arXiv:2302.00402v1 [cs.CV] 1 Feb 2023

mPLUG-2：跨文本、图像和视频的模块化多模态基础模型

Haiyang Xu [*1] **Qinghao Ye** [*1] **Ming Yan** [1] **Yaya Shi** [1] **Jiabo Ye** [1] **Yuanhong Xu** [1] **Chenliang Li** [1] **Bin Bi** [1] **Qi Qian** [1] **Wei Wang** [1] **Guohai Xu** [1] **Ji Zhang** [1] **Songfang Huang** [1] **Fei Huang** [1] **Jingren Zhou** [1]

摘要

近年来，语言、视觉和多模态预训练日益趋于融合。在这项工作中，我们提出了 mPLUG-2，这是一种采用模块化设计的全新多模态预训练统一范式，既能受益于模态协作，又能解决模态纠缠问题。不同于主要依赖序列到序列生成或基于编码器的实例判别的主流范式，mPLUG-2 通过共享通用模块实现模态协作，并将不同模态模块解耦以应对模态纠缠，从而引入了一种多模块组合网络。该模型能够灵活地为文本、图像和视频等所有模态的不同理解与生成任务选择不同模块。实证研究表明，mPLUG-2 在涵盖 30 多项下游任务的广泛任务中取得了最先进或具有竞争力的结果，这些任务包括图像-文本和视频-文本理解与生成等多模态任务，以及纯文本、纯图像和纯视频理解等单模态任务。值得注意的是，在具有挑战性的 MSRVTT 视频问答和视频描述任务中，mPLUG-2 以小得多的模型规模和数据规模，取得了新的最先进成绩：Top-1 准确率达到 48.0，CIDEr 达到 80.3。它在视觉-语言和视频语言任务上也展现出强大的零样本迁移能力。代码和模型将在 https://github.com/alibaba/AliceMind。

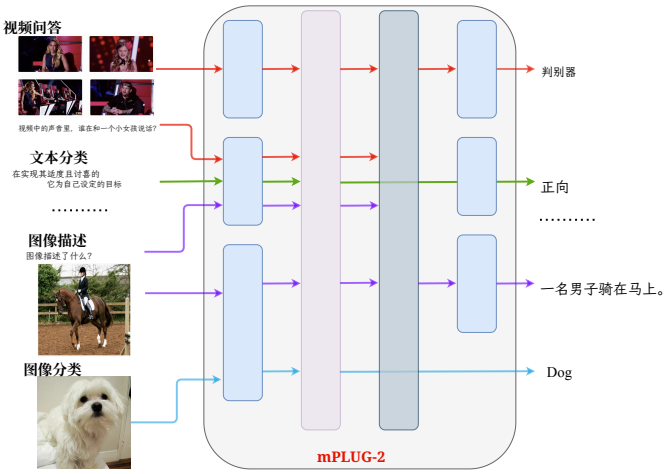

图*1*。 模块化设计构建多模态基础模型这一新范式的简要示意图。

1. 引言

大规模预训练基础模型已成为人工智能（AI）众多领域的新兴范式，涵盖语言领域（Devlin 等人，2018年；Brown 等人，2020年）、视觉领域（Dosovitskiy 等人，2020年；Liu 等人，2021b）以及多模态（Radford 等，2021；Yu 等人，2022；Wang 等人，2022e）。随着 Transformer 架构（Vaswani et al., 2017）广泛应用并取得成功，近年来，语言、视觉与多模态预训练呈现出深度融合的趋势（Yu 等人，2022；Wang 等人，2022e；Alayrac 等人，2022）。这一趋势下的一种思路是采用统一的序列到序列生成框架（如 T5（Raffel et al., 2020）、OFA（Wang 等，2022d）和 Flamingo（Alayrac 等人，2022）），将任务与模态统一起来。另一方面，BERT（Devlin 等人，2018）、Florence（Yuanet al., 2021）和 BEIT-3（Wang 等人，2022e）将所有任务都建模为实例判别，并采用纯粹的基于编码器的架构。

主流基础模型提出为多模态共享同一个单一网络（Alayrac 等人，2022），以利用模态协作带来的信息。然而，由于不同模态任务之间存在较大差异，这种策略会面临模态纠缠问题。挑战在于，多种模态可能会相互干扰（Huang 等人，2022b），尤其是在模态和任务数量较多时。单模块基础模型很难在模态协作带来的收益与模态

[*]共同贡献 [1]阿里巴巴集团达摩院。通信作者：Ming Yan < ym119608@alibaba-inc.com>。

表1. **从多种单模态和多模态下游任务的角度，对 mPLUG-2 与现有基础模型进行系统级比较。** “Cls.”表示分类。“Det.”和“Seg.”分别是“检测”和“分割”任务的简称。“VG”表示视觉定位任务。借助模块化，mPLUG-2 能够同时支持单模态（即计算机视觉和自然语言处理）与多模态（即图像-文本和视频-文本）下游任务。

	计算机视觉				自然语言处理			图像 -Text				视频 -Text		
方法	图像分类	视频分类	Det.	Seg.	文本分类	QA	摘要	检索	QA	图像描述	VG	检索	QA	图像描述
BEIT-3	✓		✓	✓				✓	✓	✓				
EVA	✓	✓	✓	✓										
CLIP	✓							✓				✓		
ALBEF								✓	✓		✓			
BLIP								✓	✓	✓		✓	✓	
VATT	✓	✓										✓		
Florence	✓	✓	✓					✓	✓			✓		
CoCa	✓	✓						✓	✓	✓		✓		
VideoCoCa		✓												
Flamingo		✓						✓	✓	✓			✓	✓
GIT2	✓				✓				✓	✓			✓	✓
FLAVA	✓				✓	✓	✓		✓					
OFA	✓				✓	✓	✓	✓	✓	✓	✓			
OmniVL	✓	✓						✓	✓	✓		✓	✓	✓
mPLUG 2.0	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

纠缠对跨多种模态的大量下游任务造成的影响之间取得平衡。

为缓解这一挑战，本文提出了一种多模态基础模型的新型统一范式，如图所示1。该范式采用基于模块的网络设计，同时考虑模态协作与模态纠缠。mPLUG-2 设计了若干共享功能模块以促进模态协作，同时保留模态专属模块来应对模态纠缠问题。随后，根据任务的模块设计，在单模态和多模态数据集上对不同模块进行有效的联合训练。因此，可灵活选择并组合不同模块，以适应大量单模态及跨模态理解与生成任务。所支持的下游任务详情见表1。据我们所知，本文提出的方法涵盖了文本、图像和视频领域中种类最多的下游任务。

具体而言，我们通过解耦空间表示和时间表示，设计了一个统一的双视觉编码器模块：视频输入与图像输入共享标准 Transformer 模块，以建模空间信息；此外，还使用一个额外的局部时序建模模块，在视频相关任务中进行时间关系建模。随后，我们引入一种新颖的通用层模块，作为不同模态之间的枢纽，通过共享自注意力模块，将视觉和语言模态投射到由语言引导的共同语义空间。此外，我们还使用一个额外的交叉注意力模块，将通用视觉表征与原始细粒度视觉表征融合。模块的详细设计如图2所示。最后，mPLUG-2 的不同模块在单模态和跨模态任务上，结合任务和模态指令进行联合预训练（Wang 等人，2022d）。在推理阶段，mPLUG-2 可以使用不同模块来处理各种单模态和跨模态任务，并采用模块化 Transformer 架构。不同任务所选用的模块见附录中的表2 。

我们在 30 多个具有挑战性的单模态和跨模态理解与生成基准测试上评估了 mPLUG-2 的全新统一范式。在模型规模和数据规模相近的情况下，它取得了最先进水平或具有竞争力的结果。凭借基于模块的网络设计，mPLUG-2 还可以通过选择和添加模块轻松扩展到其他任务。值得注意的是，在具有挑战性的 MSRVTT 视频问答和视频描述任务上，mPLUG-2 分别取得了 48.0 的 Top-1 准确率和 80.3 的 CIDEr，创造了新的最先进结果。mPLUG-2 在视觉语言和视频语言任务上也展现出强大的零样本迁移能力。

2. 相关工作

纯视觉基础模型：在视觉 Transformer（又称 ViT）（Dosovitskiy 等，2020）出现之前，卷积网络（Szegedy 等，2017；He 等，2015）长期以来一直是主流的视觉架构。凭借 Transformer 网络的卓越能力，ViT 在各种下游任务中脱颖而出（Carion 等，2020；Xu 等，2022）。除了使用 JFT-3B（Zhai 等，2021）等大规模数据集扩展朴素的 ViT 架构外，SwinV2-G（Liu 等，2021a）还通过分层架构扩展了原始 ViT。此外，EVA（Fang 等，2022a）利用大规模预训练图文模型（例如 CLIP（Radford 等，2021）），从未标注图像中蒸馏多模态知识，以扩展 ViT。近期，InternImage（Wang 等，2022f）通过可变形卷积让卷积神经网络焕发生机，并在各种视觉下游任务上达到最先进水平。此外，InternVideo（Wang 等，2022g）通过整合两个兼具生成式和判别式能力的大型视频模型，将应用扩展到视频任务

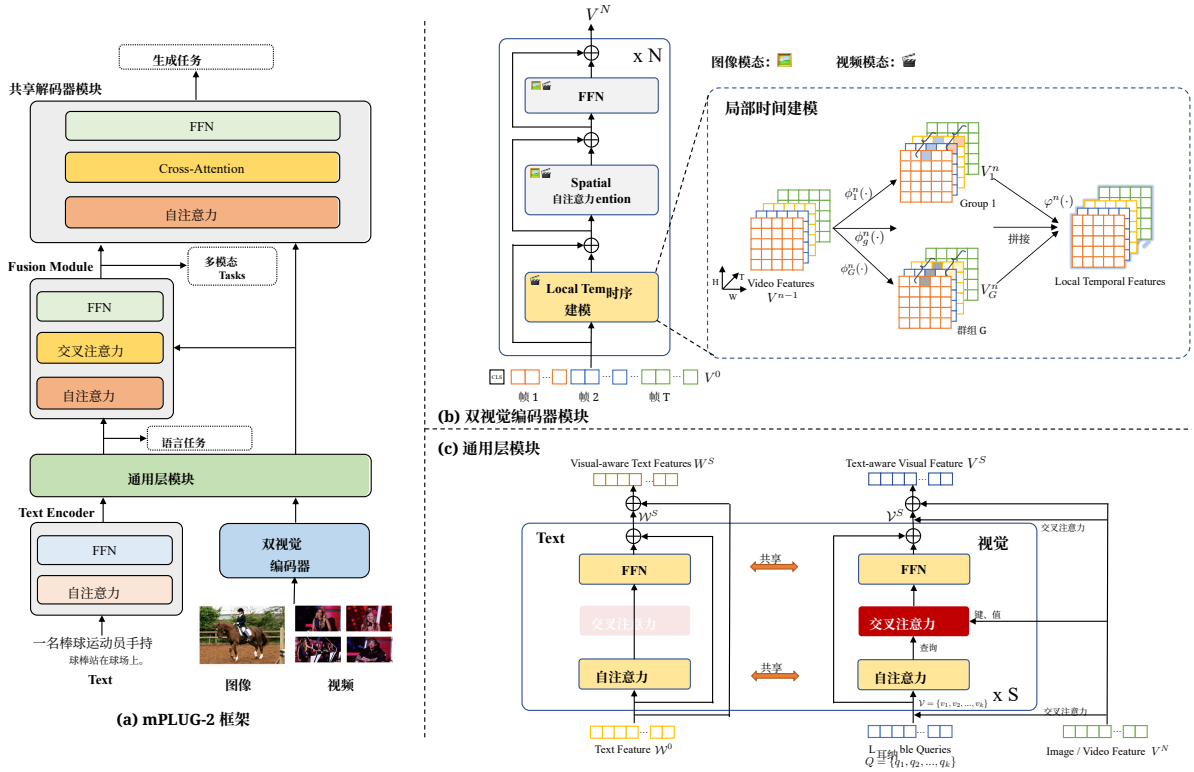

图2。 mPLUG-2 的整体框架和模块细节。

自监督视频学习。

纯语言基础模型 受 BERT 在自然语言理解领域成功实践的启发（Devlin 等人，2018年），研究者提出了大规模语言基础模型，用于自然语言处理。BART（Lewis 等人，2020年）是一种类似 BERT 的去噪自编码器，但采用编码器-解码器架构，在文本生成和理解任务中均表现出良好效果。除了 BERT系列方法（Devlin 等人，2018年；Lewis 等人，2020年；Liu 等人，2019年）之外，还有许多其他有效的架构和预训练目标。T5（Raffel 等人，2020年）提出了一个统一框架，将所有基于文本的语言任务转化为文本到文本格式。GPT-3（Brown 等人，2020年）是一种自回归语言基础模型，包含1750亿个参数，在少样本和零样本设置下的许多自然语言处理任务中表现出色。

