从多模态预训练到多模态大模型：架构、训练、评测、趋势概览

李泽君[1†], 张霁雯[1†], 王晔[1†], 杜梦飞[1], 刘晴雯[1],
王殿仪[1], 吴斌浩[1], 罗瑞璞[1], 黄萱菁[1], 魏忠钰[1*]
[1]复旦大学
zywei@fudan.edu.cn

摘要

多媒体信息在人类社会的发展历程中有着至关重要的作用，构建具有多模态信息处理能力的智能系统也是通往通用人工智能的必经之路。随着预训练技术的发展以及对于通用模型的需求，多模态的研究也从早期的任务特定的方法转移到了构建统一泛用的多模态基座模型上。初步的统一多模态模型探索受到BERT启发，从表征学习的角度出发构建能为不同下游任务提供有效初始化的多模态预训练模型，这类方法尽管有效但仍然在泛用性方面受限于预训练-微调范式，无法更广泛高效地应用。近年来随着大语言模型的发展，以大语言模型为基座的多模态大模型则展现出了巨大的潜力：此类模型有着强大的信息感知，交互，以及推理能力并且能有效泛化到多样的场景下，为新时代的通用人工智能系统提供了切实可行的思路。本文将从构建统一多模态模型的角度出发，介绍和梳理相关工作的发展，从多模态预训练到多模态大模型，介绍对应的架构，训练，评测方法以及发展趋势，为读者提供一个全面的概览。

关键词： 多模态学习 ； 多模态预训练 ； 多模态大模型

From Multi-Modal Pre-Training to Multi-Modal Large Language Models: An Overview of Architectures, Training, Evaluation, and Trends

Zejun Li[1]*†, Jiwen Zhang[1]*†, Ye Wang[1]*†, Mengfei Du[1], Qingwen Liu[1], Dianyi Wang[1], Binhao Wu[1], Ruipu Luo[1], Xuanjing Huang[1], Zhongyu Wei[1*]
[1]Fudan University
zywei@fudan.edu.cn

Abstract

Multimedia information has played a crucial role in the development of human society, and constructing intelligent systems with multi-modal information processing capabilities is an essential pathway towards achieving Artificial General Intelligence (AGI). With the advancement of pre-training techniques and the growing demand for general models, research in multi-modality has shifted from early task-specific approaches to constructing unified and generalizable multi-modal foundation models. Early explorations of unified multi-modal models were inspired by BERT and focused on representation learning to create multi-modal pre-trained models that provide effective initialization for various downstream tasks. However, these methods are still limited in generalization by the pre-training and fine-tuning paradigm, making them less efficient for broad applicability. In recent years, the development of Large Language

[†]共同一作
[*]通讯作者

Models (LLMs) has shown immense potential for Multi-modal Large Language Models (MLLMs). These models possess strong capabilities in perception, interaction, and reasoning, and can generalize effectively to diverse scenarios, offering a feasible approach to building next-generation AGI systems. This paper will provide an overview of constructing unified multi-modal models, introducing and reviewing the development from multi-modal pre-training to MLLMs. It will cover corresponding architectures, training paradigms, evaluation methods, and development trends, offering readers a comprehensive overview.

Keywords: Multimodal Learning , Multimodal Pre-training , Multimodal Large Language Model

