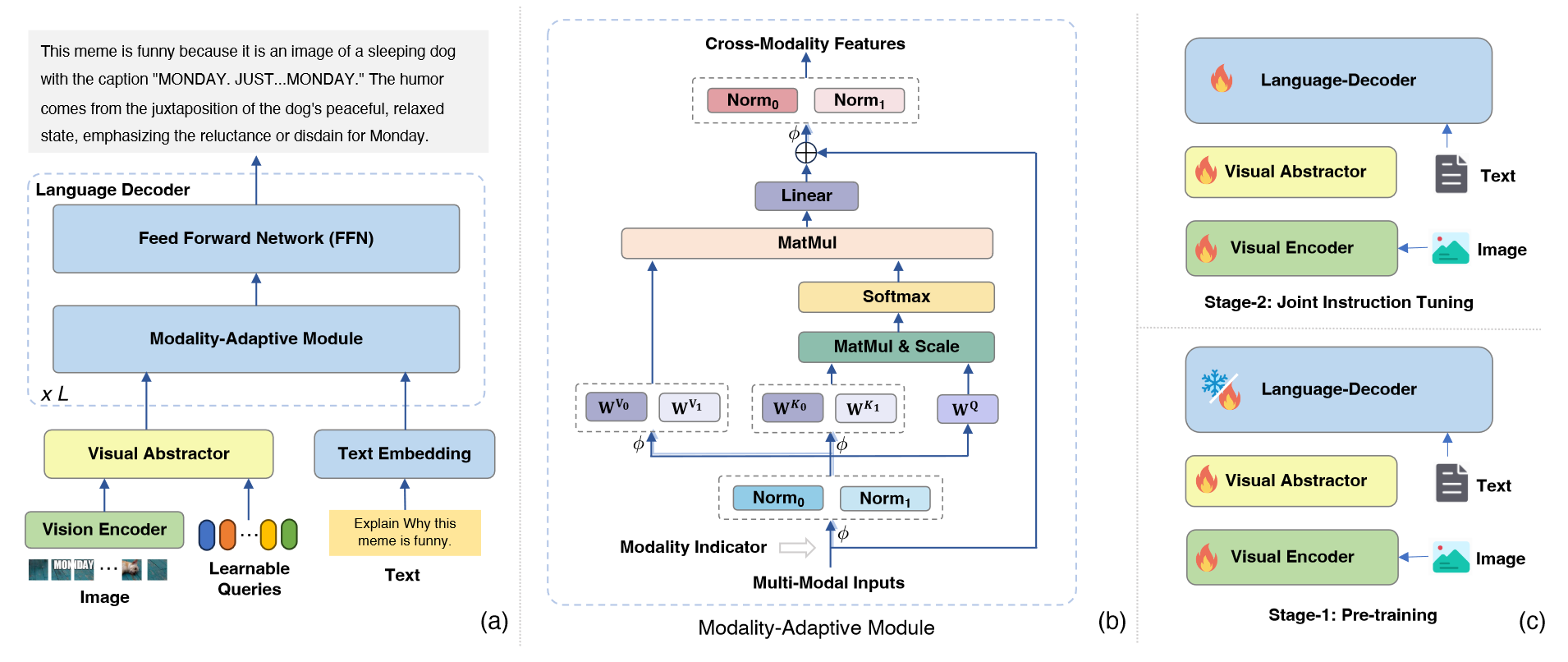


Fig2

具体来说，给定一个视觉语言序列和模态指示器，模态分离函数如下所示：

其中为模态类型。给定先前层的输出向量，其中为语言解码器的层数。本文首先将不同的模态标准化为相同的量级：

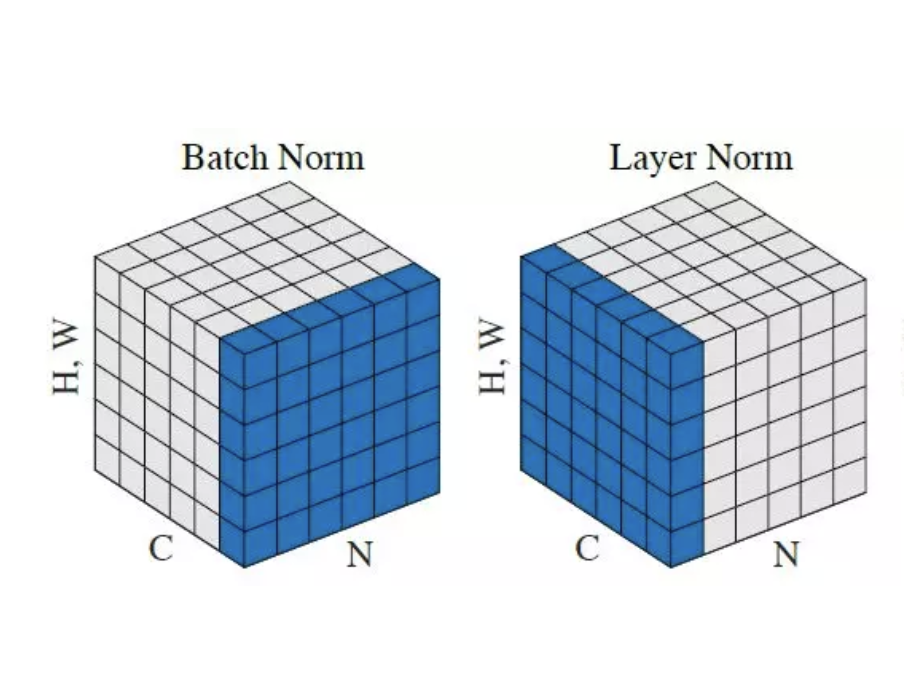


Fig3

其中和分别为视觉特征和语言特征的layer normalization。之后，本文通过对key投射矩阵和value投射举证使用分离线性映射层，构建出自注意力函数。同时，保留了query投射矩阵。公式如下所示：

其中，是可学习的映射矩阵，为第层的上下文特征。按照这种方法，我们可以计算在共享语义空间内的两个模态的相似度，也能通过不同的value映射层保留每个模态的独特模态特征。

此外，通过分离key、value映射矩阵，本文可以避免两个模态间的干扰（尤其是颗粒度不匹配）。同样地，本文也通过利用不同层的标准化层对这些特征进行建模。最后，为了在相同的特征空间内提升模态的联合，本文对所有模态使用一个共享的FFN。本文就是这样通过模态适应模块既实现了模态交互，也保留了模态特征。

### 训练范式

本文引入一个双阶段训练范式，包括了视觉语言预训练和联合视觉语言指令微调。在预训练阶段，本文会将预训练视觉编码器和语言模型进行对齐；在指令微调阶段，本文会用语言建模损失函数微调语言模型。

然而，单纯地冰封预训练视觉编码器并训练一个视觉语言映射器来将视觉数据与语言模型对齐会限制其理解复杂视觉信息（比如场景文本和视觉知识）的能力。因此，本文让视觉编码器在两个阶段一直保持可训练的状态。这一范式通过两个阶段来训练视觉编码器，使其同时有效捕捉低层级和高层级的语义视觉信息。

* 在预训练阶段：预训练语言嵌入层和部分模态适应模块被冰封，其他都保持可训练。
* 先前的研究表明，通过单模态、多模态数据同时进行训练，模型会取得更好的效果。因此本文在指令微调阶段采用一个微调整个模型的联合训练方法。通过将文本和多模态数据作为指令进行微调，可以提高文本中对视觉概念的理解程度、也可以提高复杂自然语言指令的理解程度。