视觉-语言基础模型 得益于互联网上大量的图像/视频-文本对，视觉-语言基础模型的出现能够涵盖视觉-语言预训练。CLIP（Radford 等人，2021）和 ALIGN（Jia 等人，2021）的成功表明，在噪声图文对上采用简单对比目标进行预训练的模型能够生成强大的视觉-语言表征。此外，ALBEF（Li 等人，2021b）、BLIP（Li 等人，2022c）和 mPLUG（Li 等人，2022a）通过引入多模态文本补全和文本生成任务来辅助学习。另一方面，一些基础模型则通过任务统一构建。例如，Florence（Yuan 等人，2021）统一了能够同时利用视觉数据和视觉-语言数据的对比目标。BEiT-3（Wang 等人，2022e）将预训练任务归结为文本、视觉和视觉-语言领域的掩码数据建模。SimVLM（Wang 等人，2021b）、OFA（Wang 等人，2022d）和 CoCa（Yu 等人，2022）通过生成式预训练实现视觉-语言理解与生成。与主流基础模型不同，mPLUG-2 引入了一种新型模块化Transformer框架，通过共享通用模块并解耦模态专属模块，灵活组合不同模块以处理单模态和跨模态任务，从而解决模态纠缠问题。

3. 方法

3.1. 整体框架

如图所示2，mPLUG-2 由用于图像和视频的双视觉编码器模块、文本编码器模块、作为所有任务共享的多模态枢纽的通用层模块、多模态融合模块，以及用于单模态和跨模态生成的共享解码器模块组成。我们首先使用两个单模态编码器，分别对图像/视频和文本进行编码，以表示各个模态的固有信息。对于图像/视频，我们采用双视觉编码器，通过空间建模和局部时间建模来编码视觉特征。然后，

表2. 下游任务的模块。

	输入			模块							
任务	Text	图像	视频	文本编码器	图像编码器	视频编码器	通用层	融合层	文本解码	图像解码	视频解码
视频-文本检索	✓		✓	✓		✓	✓	✓			
视频-文本问答	✓		✓	✓		✓	✓	✓			✓
视频-文本描述生成	✓		✓	✓		✓	✓	✓			✓
图像-文本检索	✓	✓		✓	✓		✓	✓			
图像-文本问答	✓	✓		✓	✓		✓	✓		✓	
图像-文本描述生成	✓	✓		✓	✓		✓	✓		✓	
视觉定位	✓	✓		✓	✓		✓	✓		✓	
视频分类			✓			✓	✓				
图像分类		✓			✓		✓				
图像检测		✓			✓		✓				
图像分割		✓			✓		✓				
文本分类	✓			✓			✓				
文本问答	✓			✓			✓				
文本摘要	✓			✓			✓		✓		

视觉和语言表示分别输入通用模块，该模块由多个通用层组成。每个通用层将不同模态投影到共享语义空间以进行跨模态对齐，同时保留不同模态的原始表示。通用层的输出用于执行单模态判别任务。对于跨模态任务，则会应用额外的融合模块来生成跨模态表示。最后，单模态和跨模态表示可作为输入送入共享的 Transformer 解码器，以执行各种生成任务，从而促进多任务预训练和迁移学习。不同下游任务所用的模块总结见表2。

双视觉编码器模块 为了捕捉各种视觉模态的视觉信息，我们提出了双视觉编码器，以同时对图像和视频进行建模。具体而言，我们将图像和视频帧拆分为一系列 L 互不重叠的视觉令牌。每个视觉令牌序列都带有可学习的空间位置嵌入和一个额外的 [CLS] 令牌，共同构成一个输入视觉序列。然而，在没有大规模视频预训练的情况下，对完整视觉序列进行建模会导致时空学习面临困难（Li 等人，2022d；Wang 等人，2022a；b）。为缓解这一问题，我们引入时间局部性，将视觉表示解耦为独立的空间表示和时间表示。如图2(b)所示，我们利用 Transformer块中的自注意力（SA）层和前馈神经网络（FFN）进行空间建模，并提出一种新颖的局部时序建模模块（LT），用于对空间表示之间的时间依赖关系建模，如下所示：

$$V_{LT}^{n} = LN(LT(V^{n-1}) + V^{n-1}), \tag{1}$$
$$V_{SA}^{n} = LN(SA(V_{LT}^{n-1}) + V_{LT}^{n-1}), \tag{2}$$
$$V^{n} = LN(FFN(V_{SA}^{n}) + V_{SA}^{n}), \tag{3}$$

其中，LN 是层归一化的缩写。局部时序建模模块通过多组融合，捕捉空间位置相同的图像块之间的相关性。公式化为：

$$V_{g}^{n} = ReLU(A_{g}^{n}\phi_{g}^{n}(V^{n-1})) \in \mathbb{R}^{T\times\frac{C}{G}} \tag{4}$$
$$LT(V^{n-1}) = \varphi^{n}(Concat[V_{1}^{n};\cdots;V_{G}^{n}]), \tag{5}$$

其中，$\phi_{g}^{n}(\cdot)$ 和 $\varphi^{n}(\cdot)$ 是线性变换函数。A_{g}^{n} 是可学习的时间关系参数，其实例化形式为卷积核。T 和 C 分别表示帧数和隐状态的大小。G 表示组数，$Concat$ 表示拼接函数。通过采用多组融合，模型能够从不同时间位置的不同表征子空间中学习丰富的时序信息。因此，除局部时序模块外，双视觉编码器模块支持图像和视频之间的权重共享，从而有效且高效地学习空间和时间表示。

文本编码器模块 对于文本编码器模块，我们使用 BERT（Devlin 等人，2018年）作为文本编码器，将输入文本和一个额外的 [CLS] 标记转换为一系列文本嵌入。CLS]标记的嵌入用于概括输入文本。

通用层模块 为了发挥模态协作的优势，我们提出通用层，在共享语义空间中对视觉和语言模态进行建模，同时保留不同模态的原始表示。

在通用模块之前，我们将双视觉编码器输出的可变数量的图像或视频特征 V^{N} 作为输入，生成固定数量的视觉令牌 k $\mathcal{V}=\{v_1, v_2, ..., v_k\}$，以降低通用层的计算复杂度。在 i_{th} 通用层中，视觉令牌$\mathcal{V}^{i-1}$ 和文本表示 $\mathcal{W}^{i-1}$ 被送入共享的自注意力层以对齐语义；随后，视觉令牌通过交叉注意力层注入原始视觉特征空间，以保留原始表示。

$$\mathcal{V}_{SA}^{i} = LN(SA(\mathcal{V}^{i-1}) + \mathcal{V}^{i-1}) \tag{6}$$

$$\mathcal{W}_{SA}^{i} = LN(SA(\mathcal{W}^{i-1}) + \mathcal{W}^{i-1}) \quad (7)$$

$$\mathcal{V}_{CA}^{i} = LN(CA(\mathcal{V}_{SA}^{i}, V^{n}) + \mathcal{V}_{SA}^{i}) \quad (8)$$

$$\mathcal{V}^{i} = LN(FFN(\mathcal{V}_{CA}^{i}) + \mathcal{V}_{CA}^{i}) \quad (9)$$

$$\mathcal{W}^{i} = LN(FFN(\mathcal{W}_{SA}^{i}) + \mathcal{W}_{SA}^{i}) \quad (10)$$

随后，$[\mathcal{V}^{i}; \mathcal{W}^{i}]$ 被反复送入下一个通用层，以获得最终的通用图像和文本表示。最后，通用层 $[\mathcal{V}^{S}; \mathcal{W}^{S}]$ 的输出通过交叉注意力层与原始表示 $[V^{N}; W^{M}]$ 相结合，从而得到文本感知的视觉表示和视觉感知的文本表示，其中 S，N，M 分别是通用模块、双视觉编码器和文本编码器的层。

融合模块 为了有效捕捉视觉和语言模态之间的跨模态交互，我们采用了 ALBEF（Li 等人，2021b）中的融合模块，该模块由一组带有交叉注意力层的 Transformer 块组成。具体而言，融合模块将通用层模块中的文本嵌入作为输入。随后，文本感知的视觉嵌入会在语言共享公共空间中与视觉感知的文本嵌入进行交叉注意力计算。通过级联带有交叉注意力层的 Transformer 块，融合模块能够生成多模态视觉-语言表示。

共享解码器模块 为了赋予模型生成能力，我们引入了共享解码器模块，使模型能够利用单模态和多模态信息生成文本。具体而言，共享解码器模块是一个可接收任意输入的 Transformer 解码器。例如，图像描述仅需视觉特征，而视觉问答则使用多模态特征。通过接收不同类型的输入，我们的共享解码器模块能够适应各种文本生成任务。共享解码器模块有助于多任务预训练和迁移学习。

3.2. 统一预训练目标

我们采用以下三个目标对 mPLUG-2 的多个模块进行联合训练。

语言损失 对于文本编码器模块，我们采用 BERT（Devlin 等人，2018年）中的掩码语言建模（MLM）来学习文本表示。我们随机遮蔽文本中 15% 的词元，并要求模型根据上下文表示预测这些被遮蔽的词元。

多模态损失 对于跨模态模块，我们采用 ALBEF（Li 等人，2021b）中的跨模态匹配损失（CML），其由视觉-语言匹配（VLM）和视觉-语言对比学习（VLC）组成。

基于指令的语言模型损失遵循 Flamingo（Alayrac 等人，2022）和 OFA（Wang 等人，2022d），我们采用基于指令的语言模型损失来统一各种生成任务。我们使用人工编写的指令来区分任务和模态，包括视频/图像-文本对、视频/图像描述生成、视频/图像问答、文本生成等。