1 引言

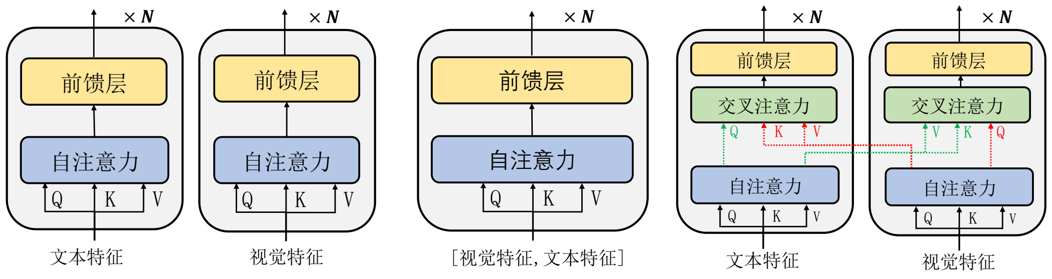
在人类社会的发展过程中，信息的交流与表达形式不断丰富和演变。从最初的语言交流到文字记录，再到现代的多媒体传播，多模态信息处理能力已成为现代社会不可或缺的一部分。随着信息技术的飞速发展，人们对于能够理解和生成结合文本、图像、声音等多种模态信息的智能系统的需求日益增长。多模态研究的重要性不仅体现在提高机器的交互能力，更在于推动社会信息交流方式的革新，促进知识的传播和文明的交流，构建具有多模态感知交互能力的智能体也是通往通用人工智能的必经之路 (Gan et al., 2022; Cui et al., 2024; Jin et al., 2024; Zhang et al., 2024)。

早期的多模态研究主要聚焦于针对特定的任务构建特定的模型，比如针对图片字幕生成任务 (Xu et al., 2015; Anderson et al., 2018a; Fan et al., 2019; Fan et al., 2021)，视觉问答任务 (Antol et al., 2015; Yu et al., 2019)，图文检索设计的架构 (Lee et al., 2018; Fan et al., 2022; Li et al., 2021b)等等。这样的方法往往会引入任务特定，甚至是数据集特定的归纳偏置，难以适应新的任务或泛化到不同的应用场景上。这限制了多模态技术在更广泛领域的应用潜力，也进一步启发了关于如何构建统一且泛用的多模态模型的研究。

受启发与BERT为代表的文本预训练模型 (Devlin et al., 2019; Liu et al., 2019)，最早对于统一多模态模型的探索聚焦在构建多模态预训练模型上。从表征学习的角度出发，利用大量的视觉文本数据进行自监督学习，在此基础上预训练模型提供的跨模态对齐的表示能够为不同的任务提供有效的初始化 (Li et al., 2019; Su et al., 2020; Chen et al., 2020b; Tan and Bansal, 2019; Li et al., 2021a; Radford et al., 2021; Li et al., 2022a)。然而预训练模型往往需要在特定任务上的微调才能进行实际的应用，并且很难有效地建立起任务之间地相关性，进而几乎无法以零样本的方式泛化到新的任务形式上。

随着ChatGPT为代表的大语言模型的出现 (OpenAI, 2023a; Touvron et al., 2023a; Touvron et al., 2023b)，研究者们发现指令微调后的大语言模型能够具有强大的对话，推理能力，并且能泛化到不同场景、任务下的不同指令 (Ouyang et al., 2022a; Chung et al., 2022; Chiang et al., 2023)，这为构建通用人工智能系统提供了切实可行的思路。多模态大模型的研究则尝试将大语言模型的成功迁移到多模态的领域下，通过给语言基座模型扩充了多个模态的编码器 (Girdhar et al., 2023; Radford et al., 2021; Zhai et al., 2023b)，并以高效的训练将其他模态的信号对齐到语言基座中 (Liu et al., 2024c; Li et al., 2023d; Bai et al., 2023b; Zhu et al., 2023a)。不同于预训练模型，得益于大语言模型的发展，多模态大模型的参数规模得益有效地扩展；除此之外，多模态大模型同样使用指令微调的方法进行训练，以自然语言的方式构建起了任务间的关联，进一步能泛化到不同的指令和场景。目前的多模态大模型能够准确地感知其他模态的信息，并根据多模态的语境和用户以自然语言进行交互，理解用户的需求并完成对应的任务 (Chen et al., 2023b; Zhao et al., 2023b; Liu et al., 2023c; Young et al., 2024; Laurençon et al., 2024; Lu et al., 2024)。近来，多模态大模型在各方面的应用和发展也展现了其成为统一化的多模态信息处理基座的潜力 (Li et al., 2023j; Chen et al., 2023e; Wu et al., 2023a; Zheng et al., 2023)。