## 实验

实验展示了mPLUG-Owl2能够同时在文本任务和多模态任务泛化，并用一个**单一通用模型**在8个经典的视觉语言benckmark上取得SOTA效果。

### 图像描述、VQA

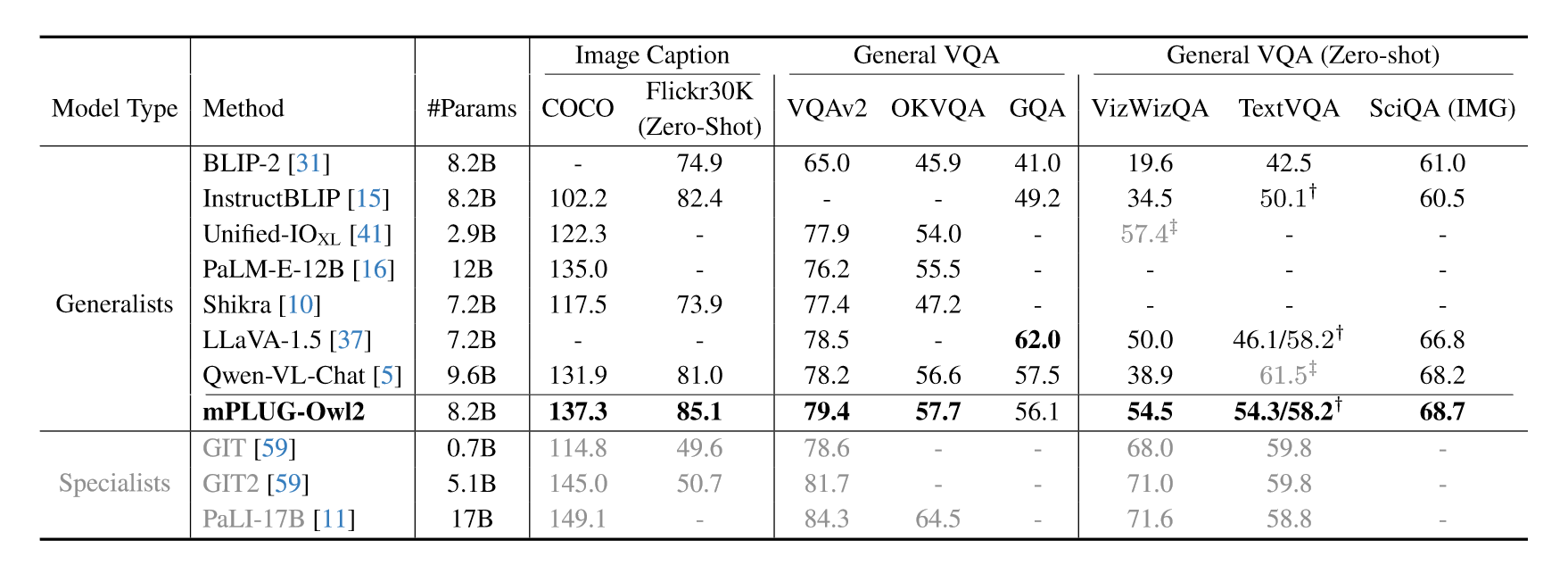


Fig3

（†表示使用了OCR输入、‡表示模型在此数据集上训练过；Specialists表示模型在数据集上微调过）

### 零样本多模态任务

此外，mPLUG-Owl2在5个最近零样本多模态benchmark上获得了数一数二的效果，展示了其在多模态指示理解和生成上的适应性和有效性。