4. 实验

4.1. 训练设置

预训练数据集 遵循以往的工作（Li 等人，2021b；2022a），我们使用与以往相同的热门图像-文本数据集对模型进行预训练，这些数据集包含 1400万张图像，包括 MS COCO（Lin等，2014）、Visual Genome（Krishna 等人，2017年）、Conceptual Captions 3M（Sharma 等人，2018年）、Conceptual Captions 12M（Changpinyo 等人，2021）和 SBU Captions（Ordonez 等人，2011年）。对于视频-文本数据集，我们采用网络来源的视频数据集 WebVid-2M（Bain 等人，2021a），其中包含 250 万个视频-文本对。文本数据集由 WikiCorpus（Devlin 等人，2018年）（约 20GB）和清理后的 Common Crawl（约 350GB）组成。后者的收集和清理方法与 C4（Raffel 等人，2020）所采用的方法基本相同。预训练的实现细节见附录。

4.2. 主要结果

我们在 30 多个基准测试上评估 mPLUG-2 的新型统一范式，包括视觉语言任务（例如多模态检索、问答和图像描述）（Xu 等人，2016年；2017年；Chen 和 Dolan，2011年）、仅语言任务（例如文本分类、问答和摘要）（Wang 等人，2018年；Rush 等人，2015a），以及纯视觉任务（例如图像分类和视频动作识别）（Deng 等人，2009；Kay 等人，2017年）。具体而言，视觉语言基准测试可分为图像-文本部分和视频文本部分。有关这些数据集的详情，请参见附录。

4.2.1. 多模态任务

文本到视频检索 我们在 MSRVTT（Xu 等人，2016年）、DiDeMo（Anne Hendricks 等人，2017年）和 LSMDC（Rohrbach 等人，2015年）数据集上，将 mPLUG-2 与多种最先进方法进行了比较。结果汇总于表 3。可以看到，mPLUG-2 在大多数数据集上的表现均优于此前最先进的方法。尤其是，与 HiTeA 相比，我们的方法在 LSMDC 数据集上的 R@1 提升了 5.7%，这表明所提出的模型能够通过所提出的方法，利用丰富电影片段中呈现的时间信息

*Ta*表 *3.* **文本到视频检索性能对比**。所有结果均以 R@1/R@5/R@1 报告 0.

方法	#预训练数据	MSRVTT			DiDeMo			LSMDC		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen（Bain 等人，2021b）	5M	31.0	59.5	70.5	31.0	59.8	72.4	15.0	30.8	39.8
BridgeFormer（Ge 等人，2022）	5M	37.6	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5
Singularity（Lei 等人，2022）	5M	36.8	65.9	75.5	47.4	75.2	84.0	-	-	-
LAVENDER（Li 等人，2022d）	30M	37.8	63.8	75.0	47.4	74.7	82.4	22.2	43.8	53.5
All-in-one（Wang 等人，2022a）	283M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
OmniVL（Wang 等人，2022b）	18M	47.8	74.2	83.8	52.4	79.5	85.4	-	-	-
HiTeA（Ye 等人，2022）	17M	46.8	71.2	81.9	**56.5**	**81.7**	**89.7**	28.7	50.3	59.0
mPLUG-2_{Base}	17M	48.3	75.0	83.2	52.3	80.8	87.5	25.5	45.8	55.8
mPLUG-2	17M	**53.1**	**77.6**	**84.7**	56.4	79.1	85.2	**34.4**	**55.2**	**65.1**

表 *4*。**视频问答性能对比**。评估采用准确率指标。mPLUG-2 通过开放词汇生成，在 MSRVTT-问答和 TGIF-FrameQA 上取得了视频问答的最先进水平。

方法	#预训练数据	MSRVTT-QA	MSVD-QA	TGIF-FrameQA
JustAsk (Yang 等人, 2021a)	69M	41.5	46.3	-
LAVENDER (Li 等人, 2022d)	30M	45.0	56.6	73.5
All-in-one (Wang 等人, 2022a)	283M	46.8	48.3	66.3
MERLOT(Zellers 等人,2021)	180M	43.1	-	69.5
OmniVL (Wang 等人, 2022b)	18M	44.1	51.0	-
HiTeA (Ye 等人, 2022)	17M	45.9	55.3	73.2
GIT (Wang 等人, 2022c)	800M	43.2	56.8	72.8
GIT2 (Wang 等人, 2022c)	12.9B	45.6	**58.2**	74.9
FrozenBiLM(Yang 等人，2022)	10M	47.0	54.8	68.6
VideoCoCa (Yan 等人，2022)	3B	46.0	56.9	-
InternVideo(Wang 等人，2022g)	12M	47.1	55.5	72.2
Flamingo (Alayrac 等人，2022)	2.3B	47.4	-	-
mPLUG-2_{Base}	17M	46.3	55.3	72.6
mPLUG-2	17M	**48.0**	58.1	**75.4**

双视觉编码器中的局部时序建模模块。

视频问答 表 4 汇总了 MSRVTT-QA（Xu 等人，2017年）、MSVD-QA（Xu 等人，2017年）和 TGIF-FrameQA（Jang 等人，2017年）上的视频问答结果。可以看到，mPLUG-2 在 MSRVTT-QA 和 TGIF-FrameQA 上大幅超越所有现有基础模型；即使使用的预训练数据量显著更少，在 MSVD-QA 上也取得了与大型基础模型 GIT2（Wang 等人，2022c）相当的结果。具体而言，mPLUG-2 在 MSRVTT 上绝对提升了 0.6%，在 TGIF-FrameQA 上绝对提升了 0.5%。此外，mPLUG-2_{Base} 以更小的模型规模取得了与大型模型（即 VideoCoCa 和 GIT2）相当的结果。

视频描述 表 55 比较了 mPLUG-2 与现有方法在 MSRVTT 和 MSVD 视频描述数据集上的表现。如表中所示，尽管预训练数据较少，mPLUG-2 在 MSRVTT 数据集上取得了显著提升，在 MSVD 数据集上的表现也具有可比性。在 MSRVTT 字幕生成任务上，我们的方法在 CIDEr 和 BLEU@4 指标上分别比最先进水平的方法 VideoCoCa（Yan 等人，2022）和 GIT2（Wang 等人，2022c）高出 4.4% 和 3.0%。此外，我们注意到，在预训练数据量相同的情况下，mPLUG-2 的表现优于 HiTeA，这表明 mPLUG-2 能够生成更强的视频-语言表征。

视觉定位 我们将 mPLUG-2 与现有最先进方法在视觉定位数据集上的表现进行了比较，这些数据集包括 RefCOCO（Yu 等人，2016年）、RefCOCO+（Yu 等人，2016年）和 RefCOCOg（Mao 等人，2016年）。表 7显示，mPLUG-2 的表现与最先进方法相当。未使用目标检测数据进行预训练时，我们的方法在 RefCOCO“testB”划分上的表现比排名第二的方法绝对提升了 0.97%。在“testB”划分中，查询可以指向各种视觉概念，而在“testA”中则只能指向人物。这一提升表明，引入通用层有助于模型理解图像中的视觉概念。

图像-文本检索 我们在 MSCOCO 和 Flickr30K 图像-文本检索数据集上评估 mPLUG-2。如表 6所示，mPLUG-2_{Base} 和 mPLUG-2 的表现均达到或优于最先进方法。Florence（Yuan 等人，2021）和 BLIP（Li 等人，2022c）预训练时分别使用了 0.9B 和 129M 条数据。相比之下，我们的 mPLUG-2 仅需 17M 条数据。这表明 mPLUG-2 的数据利用效率很高。

视觉问答 我们报告了 mPLUG-2 在视觉问答测试集上的表现。mPLUG-2 在 test-dev 和 test-std 上分别超越最先进方法 Florence（Yuan 等人，2021）0.95% 和 0.77%。我们模型所用预训练数据的规模比 Florence 少 89.11%。这表明我们的 mPLUG-2 能够高效且有效地学习多模态表征。

图像描述 我们在 MSCOCO（Lin等，2014）上将 mPLUG-2 与现有最先进方法进行比较。按照（Li 等人，2020b）的方法，我们使用交叉熵损失在 COCO 描述数据集上训练 mPLUG-2，并在相同的 Karpathy 划分上测试。如表 9所示，我们的 mPLUG-2 在 COCO 描述任务上取得了新的最先进水平。此外，尽管 LEMON（Hu 等人，2021）和 BLIP（Li 等人，2022c）等大型基础模型使用的预训练数据量近乎多出 10 倍，我们的方法仍取得了具有竞争力的结果。具体而言，在 COCO 描述任务上，我们的 mPLUG-2 的 BLEU@4 比 BLIP 高出显著的 1.2 分，CIDEr 高出 1 分。

表5. P 视频描述性能对比。B@4：BLEU@4；M：METEOR；R：ROUGE-L；C：CIDE r.

方法	#预训练数据	MSRVTT				MSVD			
		B@4	M	R	C	B@4	M	R	C
UniVL (Luo 等人，2020年)	136M	42.2	28.2	61.2	49.9	-	-	-	-
SwinBERT (Lin等，2022)	-	41.9	29.9	62.1	53.8	58.2	41.3	77.5	120.6
CLIP4Caption(Tang 等人，2021)	-	46.1	30.7	63.7	57.7	-	-	-	-
MV-GPT (Seo等人，2022)	69M	48.9	38.7	64.0	60.0	-	-	-	-
LAVENDER (Li 等人，2022d)	30M	-	-	-	60.1	-	-	-	150.7
HiTeA(Ye 等人，2022)	17M	49.2	30.7	65.0	65.1	71.0	45.3	81.4	146.9
VideoCoCa（Yan 等人，2022）	3B	53.8	-	68.0	73.2	-	-	-	-
GIT（Wang 等人，2022c）	0.8B	53.8	32.9	67.7	73.9	79.5	51.1	87.3	180.2
GIT2（Wang 等人，2022c）	12.9B	54.8	33.1	68.2	75.9	**82.2**	**52.3**	**88.7**	**185.4**
mPLUG-2$_{Base}$	17M	52.2	32.1	66.9	72.4	69.3	45.1	81.9	148.2
mPLUG-2	17M	**57.8**	**34.9**	**70.1**	**80.3**	75.0	48.4	85.3	165.8

Ta表 6. **图像-文本检索性能对比**。所有结果均以 R@1/R@5/R@10 报告 .