本文将从构建统一多模态模型的角度出发，介绍和梳理相关工作的发展。在第 2 节中我们将首先介绍早期的多模态预训练的方法。接下来，我们将聚焦于热门的多模态大模型，从多模



(a) 双塔模型 (b) 单塔模型(单流) (c) 单塔模型(双流)

Figure 1: 常见的VLP模型架构。其中的Q，K，V分别表示注意力层的Query，Key，Value输入。“[视觉特征，文本特征]”表示两种特征拼接后得到的特征。

态大模型的架构（第 3 节），多模态大模型的训练方法（第 4 节）以及多模态大模型的评估方法（第 5 节）出发进行介绍。最后，本文将在第 6 节探讨多模态大模型的新发展方向，为读者提供一个全面的研究概览。

2 多模态预训练模型

在ChatGPT (OpenAI, 2023a) 引领的大模型风潮到来之前，多模态领域的研究者主要聚焦于多模态预训练技术(Vision-Language Pre-Training，简称VLP)：尝试通过预训练的方式构建统一、泛用的多模态表征模型。通过将大量的视觉文本数据输入到基座的多模态表征模型中，并通过自监督的预训练任务帮助模型学习跨模态的信息交互，所得到的预训练模型能够为下游多样化的多模态任务提供有效的初始化表示 (Tan and Bansal, 2019; Li et al., 2020c; Radford et al., 2021; Li et al., 2022c; Wang et al., 2022c; Bao et al., 2022a; Chen et al., 2023d)。

2.1 模型架构

受启发自以BERT (Devlin et al., 2019; Liu et al., 2019)为代表的文本预训练模型，多模态预训练的架构也主要构建在多层Transformer (Vaswani et al., 2017)的基础上，通过将视觉文本数据构建成序列，并对多模态序列学习语境化的表示。

2.1.1 多模态输入表示

文本表示 对于文本的输入，多数VLP模型遵循BERT的范式 (Devlin et al., 2019)，通过将文本进行分词，并在其开头和结尾补充“[CLS]”和“[SEP]”的特殊字符来表示文本的开头和结束。所得到的字符序列通过可学习的词嵌入层得到连续化的向量表示。

视觉表示 对于视觉输入，则需要额外的视觉编码器来对输入信号进行编码。早期的VLP方法通过物体检测模型来构建物体级别的视觉token，每个token的表示为对应物体所在区域的RoI (Region-of-Interest)的特征 (Zhang et al., 2021b; Chen et al., 2020b; Li et al., 2020c; Tan and Bansal, 2019)。其中最常用的检测器为Anderson et al. (2018a)训练的Faster-RCNN (Ren et al., 2015)。由于物体检测器的参数无法在后续的预训练过程中得到更新，后续的工作专注于构建端到端的VLP模型 (Li et al., 2020a; Xu et al., 2021; Huang et al., 2020)，其中最常使用的编码器为CNN类模型 (He et al., 2016)，以及视觉Transformer类模型 (Dosovitskiy et al., 2021; Liu et al., 2021)。这类视觉编码器得到的二维网格特征会被展开成一维，得到视觉的表示特征。视频则可以看作是多张图片构成的序列，并可以进一步用三维的模型建模时序的信息 (Feichtenhofer et al., 2019; Zhang et al., 2018; Carreira and Zisserman, 2017)。

2.1.2 多模态交互的建模方式

基于第 2.1.1 节中构建的多模态序列，VLP模型以不同的架构来建模模态间和模态内的信息交互，如图 1 所示，大致可以划分为如下几类：

双塔模型 双塔模型通过浅交互的方式来建模多模态的交互。视觉特征序列和文本序列分别被单独的视觉编码器和文本编码器进行编码，最终通过对比学习的方式进行整体的图片和文本间的对齐。最具代表性的工作有CLIP (Radford et al., 2021)和ALIGN (Jia et al., 2021)。单塔模