值得注意的是，mPLUG-Owl2是第一个在同时纯文本和多模态场景中展示模态联合现象的多模态大语言模型，为未来多模态基础模型的发展开创了一条道路。

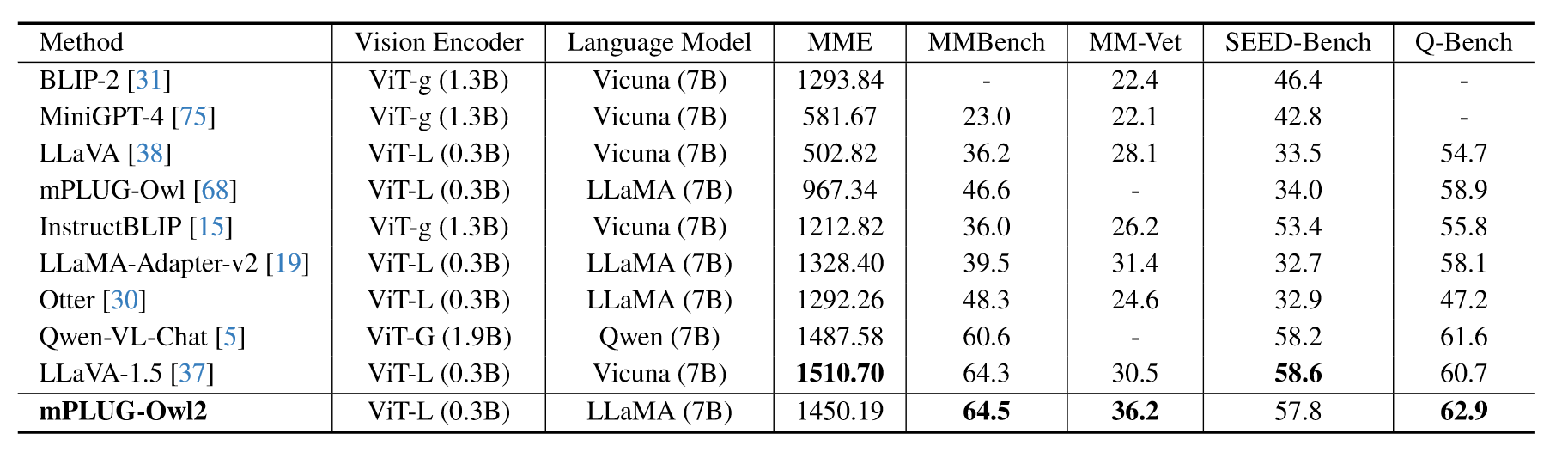


Fig4

### 自然语言理解和生成

还有，mPLUG-Owl2也在多个纯文本benchmark中也取得了SOTA结果。

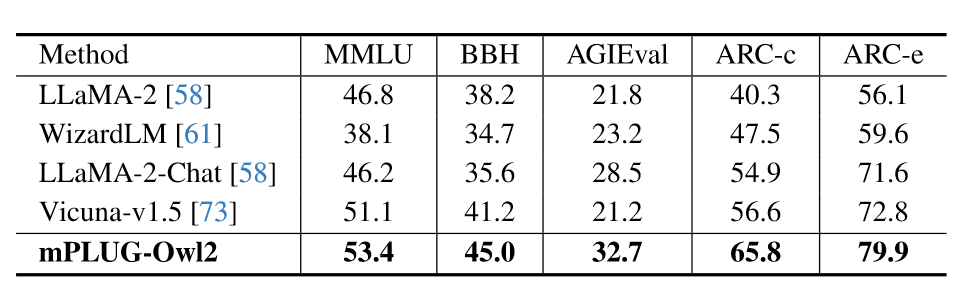


Fig5

### 零样本视频问答

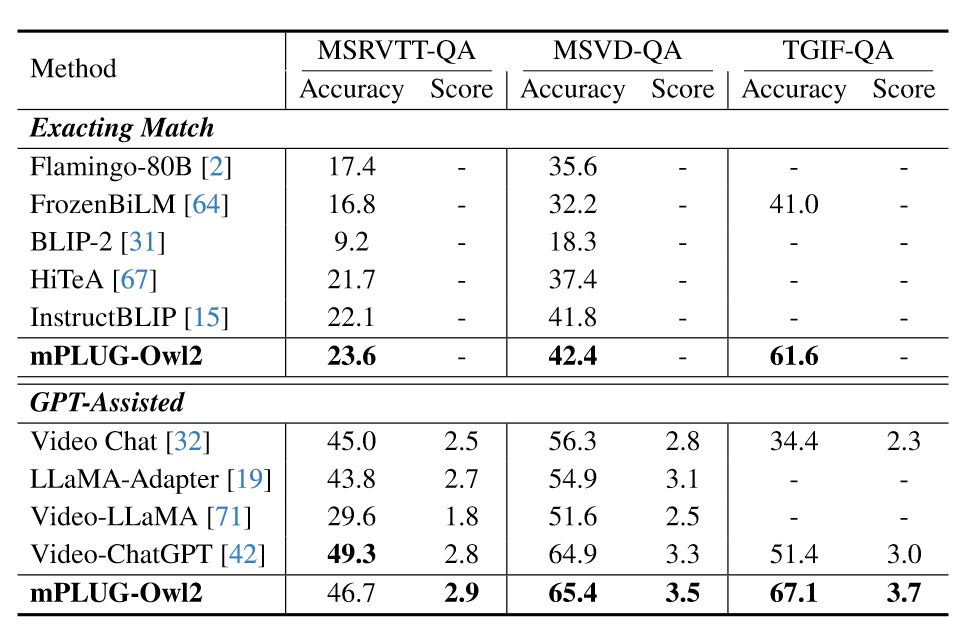


Fig6

两种评估方式：提取匹配（一种经常被使用的评估方法）、GPT协助评估。