方法	#预训练数据	MSCOCO（5K 测试集）						Flickr30K（1K 测试集）					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
E2E-VLP（Xu 等人）	4M	-	-	-	-	-	-	86.2	97.5	98.92	73.6	92.4	96.0
UNITER（Chen 等人，2020年）	400万	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR（Li 等人，2020b）	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO（Li 等人，2020a）	4M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
VLMo（Wang 等人，2021a）	4M	78.2	94.4	97.4	60.6	84.4	91.0	95.3	99.9	**100.0**	84.5	97.3	98.6
ALIGN（Jia 等人，2021）	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
ALBEF（Li 等人，2021b）	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100.0**	85.6	97.5	98.9
Florence（Yuan 等人，2021）	0.9B	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	**98.1**	-
BLIP（Li 等人，2022c）	129M	82.4	95.4	97.9	65.1	86.3	91.8	**97.4**	99.8	99.9	87.6	97.7	99.0
mPLUG-2$_{Base}$	17M	81.2	95.6	**98.1**	65.3	86.9	92.4	96.9	**100.0**	**100.0**	**88.2**	97.8	99.0
mPLUG-2	17M	**82.5**	**95.7**	98.0	**65.7**	**87.1**	**92.6**	97.2	**100.0**	**100.0**	88.1	97.6	**99.1**

表 7。**视觉定位（ReferCOCO 和 ReferCOCOg）评估结果**。我们使用 accuracy@0.5（当定位真值框与预测边界框之间的交并比大于 0.5 时，预测即为正确）来衡量模型性能。

模型	RefCOCO			RefCOCOg	
	val	testA	测试集B	验证集-u	测试集-u
UNITER (Chen 等人, 2020年)	81.41	87.04	74.17	74.86	75.77
VILLA (Gan 等人，2020年)	82.39	87.48	74.84	76.18	76.71
MDETR (Kamath 等人，2021)	86.75	89.58	81.41	81.64	80.89
UNICORN(Yang 等人，2021b)	88.29	90.42	83.06	83.44	83.93
OFA$_{Large}$ (Wang 等人，2022d)	90.05	**92.93**	85.26	84.54	**85.20**
mPLUG-2	**90.33**	92.80	**86.05**	**84.70**	85.14

表 8。**视觉问答性能比较**。评估结果以准确率表示。

方法	#预训练数据	test-dev	test-std
UNITER（Chen 等人，2020）	4M	72.70	72.91
UNIMO（Li 等人，2020a）	4M	73.79	74.02
E2E-VLP(Xu 等人)	4M	73.25	73.67
OSCAR(Li 等人, 2020b)	4M	73.16	73.44
ALBEF(Li 等人,2021b)	4M	74.54	74.70
BLIP (Li 等人, 2022c)	14M	77.54	77.62
SimVLM (Wang 等人, 2021b)	1.8B	80.03	80.34
Florence(Yuan 等人, 2021)	0.9B	80.16	80.36
OFA$_{Large}$ (Wang 等人，2022d)	18M	80.30	80.50
VLMo(Wang 等人，2021a)	-	79.94	79.98
GIT(Wang 等人，2022c)	0.8B	78.56	78.81
mPLUG-2$_{Base}$	17M	79.27	79.32
mPLUG-2	17M	**81.11**	**81.13**

4.2.2. 纯语言任务

自然语言理解 我们在 GLUE 基准测试（Wang 等人，2018年）的 6 项任务上评估 mPLUG-2 的自然语言理解能力。表10显示，mPLUG-2 的表现可与最先进的自然语言和多模态预训练模型相媲美，包括 RoBERTa（Liu 等人，2019年）、DeBERTa（He 等人，2021b）。与 DeBERTa（He 等人，2021b）相比，我们结合 DeBERTa 的方法在三项任务上取得了提升，这也证明了用于模态协作的通用模块的有效性。

自然语言生成 我们在 Gigaword 抽象摘要任务（Rush 等人，2015b）上评估 mPLUG-2 的自然语言生成能力。如表 1111所示，mPLUG-2 达到了与当前最先进模型相当的结果。

4.2.3. 纯视觉任务

视频动作识别 视频动作识别是最具代表性的视频理解任务，因为它要求模型理解视频中呈现的时空线索。表 12总结了不同方法在 Kinetics 400、Kinetics 600 和 Kinetics 700 数据集上的表现。mPLUG-2 超越了大多数最先进方法。例如，与基于 9 亿个视觉-文本对预训练的 Florence 相比，mPLUG-2 在 Kinetics 600 上的 Top-1 准确率提高了 1.9%，在 Kinetics 400 上提高了 0.6%。同时可以看到，在预训练数据量相近的情况下，mPLUG-2 的表现优于 OmniVL，这表明双视觉编码器模块对视频表征学习是有效的。

表*9*。**图像描述性能比较**。B@4：BLEU@4；M：METEOR；R：ROUGE-L；C：CIDEr。

方法	#预训练数据	COCO 描述			
		B@4	M	C	S
编码器-解码器	12M	-	-	110.9	-
E2E-VLP（Xu 等人）	4M	36.2	-	117.3	-
VinVL（Zhang 等人，2021b）	565万	38.5	30.4	130.8	
OSCAR（Li 等人，2020b）	6.5M	-	-	-	-
LEMON$_{large}$（Hu 等人，2021）	200M	40.6	30.4	135.7	23.5
BLIP（Li 等人，2022c）	129M	40.4	-	136.7	-
mPLUG-2	17M	**41.6**	**30.9**	**137.7**	**23.7**

表*10*。**GLUE 基准上的实验结果**。

模型	SST-2	RTE	MRPC	QQP	MNLI	QNLI
多模态预训练基线模型						
VisualBERT（Li 等人，2019年）	89.4	56.6	71.9	89.4	81.6	87.0
UNITER（Chen 等人，2020年）	89.7	55.6	69.3	89.2	80.9	86.0
VL-BERT（Su 等人，2019年）	89.8	55.7	70.6	89.0	81.2	86.3
VilBERT（Lu 等人，2019年）	90.4	53.7	69.0	88.6	79.9	83.8
LXMERT（Tan 和 Bansal，2019年）	90.2	57.2	69.8	75.3	80.4	84.2
Uni-Perceiver（Zhu 等人，2021）	90.2	64.3	86.6	87.1	81.7	89.9
SimVLM（Wang 等人，2021b）	90.9	63.9	75.2	90.4	83.4	88.6
FLAVA（Singh 等人，2021）	90.9	57.8	81.4	90.4	80.3	87.3
UNIMO（Li 等人，2020a）	96.8	-	-	-	89.8	-
OFA（Wang 等人，2022d）	96.6	**91.0**	91.7	92.5	90.2	94.8
自然语言预训练的最先进模型						
BERT（Devlin 等人，2018年）	93.2	70.4	88.0	91.3	86.6	92.3
RoBERTa（Liu 等人，2019年）	96.4	86.6	90.9	92.2	90.2	93.9
XLNet（Yang 等人，2019年）	**97.0**	85.9	90.8	92.3	90.8	94.9
ELECTRA（Clark 等人，2020年）	96.9	88.0	90.8	92.4	90.9	95.0
DeBERTa(He 等人，2021b)	96.8	88.3	91.9	92.3	**91.1**	**95.3**
mPLUG-2 $_{Base}$	93.5	85.2	87.3	91.3	87.6	93.2
mPLUG-2	95.1	88.0	90.1	**92.7**	90.2	94.5
mPLUG-2$_{Deberta}$	96.2	89.4	**92.1**	92.6	90.8	94.8

图像分类 我们进一步评估了 mPLUG-2 在 ImageNet-1K 上的图像分类表现。如表 13所示，mPLUG-2 在未使用 ImageNet 数据进行预训练的情况下，取得了与最先进方法相当甚至更好的结果。此外，为了有效评估 mPLUG-2 的鲁棒性和泛化能力，我们在 5 个 ImageNet 变体（即 IN-V2、IN-ReaL、IN-Adversarial、IN-Rendition 和 IN-Sketch）上进行了评估。按照标准评估流程（Fang 等人，2022a），所有模型均先在原始 ImageNet-1K 训练集上进行微调，然后直接在这 6 个变体上测试，不再进一步微调。如表 13所示，mPLUG-2 不仅在 ImageNet-1K 验证集上取得最高准确率，而且差距也相对较小（即 $\Delta_{\downarrow}$），这体现了借助通用层模块学习语言共享表示后，mPLUG-2 出色的鲁棒性和泛化能力。

4.3. 讨论

基于指令的学习的影响 基于指令的学习能够通过特定指令区分不同类型的任务。表 14 展示了基于指令的学习的有效性。在表中，我们可以观察到，基于指令的学习提升了

表*11*。**Gigaword 抽象式摘要实验结果**。 我们报告 ROUGE 评估结果。

模型	Gigaword		
	ROUGE-1	ROUGE-2	ROUGE-L
BERTSHARE（Rothe等人，2020年）	38.13	19.81	35.62
MASS（Song等人，2019年）	38.73	19.71	35.96
UniLM（Dong等人，2019年）	38.45	19.45	35.75
PEGASUS（Zhang 等人，2020年）	39.12	19.86	36.24
ProphetNet（Qi等人，2020年）	39.55	20.27	36.57
UNIMO（Li 等人，2020a）	39.71	20.37	36.88
OFA (Wang 等人，2022d)	**39.81**	20.66	**37.11**
mPLUG-2	39.65	**20.67**	36.89

表 *12.* **在微调设置下与视频动作识别最先进水平的比较**。

方法	Kinetics 400		Kinetics 600		Kinetics 700	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
TimeSformer-L (Bertasius 等人，2021)	80.6	94.7	82.2	95.6	-	-
ViViT-H (Arnab 等人，2021)	84.8	95.8	85.8	96.5	-	-
VideoSwin-L (Liu 等人，2022b)	84.9	96.7	86.1	97.3	-	-
OmniVL(Wang 等人，2022b)	79.1	94.5	-	-	-	-
TokenLearner (Ryoo 等人，2021)	85.4	96.3	86.3	97.0	-	-
VATT（Akbari 等人，2021）	82.1	95.5	83.6	96.6	-	-
MoViNet（Kondratyuk 等人，2021）	81.5	-	84.8	-	79.4	-
Florence（Yuan 等人，2021）	86.5	97.3	87.8	97.8	-	-
CoVeR（Zhang 等人，2021a）	86.3	97.2	87.9	97.8	78.5	94.2
mPLUG-2$_{Base}$	83.6	96.0	86.7	97.2	74.6	91.2
mPLUG-2	**87.1**	**97.7**	**89.8**	**98.3**	**80.4**	**94.9**

表*13*。**在 ImageNet-1K 变体上的鲁棒性和泛化能力评估方面与最先进方法的比较**。 我们在多个 ImageNet-1K 验证集上测试模型，未进行任何进一步微调。“Avg.”表示在 6 个不同 ImageNet-1K 变体上的平均 Top-1 准确率。“$\Delta_{\downarrow}$”表示 6 个变体的平均 Top-1 准确率与原始 ImageNet-1K 验证集准确率之间的差距（越低越好）。

方法	IN-1K	IN -V2	IN -ReaL	IN -Adv.	IN-Ren.	IN-Ske.	Avg.	$\Delta_{\downarrow}$
ConvNeXt（Liu 等人，2022a）	87.5	77.7	90.5	70.8	67.0	53.7	74.5	13.0
SwinV2-G（Liu 等人，2021a）	87.5	77.3	90.2	73.9	67.7	52.3	74.8	12.7
MAE（He 等人，2021a）	87.8	79.2	90.3	76.7	66.5	50.9	75.2	12.6
DeiT3（Touvron 等人，2022）	87.7	79.1	90.2	79.2	70.6	54.9	77.0	10.7
Eff-L2-NS（Tan 和 Le，2019年）	88.4	**80.5**	**90.6**	**84.8**	74.7	47.6	**77.8**	**10.6**
OFA（Wang 等人，2022d）	85.6	-	-	-	-	-	-	-
mPLUG-2	**88.5**	78.1	89.5	73.2	**75.6**	**61.2**	77.7	10.7

表 *14*。**评估所提出的基于指令的学习在下游任务上的表现**。对于检索任务，我们报告 Recall@1、Recall@5 和 Recall@10 的平均值。对于问答和图像描述任务，报告 Top-1 准确率和 CIDEr。

指令	MSRVTT -Ret.	MSVD-QA	MSRVTT -Cap.
	72.8	54.1	71.8
✓	**73.5** (+0.7)	**55.3** (+1.2)	**72.4** (+0.6)

表*15*。**双视觉编码器模块中不同时间建模模块的评估**。对于检索任务，我们报告 Recall@1、Recall@5 和 Recall@10 的平均值。对于问答和图像描述任务，我们报告 Top-1 准确率和 CIDEr。