型能够高效地以内积的形式计算多张图片和文本之间的语义相似度，在跨模态检索任务上效率非常高，但是无法解决需要复杂的跨模态理解能力的任务，比如视觉语言问答。

单塔模型 单塔模型则希望构建视觉和文本token之间的深交互，主要利用Transformer中的注意力机制来进行建模。其中如图 1 (b, c) 所示，单流模型将视觉和文本序列拼接到一起，通过自注意力层学习模态内和模态间的交互 (Li et al., 2019; Chen et al., 2020b; Su et al., 2020; Li et al., 2020c; Kim et al., 2021)；而双流模型则解耦了模态内和模态间的表示，单独用交叉注意力来建模后者 (Lu et al., 2019; Tan and Bansal, 2019; Yu et al., 2021; Dou et al., 2022b)。单塔模型具有很强的跨模态建模和推理能力，但是在运算效率上稍逊于双塔模型。

其他架构 除了上述两种常规的架构，为了结合单塔和双塔结构的优点，研究者提出了融合型的架构，ALBEF (Li et al., 2021a)，FLAVA (Singh et al., 2022)和CoCa (Yu et al., 2022)在双塔模型的上层引入深交互的单塔融合编码器 (Xu et al., 2023c)；BLIP (Li et al., 2022b)，FIBER (Dou et al., 2022a)和VLMO (Bao et al., 2022a)等则设计了动态的交叉注意力或前馈层，使得同一模型可以支持多种形式的建模并且支持参数的高效共享 (Wang et al., 2021b)。为了支持文本生成的输出形式，目前的工作也探究了解码器 (Wang et al., 2022a)，编码器-解码器 (Cho et al., 2021; Wang et al., 2022c; Chen et al., 2023d)的VLP的架构。

2.2 训练方法

2.2.1 预训练数据集

多模态预训练要求大量视觉文本数据，图文对是最常见的形式，从高质量的标注数据，包括MSCOCO (Lin et al., 2014)，Flickr30K (Young et al., 2014)和Visual Genome (Krishna et al., 2017)，到网络爬取的图文对，CC3M (Sharma et al., 2018)，SBU (Ordonez et al., 2011)，CC12M (Changpinyo et al., 2021)，RedCaps (Desai et al., 2021)，再到更大规模的LAION (Schuhmann et al., 2021)等等。对于视频文本数据，则常用的数据为Kinetics-400 (Kay et al., 2017)，HowTo100M (Miech et al., 2019)以及WebVid-2M (Bain et al., 2021)。一般而言，数据集的规模和数据质量之间存在负相关的关系，需要综合考量。

2.2.2 预训练任务

多模态预训练任务主要为自监督的任务，通过不同的形式学习视觉文本的相关联系。

遮盖恢复建模 此类任务受启发于BERT中提出的遮盖语言模型（Masked Language Modeling，简称MLM）任务 (Devlin et al., 2019)，随机遮盖部分输入序列，要求模型预测原本的信息来学习语境的建模能力。对于文本，MLM也是VLP方法最常用的预训练任务 (Su et al., 2020; Li et al., 2019)。对于视觉信号的遮盖，基于物体检测输入的模型训练目标为预测被遮盖物体的类别和输入特征 (Chen et al., 2020b; Tan and Bansal, 2019)；SOHO (Huang et al., 2021)，VL-BEiT (Bao et al., 2022b)和BEiT-3 (Wang et al., 2023e)则利用视觉字典的方法给图片块构建了离散的表示，使得其也可以和MLM一样以分类的形式训练。