### 消融实验

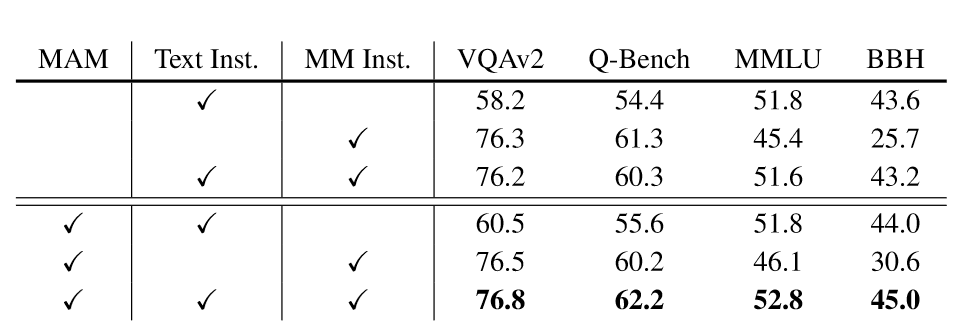


Fig7

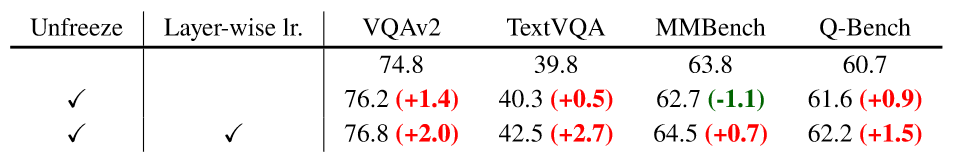


Fig8

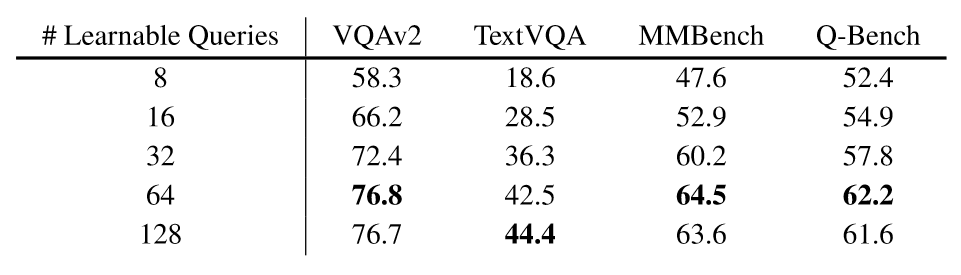


Fig9

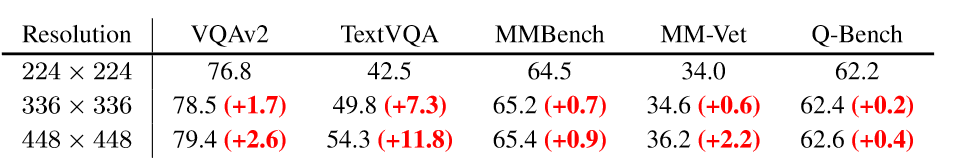


Fig10

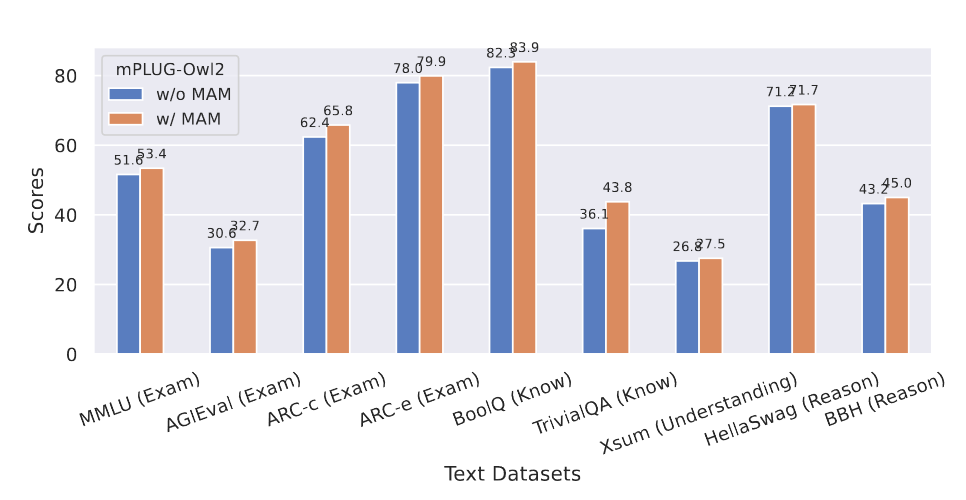


Fig11

1. 模态适应模块在多模态任务和纯文本任务上都具有有效性。