时序模块 MSRVTT	-Ret.	MSVD-问答 MSRVTT	-Cap.
时间自注意力	70.3	55.1	71.1
时间卷积	71.4 (+1.1)	55.0 (-0.1)	71.7 (+0.6)
局部时间建模	**73.5** (+3.2)	**55.3** (+0.2)	**72.4** (+1.3)

检索和问答的性能分别在平均召回率和准确率上至少提升了 0.7% 和 1.2%。借助基于指令的学习，mPLUG-2 能够在使用不同

表16. 通用层对提升视觉任务性能的影响评估。

方法	ImageNet	CIFAR10	CIFAR100	Cars	DTD	SUN	Food101	平均值
CLIP-ViT-L/14	86.2	98.6	92.2	91.6	81.9	80.7	94.4	89.4
+通用层	86.6 (+0.4)	99.3 (+0.7)	93.1 (+0.9)	94.4 (+2.8)	85.1 (+3.2)	80.4 (-0.4)	95.4 (+1.0)	**90.6** (+1.2)

表17。通用层对提升语言和视觉语言任务性能的影响评估 。

模型	SST-2	RTE	MRPC	QQP	MNLI	QNLI	视觉问答测试-dev
$\text{BERT}_{\text{预训练}}$	91.7	71.4	86.3	90.8	84.3	89.3	78.6
+联合训练	92.5	82.3	86.6	90.6	86.2	92.1	78.9
+通用层	**93.5**	**85.2**	**87.3**	**91.3**	**87.6**	**93.2**	**79.3**

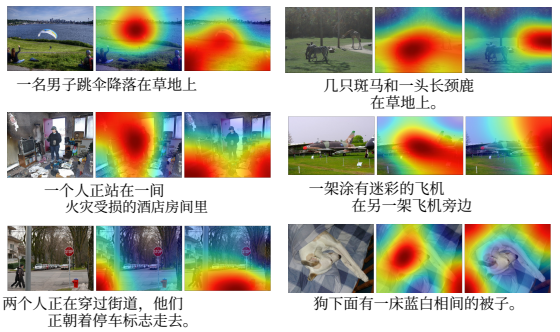

图 3. 通用层中潜在查询的 Grad-CAM 可视化。

指令时调用不同的模块，以提升性能。

局部时序建模模块的影响 为了验证我们提出的双视觉编码器局部时序建模模块的有效性，我们采用不同的时序建模结构进行了实验。具体而言，我们尝试了时间自注意力和时间卷积进行比较。结果汇总于表 15。我们可以看到，通过引入时间局部性建模，局部时序建模模块的表现优于时间自注意力模块。同时，借助多组融合机制，局部时序建模模块能够在不同的表征子空间中学习多样化的时间表示，而时间卷积则受限于相同的时序表征空间，因此前者表现更好。

通用层的影响 为了验证我们提出的通用层模块的有效性，我们在所有单模态和多模态任务中对该模块进行了消融实验。如表 16 和表 17所示，在本实验中，我们分别将第 1/2/2 行设为视觉、语言和视觉语言任务的基线。我们发现，与基线相比，共享通用层有助于所有模态任务，因为它促进了不同模态之间的协作。

在图 3中，我们将 Grad-CAM 可视化在交叉

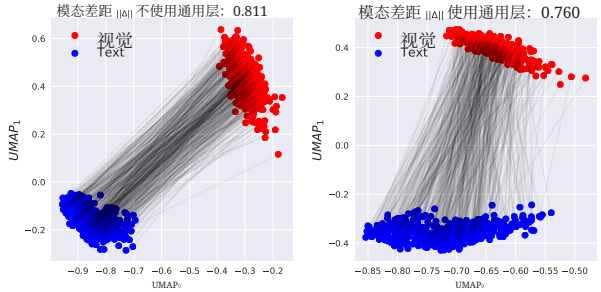

图 4。预训练 mPLUG-2 生成的视觉和语言嵌入的 UMAP 可视化结果。黑线表示视觉语言对。

第一层通用层中的注意力图。对于每个样本，我们展示两张关注不同视觉概念的交叉注意力图。结果表明，通用层能够通过关注图像中各种视觉概念所在的区域，促进视觉块特征与语言特征之间的模态协作和模态纠缠。

用于模态协作的通用层 在此，我们研究通用层对模态协作的影响。我们随机抽取一些视觉语言对，并在图 4中绘制预训练 mPLUG-2 生成的嵌入的 UMAP 可视化结果。可以观察到，在通用层的帮助下，视觉样本与文本样本之间的距离更近，而不是仅仅形成两个聚集的簇。此外，我们定量计算了模态差距 $\|\Delta\|$ (Liang 等人，2022），其中 Δ 表示视觉嵌入中心与文本嵌入中心之间的差异。可以看出，与不含通用层的模型相比，含通用层的模型能够促进视觉和语言模态之间的协作，从而缩小模态差距。

5. 结论

本文介绍了 mPLUG-2，这是一种采用模块化设计来构建多模态基础模型的新型统一范式。mPLUG-2 引入了基于模块的网络设计，通过共享通用模块实现模态协作，并将模态专属模块解耦，以解决模态纠缠问题。实验结果表明，mPLUG-2 的这一新型统一范式在文本、图像和视频模态的 30 多项任务中均表现出色。通过选择

并添加模块。

参考文献

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., andBatra, D. 视觉问答：视觉问答。国际计算机视觉期刊，123(1):4–31，2017年5月。

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. VATT：利用Transformer从原始视频、音频和文本进行多模态自监督学习。神经信息处理系统进展，34，2021。

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Si- monyan, K. Flamingo：一种用于少样本学习的视觉语言模型。*arXiv*，abs/2204.14198，2022。

Anne Hendricks, L.、Wang, O.、Shechtman, E.、Sivic, J.、Darrell, T. 和 Russell, B. 使用自然语言定位视频片段。载于*IEEE/CVF*国际计算机视觉会议论文集，第5803–5812页，2017年。

Arnab, A.、Dehghani, M.、Heigold, G.、Sun, C.、Lucic, M. 和 Schmid, C. Vivit：视频视觉Transformer。载于*2021 IEEE/CVF*国际计算机视觉会议，*ICCV 2021*，加拿大魁北克省蒙特利尔，*2021*年*10*月*10*日至*17*日，第6816–6826页。IEEE，2021。

Bain, M.、Nagrani, A.、Varol, G. 和 Zisserman, A. Frozen in Time：用于端到端检索的联合视频图像编码器。*2021 IEEE/CVF*国际计算机视觉会议（*ICCV*），第1708–1718页，2021a。

Bain, M., Nagrani, A., Varol, G., 和 Zisserman, A. Frozen in Time：用于端到端检索的联合视频图像编码器。载于*IEEE/CVF*国际计算机视觉会议论文集，第1728–1738页，2021b。

Bertasius, G., Wang, H., 和 Torresani, L. 时空注意力是否就是视频理解所需的一切？载于 *ICML*，第2卷，第4页，2021。

Brown, T. B.、Mann, B.、Ryder, N.、Subbiah, M.、Kaplan, J.、Dhariwal, P.、Neelakantan, A.、Shyam, P.、Sastry, G.、Askell, A.、Agarwal, S.、Herbert-Voss, A.、Krueger, G.、Henighan, T.、Child, R.、Ramesh, A.、Ziegler, D. M.、Wu, J.、Winter, C.、Hesse, C.、Chen, M.、Sigler, E.、Litwin, M.、Gray, S.、Chess, B.、Clark, J.、Berner, C.、McCandlish, S.、Radford, A.、Sutskever, I. 和 Amodei, D. 语言模型是少样本学习者。载于 Larochelle, H.、Ranzato, M., Hadsell , R., Balcan , M., 和 Lin , H. (eds.), 进展神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统大会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行，2020年。Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., 和 Zagoruyko, S. 使用 Transformer 实现端到端目标检测。载于 Vedaldi, A., Bischof, H., Brox, T., 和 Frahm, J.（编），计算机视觉—*ECCV 2020*—第*16*届欧洲会议，英国格拉斯哥，*2020*年*8*月*23*日至*28*日，论文集，第一部分，第12346卷，第213–229页，2020年。

Changpinyo, S., Sharma, P., Ding, N., 和 Soricut, R. Con- ceptual 12m：推动网络规模图文预训练识别长尾视觉概念。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3558–3568 页，2021。

Chen, D. 和 Dolan, W. B. 收集用于释义评估的高度平行数据。载于计算语言学协会人类语言技术第*49*届年会，第 190–200 页，2011年。

Chen, T., Saxena, S., Li, L., Fleet, D. J., 和 Hinton, G. Pix2seq：用于目标检测的语言建模框架。*arXiv* 预印本 *arXiv:2109.10852*，2021。

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., 和 Liu, J. Uniter：通用图像-文本表示学习。载于欧洲计算机视觉会议，第 104–120 页。施普林格，2020年。

Clark, K., Luong, M., Le, Q.V. 和 Manning, C.D. ELEC- TRA：将文本编码器作为判别器而非生成器进行预训练。载于第八届国际学习表征会议，*ICLR 2020*。OpenReview.net，2020年。

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. 和 Fei-Fei, L. ImageNet：一个大规模层次化图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。IEEE，2009年。

Devlin, J., Chang, M.-W., Lee, K. 和 Toutanova, K. BERT：用于语言理解的深度双向Transformer预训练。*arXiv*预印本 *arXiv:1810.04805*，2018年。

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M. 和 Hon, H. 用于自然语言理解和生成的统一语言模型预训练。载于 *NeurIPS2019*，第13042–13054页，2019年。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等。图像胜过 16x16

个词：用于大规模图像识别的 Transformer。*arXiv* 预印本 *arXiv:2010.11929*，2020年。

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., 和 Cao, Y. EVA：探索大规模掩码视觉表征学习的极限。*arXiv* 预印本 *arXiv:2211.07636*，2022a。

Fang, Y., Wang, W., Xie, B., Sun, Q.-S., Wu, L. Y., Wang, X., Huang, T., Wang, X., 和 Cao, Y. EVA：探索大规模掩码视觉表征学习的极限。*ArXiv*，abs/2211.07636，2022b。

Fu, T.-J., Li, L., Gan, Z., Lin, K., Wang, W. Y., Wang, L., 和 Liu, Z. VIOLET：采用掩码视觉词元建模的端到端视频语言转换器。*arXiv* 预印本 *arXiv:2111.12681*，2021。

Gan, Z.、Chen, Y.、Li, L.、Zhu, C.、Cheng, Y. 和 Liu, J. 面向视觉与语言表征学习的大规模对抗训练。载于 Larochelle, H.、Ranzato, M.、Hadsell, R.、Balcan, M. 和 Lin, H.（编），神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统大会年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。

Ge, Y.、Ge, Y.、Liu, X.、Li, D.、Shan, Y.、Qie, X. 和 Luo, P. 通过多项选择题连接视频文本检索。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第16167–16176页，2022。

Ghiasi, G.、Cui, Y.、Srinivas, A.、Qian, R.、Lin, T.、Cubuk, E. D.、Le, Q. V. 和 Zoph, B. 简单的复制粘贴是实例分割的一种强大数据增强方法。载于 *IEEE*计算机视觉与模式识别会议，*CVPR 2021*，线上，*2021*年*6*月*19*日至*25*日，第2918–2928页。计算机视觉基金会 / IEEE，2021。