文本生成任务 不同于MLM进行双向的语境建模，文本生成任务要求模型进行单向的逐词的生成 (Li et al., 2022a; Wang et al., 2022a; Yu et al., 2022)。不同于MLM任务在每个样本里仅学习被遮盖部分的信息，文本生成任务可以利用每个token的信息，训练效率更高。Wang et al. (2022c)提出PrefixLM来让模型同时学习双向和单向的语境信息。VL-BART (Cho et al., 2021)和OFA (Wang et al., 2022b)进一步通过文本生成的形式统一表示多种任务。

图文对齐任务 用于帮助模型对齐模态间的表示，主要的任务有图文对比学习和图文匹配。前者主要作用在单塔模型部分 (Radford et al., 2021; Li et al., 2021a)，使用InfoNCE损失 (van den Oord et al., 2019)在训练批次内进行对比学习；后者则作用在融合模型部分，要求模型进行图文对是否匹配的二分类判断，其中的负样本可以来自于随机的采样 (Chen et al., 2020b)，批次内的困难负样本 (Li et al., 2021a)，数据集的困难负样本 (Chen et al., 2020a)。除了全局的对齐，也有工作 (Chen et al., 2020b; Kim et al., 2021)尝试用最优传输的理论对齐局部的token。

其他训练目标 除了上述的优化目标，视频还可以利用时序的信息进行自监督训练，比如视频帧的排序建模任务 (Li et al., 2020b)。除此之外，视觉问答 (Tan and Bansal, 2019; Li et al., 2020c)，物体标记框预测 (Xu et al., 2021; Zeng et al., 2022)也常被作为训练的任务。

2.3 多模态预训练模型的评测

多数的VLP模型仅仅作为初始化的表征参数，需要经过下游的微调后才能进行评测，并且需要在预训练模型的基础上增加任务相关的层来满足任务所需的输出形式。常见的任务包括：（1）跨模态检索任务：要求模型给定某模态内的一个样本从另一模态的候选集选择与其最语义匹配的样本，包括图文检索 (Lin et al., 2014; Young et al., 2014)，视频检索 (Xu et al., 2016; Chen and Dolan, 2011)等等；（2）视觉推理任务：需要模型根据对应的问题和视觉信息进行推理，包括视觉问答（VQA）(Goyal et al., 2017; Hudson and Manning, 2019; Xu et al., 2017; Jang et al., 2017)，推理 (Xie et al., 2019; Suhr et al., 2017; Zellers et al., 2019)，等等；（3）视觉描述任务：要求模型用语言描述对应的视觉信号 (Agrawal et al., 2019; Wang et al., 2019)。

尽管VLP模型能够对多种下游任务提供良好的初始化表示，预训练微调的范式也有明显的局限性：模型需要微调才能进行应用，并且无法适应没见过的任务形式，极大地限制了模型的泛用性。而VL-BART (Cho et al., 2021)和OFA等模型 (Wang et al., 2022b)中通过文本描述不同任务的方法则展现出了巨大的潜力，也成为了本文后续论述的多模态大模型的雏形。

3 多模态大模型架构

随着以ChatGPT为代表的大模型时代到来，多模态序列建模逐渐成为研究热点。当今的多模态大模型主要关注于多模态内容的理解和文本生成。这些模型通常将输入的模态信号进行抽象编码，形成一系列模态特征标记的序列表示，随后传递到大语言模型中进行统一处理，并解码生成文本。本章将从多模态序列表示开始，逐步介绍多模态大模型的具体架构设计以及架构优化方案。

3.1 多模态序列表示

一般当下多模态大模型主要关注多模态内容理解和文本生成。通常将输入的模态信号进行抽象编码成一系列模态特征标记的序列表示，然后再传递到大语言模型中进行统一的处理并解码出响应文本。建模数据格式为:Text + X− > Text。其中输入数据为文本和多模态内容对X(可以是图像文本对，音频文本对，视频文本对等)，输出是文本响应。