2. 指令微调所用的样本同时考虑多模态和纯文本会更好。
3. 让视觉encoder可训练会提升性能，加入层级学习率会消除遗忘和对视觉表示的损害。
4. 可学习的Query数量在64会比较好。
5. 图像分辨率越高越好。

### 量化分析

另外，本文通过深度分析展示并验证了本文提出的模态适应模块的效果，尤其是在提升包括理解、知识、推理在内的文本任务上。

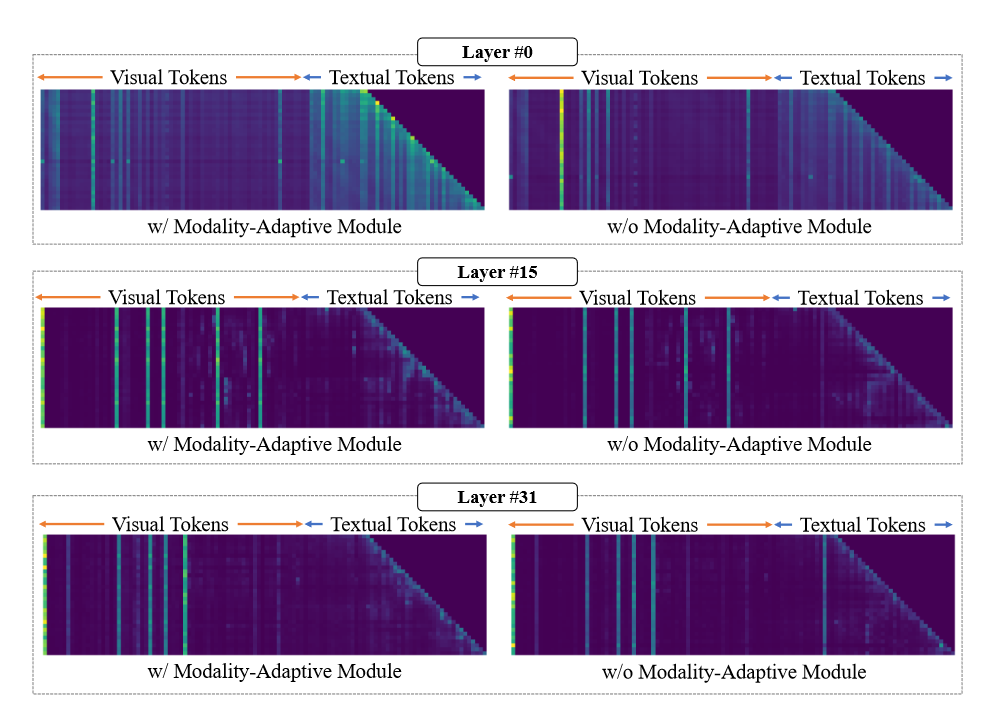


Fig12

在前几层模型更加关注文本token，在后几层更加关注视觉token。直观的解释是，多模态大模型会先通过句法信息理解指令，之后考虑文本输入来识别相关的视觉上下文token。