He, K., Zhang, X., Ren, S. 和 Sun, J. 用于图像识别的深度残差学习。*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR*），第770–778页，2015年。

He, K., Gkioxari, G., Doll´ar, P. 和 Girshick, R. B. Mask R-CNN。*2017*年*IEEE*国际计算机视觉会议（*ICCV*），第2980–2988页，2017年。

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P. 和 Girshick, R. B. 掩码自编码器是可扩展的视觉学习器。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第15979–15988页，2021a。

He, P., Liu, X., Gao, J., and Chen, W. Deberta：采用解耦注意力的增强型 BERT 解码器。载于第九届国际学习表征会议（*ICLR 2021*）。OpenReview.net，2021b。

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. 扩展视觉语言预训练以实现图像描述。*2022 IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），第17959–17968页，2021。Huang, J., Li, Y., Feng, J., Sun, X., and Ji, R. Clover：迈向统一的视频语言对齐与融合模型。*arXiv* 预印本 *arXiv:2207.07885*，2022a。Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. 模态竞争：是什么导致多模态网络的联合训练在深度学习中失败（可证明）？ 载于国际机器学习会议，第9226–9259页。PMLR，2022b。

Jang, Y., Song, Y., Yu, Y., Kim, Y. 和 Kim, G. Tgif-问答：迈向视觉问答中的时空推理。载于*IEEE*计算机视觉与模式识别会议论文集，第2758–2766页，2017年。

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., and Duerig, T. 使用噪声文本监督扩展视觉与视觉语言表征学习。发表于国际机器学习会议，2021年。

Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. MDETR 调制的检测方法，用于端到端多模态理解。发表于*IEEE/CVF*国际计算机视觉会议论文集，第 1780–1790 页，2021。

Karpathy, A. and Fei-Fei, L. 用于生成图像描述的深度视觉-语义对齐。发表于*IEEE*计算机视觉与模式识别会议论文集，第 3128–3137 页，2015年。

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., 和 Zisserman, A. Kinetics 人类动作视频数据集。*ArXiv*，abs/1705.06950，2017年。

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., 和 Gong, B. Movinets：用于高效视频识别的移动视频网络。载于 *IEEE*计算机视觉与模式识别会议，*CVPR 2021*，线上举行，*2021*年*6*月*19*日至*25*日，第16020–16030 页。计算机视觉基金会 / IEEE，2021。

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., 等。Visual Genome：利用众包密集图像标注连接语言与视觉。国际计算机视觉杂志，123(1):32–73，2017年。

Lei, J., Berg, T. L., 和 Bansal, M. 揭示视频与语言学习中的单帧偏差。*arXiv* 预印本 *arXiv:2206.03428*，2022。

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART：面向自然语言生成、翻译和理解的去噪序列到序列预训练。载于 Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R.（编），第*58*届计算语言学协会年会论文集，*ACL 2020*，线上会议，*2020*年*7*月*5*日至*10*日，第7871–7880页。计算语言学协会，2020年。

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., da Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., and Si, L. mplug：通过跨模态跳跃连接实现高效有效的视觉语言学习。*arXiv*，abs/2205.12005，2022a。

Li, D., Li, J., Li, H., Niebles, J. C., 和 Hoi, S. C. 对齐与提示：利用实体提示进行视频与语言预训练。*arXiv* 预印本 *arXiv:2112.09583*，2021a。

Li, D., Li, J., Li, H., Niebles, J. C., 和 Hoi, S. C. 对齐与提示：利用实体提示进行视频与语言预训练。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第4953–4963 页，2022b。

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., 和 Hoi, S. C. H. 先对齐再融合：通过动量蒸馏进行视觉与语言表征学习。神经信息处理系统进展，34，2021b。

Li, J., Li, D., Xiong, C., 和 Hoi, S. Blip：用于统一视觉-语言理解与生成的自举式语言-图像预训练。*arXiv* 预印本 *arXiv:2201.12086*，2022c。

Li, L., Gan, Z., Lin, K., Lin, C.-C., Liu, Z., Liu, C., and Wang, L. Lavender：将视频-语言理解统一为掩码语言建模。*arXiv*预印本 *arXiv:2206.07160*，2022d。

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert：面向视觉与语言的简单且高效的基线方法。*arXiv*预印本 *arXiv:1908.03557*，2019年。Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.- W., and Gao, J. 接地语言-图像预训练。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第10955–10965页，2021c。

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. UNIMO：通过跨模态对比学习实现统一模态理解与生成。*arXiv*预印本 *arXiv:2012.15409*，2020a。Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., 等。OSCAR：面向视觉语言任务的目标-语义对齐预训练。载于欧洲计算机视觉会议，第121–137页。施普林格，2020b。Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., and Ling, H. CBNet：用于目标检测的复合骨干网络架构。*IEEE*图像处理汇刊，31：6893–6906，2021。

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. 正视差距：理解多模态对比表征学习中的模态差距。*arXiv*预印本 *arXiv:2203.02053*，2022。

Lin, K., Li, L., Lin, C.-C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., and Wang, L. Swinbert：采用稀疏注意力进行视频描述的端到端Transformer。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第17949–17958页，2022。

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P. 和 Zitnick, C. L. Microsoft coco：上下文中的常见物体。在 欧洲计算机视觉会议，第740–755页。施普林格，2014年。

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. 和 Stoyanov, V. Roberta：一种经过稳健优化的BERT预训练方法。*arXiv*预印本 *arXiv:1907.11692*，2019年。

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F. 和 Guo, B. Swin Transformer V2：扩展容量和分辨率。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第11999–12009页，2021a。

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. 和 Guo, B. Swin Transformer：采用移位窗口的分层视觉Transformer。国际计算机视觉会议 *(ICCV)*，2021b。

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T. 和 Xie, S. 面向 2020年代的卷积网络。*2022*年*IEEE/CVF*计算机视觉与模式识别会议 *(CVPR)*，第 11966–11976 页，2022a。

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S. 和 Hu, H. 视频Swin Transformer。载于*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2022*，新奥尔良，路易斯安那州，美国，*2022*年*6*月*18*日至*24*日，第3192–3201页。IEEE，2022b。

Loshchilov, I. 和 Hutter, F. 解耦权重衰减正则化。载于 *ICLR*，2019年。

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert：为视觉与语言任务预训练任务无关的视觉语言表征。载于 神经信息处理系统进展，第 13–23 页，2019年。

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., and Zhou, M. Univl：一种用于多模态理解与生成的统一视频和语言预训练模型。*arXiv* 预印本 *arXiv:2002.06353*，2020年。

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip：一项关于将 CLIP 用于端到端视频片段检索和图像描述的实证研究。*Neurocomputing*，508：293–304，2022。

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. X-clip：面向视频文本检索的端到端多粒度对比学习。*arXiv* 预印本 *arXiv:2207.07285*，2022。

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. 无歧义物体描述的生成与理解。载于*IEEE*计算机视觉与模式识别会议论文集，第 11–20 页，2016年。

Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text：使用 100 万张带描述的照片描述图像。载于 *NIPS*，2011 年。

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., K¨opf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch：一种命令式风格、高性能的深度学习库。载于神经信息处理系统，2019年。

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k Entities：收集区域与短语对应关系，以构建更丰富的图像到句子模型。载于 *IEEE*国际计算机视觉会议论文集，第 2641–2649 页，2015年。

Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. ProphetNet：预测序列到序列预训练中的未来 n元语法。载于 *2020*年自然语言处理实证方法会议：研究发现论文集，第 2401–2410 页，2020年。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等. 从自然语言监督中学习可迁移的视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021。Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. 探索统一文本到文本转换器迁移学习的极限。*J. Mach. Learn.Res.*，21:140:1–140:67，2020。Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. 一份用于电影描述的数据集。载于*IEEE*计算机视觉与模式识别会议论文集，第 3202–3212 页，2015年。

Rothe, S., Narayan, S., and Severyn, A. 利用预训练检查点执行序列生成任务。计算语言学协会会刊，8：264–280，2020年。

Rush, A. M., Chopra, S., 和 Weston, J. 用于抽象式句子摘要的神经注意力模型。载于自然语言处理实证方法会议，2015a。

Rush, A. M., Chopra, S., 和 Weston, J. 用于抽象式句子摘要的神经注意力模型。载于*2015*年自然语言处理实证方法会议论文集，第 379–389 页，2015b。

Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., 和 Angelova, A. TokenLearner：8 个学到的 Token 能为图像和视频做什么？*arXiv*，abs/2106.11297，2021。

Seo, P. H., Nagrani, A., Arnab, A., 和 Schmid, C. 面向多模态视频字幕生成的端到端生成式预训练。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 17959–17968 页，2022。

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., 和 Sun, J. Objects365：一个用于目标检测的大规模高质量数据集。*2019*年 *IEEE/CVF* 国际计算机视觉会议 *(ICCV)*，第 8429–8438 页，2019年。

Sharma, P., Ding, N., Goodman, S., 和 Soricut, R. Conceptual captions：一个经过清洗、上位词处理的图像替代文本数据集，用于自动图像描述。载于计算语言学协会第 *56* 届年会论文集（第 *1* 卷：长篇论文），第 2556–2565 页，2018年。

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. FLAVA：一种基础语言与视觉对齐模型。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 15617–15629 页，2021。

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. MASS：用于语言生成的掩码序列到序列预训练。载于 *ICML2019*，第5926–5936 页，2019年。

Su, W.、Zhu, X.、Cao, Y.、Li, B.、Lu, L.、Wei, F. 和 Dai, J. VL-BERT：通用视觉语言表征的预训练。*arXiv* 预印本 *arXiv:1908.08530*，2019年。

Szegedy, C.、Ioffe, S.、Vanhoucke, V. 和 Alemi, A. A. Inception-v4、Inception-ResNet 以及残差连接对学习的影响。在第三十一届 *AAAI*人工智能会议，2017年。

Tan, H. 和 Bansal, M. LXMERT：从 Transformer 中学习跨模态编码器表征。*arXiv* 预印本 *arXiv:1908.07490*，2019年。

Tan, M. 和 Le, Q. V. EfficientNet：重新思考卷积神经网络的模型缩放。*arXiv*，abs/1905.11946，2019年。Tang, M., Wang, Z., Liu, Z., Rao, F., Li, D. 和 Li, X. CLIP4Caption：用于视频描述的 CLIP。载于第*29*届 *ACM*多媒体国际会议论文集，第 4858–4862 页，2021。Touvron, H., Cord, M. 和 J'egou, H. DeiT III：ViT 的复仇。载于欧洲计算机视觉会议，2022。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. 和 Polosukhin, I. 注意力就是你所需要的一切。载于神经信息处理系统进展，第 5998–6008 页，2017年。Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. 和 Bowman, S. R. GLUE：用于自然语言理解的多任务基准测试与分析平台。载于*BlackboxNLP@EMNLP*，2018年。Wang, A. J., Ge, Y., Yan, R., Ge, Y., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X. 和 Shou, M. Z. 一体化：探索统一视频语言预训练。*arXiv*预印本 *arXiv:2203.07303*，2022a。Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., Xie, Y., Liu, C., Jiang, Y.-G. 和 Yuan, L. OmniVL：用于图像-语言和视频语言任务的单一基础模型。*arXiv*，abs/2209.07526，2022b。Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C. 和 Wang, L. GIT：用于视觉与语言的生成式图像到文本 Transformer。*arXiv*，abs/2205.14100，2022c。Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J. 和 Yang, H. 通过简单的序列到序列学习框架统一架构、任务和模态。载于国际机器学习会议，2022d。