- 文本序列表示:按照语言模型的词表将输入文本分成一系列的文本标记。
- 图像序列表示:将输入图片切成M ∗ M的图像补丁通过视觉编码器变成一系列视觉标记。
- 视频序列表示:如果输入是视频，和图像方法一致，将视频按帧抽出图片表示，将每帧图片切成M ∗ M的图像补丁通过视觉编码器变成一系列视觉标记。并将每帧所有标记进行拼接。此外，会额外加一个表征时序信息的序列标记。而Valley(Luo et al., 2023b)则在每帧的视觉标记上作时间维度的平均，来保留空间信息。
- 音频序列表示:使用固定大小窗口进行滑动，将窗口内音频信号通过编码器进行编码处理，每个窗口的编码结果作为一个输入标记，从而得到一系列音频的序列标记表示。

3.2 多模态大模型的基座模型

典型的多模态大模型的基座模型一般可以抽象为三个模块，即预训练的视觉编码器、预训练大语言模型和模态连接件。预训练的视觉编码器作用是将输入的视觉信号编码成抽象的特征表示，类比于人类的眼睛，用于接收和预处理视觉输入。而大语言模型作为中央大脑，管理接收到的输入模态信号并理解和执行推理。视觉输入特征和大语言模型的文本特征相差较大，难以被大语言模型直接理解处理，而模态连接件充当一个桥梁来连接视觉编码器和大语言模型，使不同模态信息得到对齐。一些多模态大模型还包括一个生成器来输出除文本之外的其他模态信息，具体请参阅第 6.2 节中的讨论。

3.2.1 视觉编码器

视觉编码器(Vision Encoder，简称VE)的作用是将原始的图像输入$\mathrm{X_v}$进行编码压缩成紧凑的视觉特征表示$\mathrm{F_v}$，公式如下:

$$\mathrm{F}_v = \mathrm{VE}(X_v) \tag{1}$$

预训练视觉编码器一般为ViT(Dosovitskiy et al., 2020)架构。比如常用的CLIP (Radford et al., 2021)通过图像-文本对进行大规模预训练，将视觉编码器在语义上与文本特征实现了对齐，因此使用这种最初预对齐的编码器通过预训练与大语言模型进行对齐会更加容易。除了常用的CLIP(Radford et al., 2021)视觉编码器，一些工作还探索了使用CLIP(Radford et al., 2021)的其他变体。MiniGPT-4(Zhu et al., 2023b) 、CogVLM(Wang et al., 2023d)采用EVA-CLIP (Sun et al., 2023b)编码器，该编码器在训练技术方面进行改进。SigLIP(Zhai et al., 2023b)改进了图像文本预训练损失来得到更好的对齐效果，被近期性能较强的多模态大模型广泛使用(Hu et al., 2023b; Lu et al., 2024; He et al., 2024; Laurençon et al., 2024; Team et al., 2024)。ImageBind(Girdhar et al., 2023)还将视觉、文本、音频和深度图等进行了对齐，扩展了输入模态的类别。

一些工作使用卷积架构作为视觉编码器，Osprey(Dehghani and Trojovskỳ, 2023)，ConvLLaVA(Ge et al., 2024)使用基于卷积的ConvNext-L 编码器(Liu et al., 2022b)来获取更高分辨率信息和多级特征，在相同分辨率下ConvNext(Liu et al., 2022b)生成的视觉特征标记更少，计算成本下降显著。此外，Fuyu-8b(Bavishi et al., 2023)探索了无编码器的结构，将图像块在输入到大语言模型之前直接投影，模型自然支持灵活的图像分辨率输入。