再加入模态适应模块后，模型在前期对文本的关注度更高、后期对视觉的关注度更高。说明模态适应模块起到了阻止模态被“一视同仁”的情况。

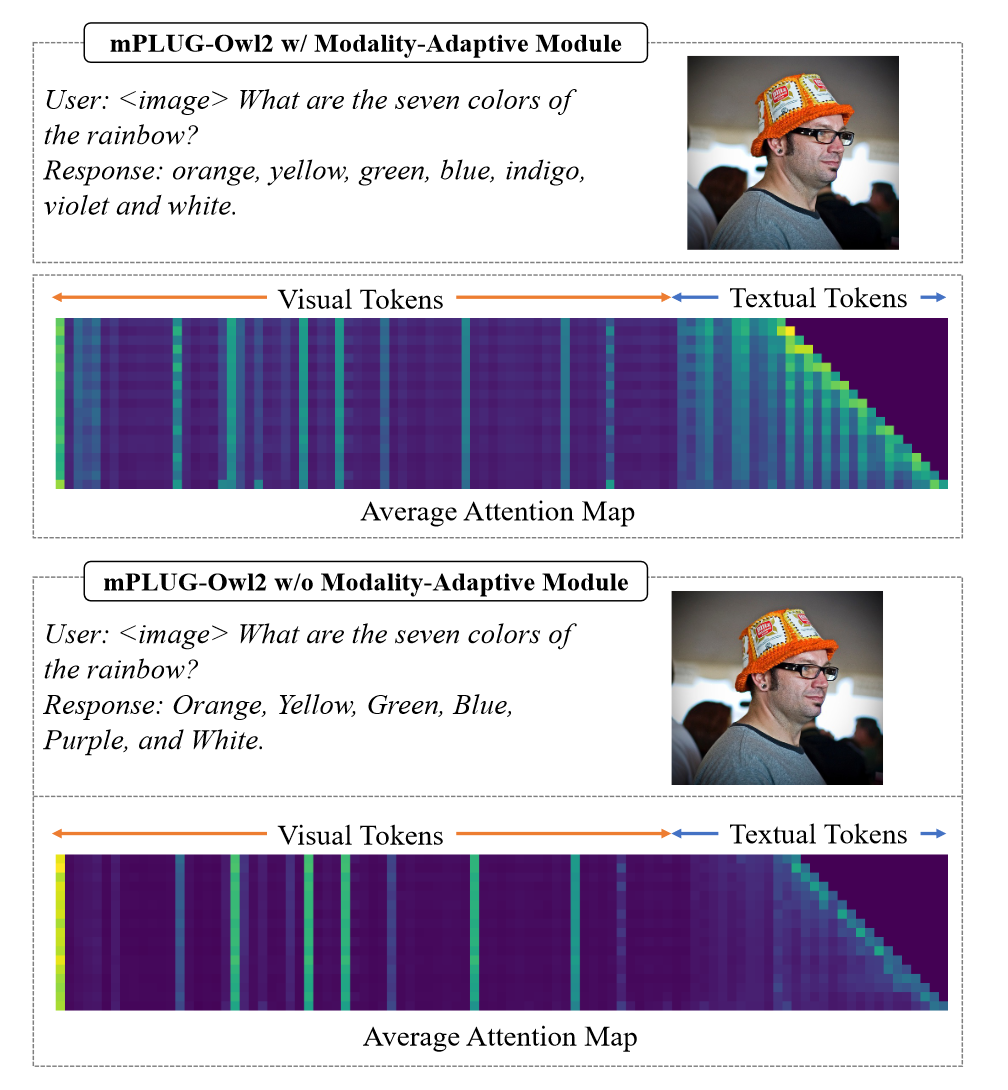


Fig13

当问的问题和图片毫无关系的时候，加入模态适应模块更能将注意力引导向文本，回答更加正确。

## 个人感想

特色：

1. 加入模态适应模块，本质上就是一种模态分离的注意力机制。防止颗粒度不对齐（打破维度、横向突破）。
2. 加入视觉提炼器，提升视觉信息密度、降低计算量。
3. 利用注意力地图量化分析，揭露问题从而方便找到创新点。

创新方向：

* mPLUG-Owl2理解能力强，但是推理能力弱。而LLaVA推理能力很强。有没有合并的可能？
* 接受更多模态，提高通用性，做一个万能大模型（3D点云、图片seq）。
* 能否将分离的思想应用到处理多张照片中，获得图片之间的联系？