Wang, W., Bao, H., Dong, L. 和 Wei, F. VLMo：通过模态专家混合实现统一视觉-语言预训练。*arXiv*预印本 *arXiv:2111.02358*，2021a。

Wang, W.、Bao, H.、Dong, L.、Bjorck, J.、Peng, Z.、Liu, Q.、Aggarwal, K.、Mohammed, O.、Singhal, S.、Som, S. 和 Wei, F. 图像即外语：面向所有视觉和视觉语言任务的 BEiT 预训练。*ArXiv*，abs/2208.10442，2022e。

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., hua Hu, X., Lu, T., Lu, L., Li, H., Wang, X., 和 Qiao, Y. In- ternimage：使用可变形卷积探索大规模视觉基础模型。*ArXiv*，abs/2211.05778，2022f。

Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., and Qiao, Y. Internvideo：通过生成式和判别式学习实现通用视频基础模型。*arXiv*，abs/2212.03191，2022g。Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm：基于弱监督的简单视觉语言模型预训练。*CoRR*，abs/2108.10904，2021b。Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. 通过对外观和运动逐步精炼的注意力进行视频问答。载于第*25*届*ACM*多媒体国际会议论文集，第1645–1653页，2017年。Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., and Huang, F. E2E-VLP：通过视觉学习增强的端到端视觉-语言预训练。载于 Zong, C., Xia, F., Li, W., and Navigli, R.（编），第*59*届计算语言学协会年会暨第*11*届自然语言处理国际联合会议论文集（*ACL/IJCNLP 2021*），第503–513页。Xu, J., Mei, T., Yao, T., and Rui, Y. MSR-VTT：用于连接视频与语言的大型视频描述数据集。载于*IEEE*计算机视觉与模式识别会议论文集，第5288–5296页，2016年。Xu, J., Mello, S. D., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. GroupViT：语义分割源自文本监督。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第18113–18123页，2022年。Yan, S., Zhu, T., Wang, Z., Cao, Y., Zhang, M., Ghosh, S., Wu, Y., and Yu, J. 通过对比式图像描述模型的零样本迁移进行视频-文本建模。*arXiv*，abs/2212.04979，2022年。

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. Just Ask：从数百万个带旁白的视频中学习回答问题。载于*IEEE/CVF*国际计算机视觉会议论文集，第1686–1697页，2021a。

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. 基于冻结双向语言模型的零样本视频问答。*arXiv*预印本*arXiv:2206.08155*，2022。

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. XLNet：面向语言理解的广义自回归预训练。载于 Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alch´e-Buc, F., Fox, E. B., and Garnett, R.（编），神经信息处理系统进展*32*：神经信息处理系统年度会议*2019*，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第5754–5764页，2019。

Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. 跨越文本与框的格式边界：迈向统一视觉语言建模。*ArXiv*，abs/2111.12085，2021b。Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J. C., and Huang, F. HiTeA：分层时序感知视频语言预训练。*ArXiv*，abs/2212.14546，2022。

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. CoCa：对比式图像描述模型是图像-文本基础模型。*ArXiv*，abs/2205.01917，2022。Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. 对指代表达中的上下文进行建模。载于欧洲计算机视觉会议，第69–85页。施普林格，2016。Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., 等。Florence：面向计算机视觉的新型基础模型。*arXiv*预印本*arXiv:2111.11432*，2021。

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., Farhadi, A., and Choi, Y. MERLOT：多模态神经脚本知识模型。神经信息处理系统进展，34，2021。Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. 视觉Transformer的规模扩展。*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1204–1213页，2021。Zhang, B., Yu, J., Fifty, C., Han, W., Dai, A. M., Pang, R., and Sha, F. 通过视频和图像协同训练Transformer提升动作识别性能。*ArXiv*，abs/2112.07175，2021a。

Zhang, J., Zhao, Y., Saleh, M. 和 Liu, P. Pegasus：使用提取的间隙句进行抽象式摘要预训练。发表于 国际机器学习会议，第 11328–11339 页。PMLR，2020年。Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y. 和 Gao, J. Vinvl：让视觉表征在视觉语言模型中发挥作用。*CoRR*，abs/2101.00529，2021b。Zhu, X., Zhu, J., Li, H., Wu, X., Wang, X., Li, H., Wang, X. 和 Dai, J. Uni-perceiver：为零样本和少样本任务的通用感知预训练统一架构。*arXiv* 预印本 *arXiv:2112.01522*，2021。Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A. 和 Fidler, S. 对齐书籍与电影：通过观看电影和阅读书籍，实现类故事的视觉解释。*2015*年 *IEEE* 国际计算机视觉会议（*ICCV*），第 19–27 页，2015年。

表*18*。 **在 COCO 数据集上，目标检测和实例分割任务与最先进结果的系统级比较**。 我们报告标准边界框 AP（AP_{box}）和掩码平均精度（AP_{mask}）。检测器包括 Cascade Mask R-CNN（级联）、Dynamic Head（DyHead）、Hybrid Task Cascade（HTC）及其扩展版本（HTC++）。

方法	检测器	AP_{box}	AP_{mask}
Mask R-CNN (He 等人，2017年)	-	46.3	40.1
DETR (Carion 等人，2020年)	-	44.9	33.0
Pix2seq(Chen 等人，2021)	-	45.0	-
复制粘贴 (Ghiasi 等人，2021)	级联	57.0	48.9
Swin-L (Liu 等人，2021b)	HTC++	58.0	50.4
CBNetV2(Liang 等人,2021)	HTC	59.6	51.8
GLIP (Li 等人, 2021c)	DyHead	60.8	-
SwinV2-L (Liu 等人, 2021a)	HTC++	60.2	**52.1**
Florence（Yuan 等人，2021）	DyHead	**62.0**	-
mPLUG-2	级联	46.9	40.6

表*19*。 **文本到视频检索上的零样本评估**。 所有结果均以 R@1/R@5/R@10 报告。

		MSRVTT			DiDeMo			LSMDC		
方法	# 预训练数据	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen (Bain 等人, 2021b)	5M	18.7	39.5	51.6	21.1	46.0	56.2	9.3	22.0	30.1
ALPRO (Li 等人, 2021a)	5M	24.1	44.7	55.4	23.8	47.3	57.9	-	-	-
Singularity（Lei 等人，2022）	5M	28.4	50.2	59.5	36.9	61.6	69.3	-	-	-
VIOLET（Fu 等人，2021）	183M	25.9	49.5	59.7	23.5	49.8	59.8	-	-	-
Florence（Yuan 等人，2021）	900M	37.6	63.8	72.6	-	-	-	-	-	-
mPLUG（Li 等人，2022a）	14M	38.1	59.2	68.2	-	-	-	-	-	-
HiTeA（Ye 等人，2022）	17M	34.4	60.0	69.9	43.2	69.3	79.0	18.3	36.7	44.2
OmniVL（Wang 等人，2022b）	18M	42.0	63.0	73.0	40.6	64.6	74.3	-	-	-
mPLUG-2	17M	**47.1**	**69.7**	**79.0**	**45.7**	**71.1**	**79.2**	**24.1**	**43.8**	**52.0**

表*20*。**视频问答的零样本评估**。报告准确率。

方法	# 预训练数据	MSRVTT-QA	MSVD-QA
Just Ask(Yang 等人, 2021a)	69M	2.9	7.5
LAVENDER (Li 等人, 2022d)	5M	4.5	11.6
MERLOT Reserve(Zellers 等人, 2021)	1B	5.8	-
FrozenBiLM(Yang 等人, 2022)	10M	6.4	11.7
BLIP(Li 等人,2022c)	129M	19.2	35.2
mPLUG (Li 等人, 2022a)	400M	21.1	37.2
HiTeA (Ye 等人, 2022)	5M	21.7	37.4
mPLUG-2	17M	**43.8**	**55.3**

A. 更多结果

A.1. 检测与分割

我们评估了 mPLUG-2 在 COCO 数据集（Lin等，2014）上的目标检测和实例分割性能。该数据集广泛用于包含 80 个常见类别的目标级检测与分割。表 18 展示了 COCO 数据集上的结果。我们发现，mPLUG-2 的表现优于典型的最先进基于 ResNet 的检测方法（例如 DETR（Carion 等人，2020年）和 Pix2seq（Chen 等人，2021））。针对计算机视觉优化的基础模型（例如 Florence（Yuan 等人，2021）和 Swin-Transformer（Liu 等人，2021b））与 mPLUG-2 之间存在性能差距。请注意，mPLUG-2 并未使用仅包含视觉任务和数据进行预训练。其性能低于在 ImageNet 上预训练的模型，是意料之中的。

A.2. 零样本迁移能力

文本到视频检索为测试预训练 mPLUG-2 的迁移能力，我们在文本到视频检索任务上进行了零样本评估，结果汇总于表19中。我们发现，mPLUG-2 在 MSRVTT、DiDeMo 和 LSMDC 数据集上均取得了最先进水平的结果，并大幅领先于以往方法，例如在 MSRVTT 数据集上的 R@1 提高了 5.1 个百分点。这些结果证明，我们的 mPLUG-2 具有出色的零样本迁移能力。



图 *5*。视觉定位任务的可视化结果。绿色表示真实框，红色表示预测边界框。

视频问答我们测试了预训练 mPLUG-2 在视频问答任务上的迁移能力，结果汇总于表20中。可以看出，mPLUG-2 在 MSRVTT-QA 和 MSVD-QA 数据集上均取得了最佳零样本性能，这表明在通用模块和基于指令的学习的帮助下，我们的模型具有很强的零样本迁移能力。

A.3. 视觉定位

图 5展示了若干视觉定位任务的示例。第一行表明，我们的 mPLUG-2 能够理解各种视觉概念及其关系，也能够实现视觉与语言之间的细粒度对齐。第二行展示了几个失败案例。在第一个示例中，“trunk”一词存在歧义，导致预测错误。在第二个示例中，mPLUG-2 未能识别模糊的“donut”。在第三个示例中，mPLUG-2 未意识到镜像中的左右方向是反的，因此预测了“left”这一项。

B. 实现细节

B.1. 预训练

我们的模型采用 PyTorch 框架（Paszke 等人，2019）实现。具体而言，文本编码器采用在维基百科和 Bookcorpus（Zhu 等人，2015）上预训练的 BERT（Devlin 等人，2018）模型初始化。视觉编码器则采用在 4 亿个噪声图文对上预训练的 CLIP-ViT（Radford 等，2021）初始化。对于基础尺寸的模型，即 mPLUG-2_{Base}，我们使用 ViT-B/16 作为视觉编码器，并使用 BERT-Base（Devlin 等人，2018）作为文本编码器和文本解码器。对于 mPLUG-2，我们分别采用 ViT-L/14（Dosovitskiy 等，2020）和 BERT-Large（Devlin 等人，2018）扩展视觉和文本编码器。$C = 768$ 和 $C = 1024$ 分别用于 mPLUG-2_{Base} 和 mPLUG-2。考虑到通用层的实证表现较好，我们将 $S = 2$设为通用层，并根据经验为局部时序建模模块中的多组机制选择 $G = C$。mPLUG-2_{Base} 的融合模块层数设为 3，mPLUG-2 的设为 6；两种模型的共享解码器层数均设为 12。我们在 8 块 NVIDIA A100 GPU 上以总批量大小 1024 对 mPLUG-2_{Base} 进行 30 个周期的预训练，并在 16 块 NVIDIA A100 GPU 上使用批量大小 512。我们采用 AdamW（Loshchilov & Hutter，2019）优化器，权重衰减系数为 0.02，betas 参数为（0.9，0.98），以稳定学习过程。学习率在前 5000 次迭代中先预热至 lr_{max}，随后按照余弦退火计划衰减。lr_{max} 对于 mPLUG-2_{Base} 设为 1e-4，对于 mPLUG-2 设为 5e-5。在