除了无编码器和单一编码器，近期一些工作还采用了双编码器设计 (Hong et al., 2023; Li et al., 2024)，使用两个编码器分别处理高分辨率图像和低分辨率图像。双编码器设计除了有助于支持高分辨外，还能够利用不同视觉编码器的归纳偏差进行互补，这有助于捕获更广泛的视觉表示，从而增强模型对视觉数据的理解。例如，Cobra(Zhao et al.,)和Deepseek-VL(Lu et al., 2024)分别采用了DINOv2(Oquab et al.,)、SAM-B(Kirillov et al., 2023)和SigLIP(Zhai et al., 2023b)的双编码器设计。这些模型通过结合DINOv2和SAM自监督学习的低级空间特征以及SigLIP通过弱监督提供的高级语义属性，实现了性能的互补增强。类似地，SPHINX-X(Gao et al., 2024)采用了DINOv2(Oquab et al.,)和CLIP-ConvNeXt(Cherti et al., 2023)的组合。Prismatic(Karamcheti et al., 2024)的研究进一步证明了SigLIP(Zhai et al., 2023b)和DINOv2(Oquab et al.,)可以实现最佳的互补效果。

3.2.2 大语言模型

通过对网络庞大语料库的自监督预训练，大语言模型嵌入了丰富的语言世界知识，拥有强大的泛化涌现、问答推理以及指令遵循能力。多模态大模型以大语言模型为骨干，大语言模型作为中央处理器统一接收视觉语言模态的特征入并统一解码推理，输出响应。如公式1,2。

目前，大多数的大语言模型属于因果解码器类别。LLAMA系列(Touvron et al., 2023a; Touvron et al., 2023b)和Vicuna(Chiang et al., 2023)是最具代表性的开源大语言模型，支撑了经典多模态大模型工作LLaVA系列(Liu et al., 2024c; Liu et al., 2023c; Liu et al., 2024b)、MiniGPT-4(Zhu et al., 2023b)等的研究。然而，这些模型主要在英语语料库上进行预训练，因此不支持多语言处理。双语大语言模型如Qwen(Bai et al., 2023a)和InternLM(Cai et al., 2024)则能很好地支持中文和英语，分别成为了Qwen-VL(Bai et al., 2023b)和多模态书生系列(Zhang et al., 2023a; Dong et al., 2024b)的大语言模型骨干。其他如Mistral(Jiang et al., 2023a)、Yi(Young et al., 2024)和Deepseek(Bi et al., 2024)等强大开源大语言模型也被用于构建新近推出的多模态模型，如Mini-Gemini(Li et al., 2024)、Yi-VL(Young et al., 2024)、LLaVA-next(Liu et al., 2024b)、IDEFICS2(Laurençon et al., 2024)和Deepseek-VL(Lu et al., 2024)等。

此外，对大语言模型专家混合体系结构(MoE)的研究也引起了广泛关注。与密集模型相比，MoE的稀疏架构可以通过选择性地激活部分参数，在不增加计算成本的情况下扩大模型总参数大小。在多模态大模型领域，如MM1(McKinzie et al., 2024)和MoE-LLaVA(Lin et al., 2024)将MoE机制引入并实现在多数基准测试中超越了相应的密集模型架构。Mini-Gemini(Li et al., 2024)拥有一个基于Mixtral 8x7B(MistralAITeam, 2023)构建的MoE版本。

同时，与动辄参数大小上百亿的多模态大模型相比，另一条路线是开发参数量更少、但性能不减的高效多模态大模型，这些模型通常使用参数少于3B的语言模型作为骨干。如MiniCPM-V系列(Hu et al., 2023b)和PaliGemma(Team et al., 2024)分别使用MiniCPM-2.4B(Hu et al., 2024b)和谷歌的Gemma-2B(Team et al., 2024)。而微软的Phi-3-V(Abdin et al., 2024)使用了拥有接近于Mixtral 8x7B (MistralAITeam, 2023)模型相当性能的Phi-3-mini(Abdin et al., 2024)。除了利用预训练的大语言模型外，MobileVLM系列(Chu et al., 2023; Chu et al., 2024)缩小了LLAMA(Touvron et al., 2023b)的参数尺寸，并使用开源数据集从头开始训练。