预训练期间，我们将图像和视频帧随机裁剪至 224×224 分辨率，并为每个视频稀疏采样 4 帧，同时保持帧的先后顺序。对于视觉-文本对比学习，队列大小和动量系数分别设为 65,536 和 0.995。

B.2. 下游任务

B.2.1. 纯视觉任务

视频动作识别 我们首先在 Kinetics-710 数据集上训练 mPLUG-2 40 个 epoch。该数据集由 Kinetics-400、Kinetics-600 和 Kinetics-700 合并而成，并移除了验证集和测试集中包含的视频。具体而言，mPLUG-2_{Base} 的基础学习率设为 1e-5，mPLUG-2 的基础学习率设为 5e-6，批次大小分别为 256 和 128。随后，分别在 Kinetics-400、Kinetics-600 和 Kinetics-700 上以相同的学习率和批次大小微调 5 个 epoch。

图像分类 我们以 6e-5 的学习率和 4096 的批次大小对 mPLUG-2 进行 30 个 epoch 的微调。我们使用随机裁剪、水平翻转、RandAug 和 RandErase 变换进行数据增强。

目标检测与分割 我们沿用 EVA（Fang 等人，2022b）中的相同设置，在目标检测与分割任务上训练 mPLUG-2。不同之处在于，我们在 MSCOCO 上微调之前，不会在 Object365（Shao 等人，2019年）上对 mPLUG-2 进行预训练。

B.2.2. 纯语言任务

自然语言理解 遵循（Wang 等人，2022d），我们在合适的范围内选择最佳超参数进行微调。我们调整训练轮数（5、7、10）、学习率（3e-5、5e-5、6e-5、7e-5、1e-4）和批量大小（32、64、128）。我们报告各项任务在开发集上的最佳性能。

自然语言生成 遵循（Wang 等人，2022d），我们以 3e-5 的学习率和 256 的批量大小对 mPLUG-2 微调 50,000 步。推理时，束宽设为 5，最大生成长度设为 512。

B.2.3. 视频-文本多模态任务

对于所有视频语言下游任务，我们将视频帧调整为 224×224。微调期间，对于文本到视频检索，我们随机采样 12 帧；对于视频问答和视频字幕，则随机采样 16 帧。推理期间，我们采用均匀采样。我们使用最小比例为 0.5 的随机裁剪，以及概率为 0.5 的水平翻转进行数据增强。

文本到视频检索 我们在 MSRVTT/DiDeMo/LSMDC 的训练集上训练 mPLUG-2_{Base} 和 mPLUG-2 10 个 epoch，学习率为 2e-5，批量大小为 192。

视频问答 我们在 MSRVTT-QA/MSVD-QA/TGIF- FrameQA 的训练集上训练 mPLUG-2_{Base} 和 mPLUG-2 10 个 epoch，学习率为 2e-5，批量大小为 128。

视频描述对于视频描述任务，我们使用前缀提示“视频描述了什么？”来提升生成的图像描述质量。MSRVTT 和 MSVD 数据集采用相同的训练参数。具体而言，我们在训练集上使用交叉熵损失对 mPLUG-2_{Base} 和 mPLUG-2 进行微调，训练 10 个周期，学习率为 2e-5，批量大小为 128。随后，我们进行额外 5 个周期的 CIDEr优化，学习率为 1e-6，批量大小为 16。最后，我们使用束宽为 5、最大生成长度为 25 的设置对测试集进行评估。

B.2.4. 图像-文本多模态任务

对于检索/视觉问答/图像描述/定位任务，我们将图像帧缩放至 336/576/384/336。我们使用最小比例为 0.5 的缩放裁剪和概率为 0.5 的水平翻转进行数据增强。推理时，我们进行中心裁剪。

图像-文本检索 我们在 MSCOCO/Flickr30K 的训练集上训练 mPLUG-2 8 个 epoch，学习率为 1e-5，批次大小为 512。

视觉问答 我们在视觉问答数据集上训练 mPLUG-2_{Base} 8 个 epoch，学习率为 3e-5，批次大小为 512。

图像描述 对于图像描述任务，我们使用前缀提示“图像描述了什么？”来提升生成图像描述的质量。我们首先在 COCO 训练集上使用交叉熵损失对 mPLUG-2 进行微调，训练 5 个 epoch，学习率为 1e-5，批量大小为 256。随后，我们在 COCO Caption Karpathy 验证集划分上进行评估，并直接将其用于预测 Nocaps 验证集。推理时，我们使用束搜索，束宽设为 5，最大生成长度设为 25。

视觉定位 我们首先使用 RefCOCO 系列数据集训练模型 120 个 epoch，学习率为 2e-5。然后，我们继续在每个数据集上微调模型 30 个 epoch，学习率为 2e-6。对于 RefCOCO 和 RefCOCOg，我们分别将查询长度限制为 20/40。

B.3. 数据集描述

文本到视频检索 我们在三个常用的文本到视频检索数据集上评估 mPLUG-2，包括 MSRVTT（Xu 等人，2016年）、DiDeMo（Anne Hendricks 等人，2017年）和 LSMDC（Rohrbach 等人，2015年）。

- **MSRVTT** 由 1 万个来源于 YouTube 的视频和 20 万条文本描述组成。遵循（Li 等人，2022d；Luo 等人，2022；Huang 等人，2022a）的做法，该数据集被划分为 9K 个视频用于训练，1K 个视频用于测试。
- **DiDeMo** 由来自 Flickr 的 10K 个视频组成，每个视频配有 4 条描述。遵循（Li 等人，2022b；Ma 等人，2022；Li 等人，2022d）的做法，我们将一个视频的所有描述拼接成一个段落，并评估段落到视频的检索性能。该数据集被划分为 8K 个视频用于训练、1K 个视频用于验证集，以及 1K 个视频用于测试。
- **LSMDC** 包含来自 202 部电影的 118,081 个视频片段。按照（Rohrbach 等人，2015年）中的标准划分，该数据集分为 101K 个视频用于训练，1K 个视频用于测试。

视频问答 我们在三个热门的视频问答数据集上评估 mPLUG-2，包括 MSRVTT-QA（Xu 等人，2017年）、MSVD-QA（Xu 等人，2017年）和 TGIF-FrameQA（Jang 等人，2017年）。

- **MSRVTT-QA** 基于 MSRVTT 数据集（Xu 等人，2016年）。问答对根据描述自动生成。该基准测试包含针对 10K 个视频的 243K 个开放式问题。
- **MSVD-QA** 基于 MSVD 数据集（Chen 和 Dolan，2011年），并使用自动生成的问答对。它包含 2K 个视频和 47K 个问题。
- **TGIF-FrameQA** 收集了仅凭视频中的单帧即可回答的问题，并划分为训练集（包含 3.5 万个问题）和测试集（包含 1.4 万个问题）。

视频描述 我们使 用 MSRVTT（Xu 等人，2016年）和 MSVD（Chen 和 Dolan，2011年）评估视频描述 n.

- **MSRVTT** 由 1 万个视频组成，每个视频配有 20 条图像描述，如上所述。我们采用与文本到视频检索任务相同的数据划分。
- **MSVD** 包含 1,970 个 YouTube 短视频片段。遵循（Lin 等人，2022；Li 等人，2022d）中的标准划分，我们将数据集分为 1,200 个训练视频、100 个验证视频和 670 个测试视频。

Visu类问答 我们在 VQA 2.0 数据集上评估我们的方法（Agrawal等，2017年）。

- **VQA 2.0** 是一个包含关于图像的开放式问题的数据集，每张图像至少有 3 个问题（平均 5.4 个）。该数据集包含 8.3 万/4.1 万/8.1 万张图像，分别用于训练/验证集/测试。

图像-文本检索 我们使用两个广受欢迎的图像-文本检索基准测试：COCO（Lin 等人，2014）和 Flickr30K（Plummer 等人，2015），以评估模型。对于 COCO 和 Flickr30K，我们均采用广泛使用的 Karpathy 划分（Karpathy 和 Fei-Fei，2015年）。

- **COCO** 包含超过 33 万张图像，并为每张图像提供了 5 条由不同人员独立撰写的图像描述。训练集、验证集和测试集分别包含 113k、5k 和 5k 张图像。
- **Flickr30K**包含来自 Flickr 的 31k 张图像，每张图像配有 5 条人工标注的句子。训练集、验证集和测试集分别包含 29k、1k 和 1k 张图像。

图像描述 我们在 COCO（Lin等，2014年）数据集上评估我们的方法。

- **COCO**采用与图像-文本检索任务相同的数据划分。

自然语言理解 为了验证我们的 mPLUG-2 的自然语言理解能力，我们从 GLUE（Wang 等人，2018年）基准测试中选取了 6 个语言理解数据集，其中包括单句分类任务和句子对分类任务。

- **SST-2** 斯坦福情感树库包含电影评论中的句子及人工标注的情感。任务是预测给定句子的情感。

- **RTE**文本蕴涵识别数据集源自一系列年度文本蕴涵挑战赛。
- **MRPC** 微软研究院释义语料库由从在线新闻来源收集的句子对构成，并由人工标注句对中的句子在语义上是否等价。
- **QQP** Quora 问题对数据集收录了来自社区问答网站 Quora 的问题对。任务是预测一对问题在语义上是否等价。

- **MNLI**多体裁自然语言推理语料库包含句子对（前提、假设）及其文本蕴涵标注。任务是预测前提与假设之间是否存在蕴涵关系。
- **QNLI** 斯坦福问答数据集是一个问答数据集，其中段落中的一个句子（摘自维基百科）包含对应问题（由标注者撰写）的答案。任务是判断上下文句子是否包含问题的答案。

自然语言生成 **任务** 我们使用 Gigaword 数据集（Rush 等人，2015a）进行文本摘要任务，以验证 e
自然语言生成 能力。我们的 mPLUG-2

- **Gigaword** 基于 Gigaword 语料库中的文章对进行标题生成。该语料库包含约 400 万篇文章。训练集、验证集和测试集分别包含 3803957、189651 和 1951 个样本。

视频动作识别 我们采用三个常用的基准测试：Kinetics 400/600/700 数据集（Kay 等人，2017年），以评估我们的模型。

这三个基准测试中的视频均采集自 YouTube。每段视频片段时长约为 10 秒，并标注一个动作类别。视频内容包括演奏乐器等人-物交互，以及握手、拥抱等人与人之间的互动。

- **Kinetics 400**包含 24 万个训练视频和 2 万个验证视频，涵盖 400 类人类动作 es.
- **Kinetics 600**包含 39.2 万个训练视频和 3 万个验证视频，涵盖 600 类动作。
- **Kinetics 700**包含 54.5 万个训练视频和 3.5 万个验证视频，涵盖 700 类动作。

图像分类 我们在 ImageNet-1K 上评估 mPLUG-2 的图像分类性能（Deng 等人，2009）。

- **ImageNet-1K**包含 128 万张训练图像和 5 万张验证图像，分属 1,000 个类别。