3.2.3 模态连接件

模态连接件(Projector)的作用是将视觉编码器输出的抽象视觉特征F_v映射到与大语言模型中的词嵌入语义空间具有相同维度的标记Z_v完成对齐，公式如下:

$$Z_v = \text{Projector}(F_v) \tag{2}$$

- **基于MLP架构**：多模态大模型的模态连接件通常使用可学习的线性层或多层感知器(MLP) 来实现。使用MLP架构的多模态大模型代表作是LLaVA系列(Liu et al., 2024c; Liu et al., 2023c; Liu et al., 2024b)。MLP架构的模态连接件的优点是可以较好的建模视觉特征的上下文信息，但缺点是视觉表示冗长，表征效率不高。

- **基于Attention架构**：BLIP-2(Li et al., 2023d)引入了Q-Former，作为一种轻量级的Transformer使用一组可学习的查询向量，从冻结的视觉编码器中提取视觉特征。后续Qwen-VL(Bai et al., 2023b)和MiniCPM-V(Hu et al., 2023b)系列同样使用了基于注意力架构的重采样器(Perceiver Resampler)作为连接件，可学习的查询向量作为Q，视觉编码器输出的图像特征为K和V，通过交叉注意力计算，输出视觉特征的聚合表示。

- **基于CNN架构**：MobileVLMv2(Chu et al., 2024)提出了LDPv2,一种基于卷积架构的新型模态连接件，包含特征变换、标记压缩和位置信息增强三个组件。通过使用逐点卷积层来匹配大语言模型的嵌入空间维度，平均池化层压缩视觉标记数量以及PEG模块来增强位置信息。较先前版本减少了99.8% 的参数量，实现更好的效率。

- **混合架构**：Honeybee(Cha et al., 2023)使用两个模态连接件，分别是C-Abstractor 和D-Abstractor。C-Abstractor为卷积架构由ResNet模块组成，专注于建模视觉特征的上下文，其中自适应平均池化有助于灵活管理视觉标记的数量。D-Abstractor利用可变形的注意力架构用于维护视觉特征的局部上下文。

另一条路线不仅仅使用模态连接件将视觉编码器和大语言模型进行模态对齐。而是在大语言模型内部插入额外参数模块以实现文本特征和视觉特征的交互融合。比如，Flamingo(Alayrac et al., 2022)、Open-Flamingo(Awadalla et al., 2023)和MMGPT(Gong et al., 2023) 在大语言模型内部层与层之间插入额外的交叉注意力层，将重采样器输出的视觉特征标记作为查询，与新插入到大语言模型中的层计算交叉注意力，从而将视觉信息注入到文本的生成过程中。CogVLM(Wang et al., 2023d)在每个Transformer 层中插入视觉专家模块，视觉专家模块包含QKV权重矩阵和FFN层，参数初始化来自预训练大语言模型，以实现视觉语言的双重交互融合。LLaMA-Adapter系列(Zhang et al., 2023b; Gao et al., 2023a)将可学习的提示引入Transformer 层。这些提示结合了视觉特征，作为前缀与文本特征进行拼接。

3.3 多模态大模型架构优化

尽管经典的多模态大模型已经能在各类通用任务上取得优异的表现 (Lin et al., 2014; Agrawal et al., 2015; He et al., 2015; Kafle and Kanan, 2017)，但在处理PDF文档、4K视频等对分辨率要求高、对时间/空间建模能力强的场景下仍然具有很大的不足 (Mathew et al., 2021; Li et al., 2023i; Ma et al., 2023; Liu et al., 2023f)。目前主流的开源模型，如LLaVA、QWen-VL等 (Liu et al., 2024c; Bai et al., 2023b)多数采用一个低分辨率的视觉编码器，其输入图像分辨率从224x224 (Dosovitskiy et al., 2021; Chen et al., 2020a; Sun et al., 2023b)、336x336 (如CLIP-ViT-Large-336 (Radford et al., 2021))到448×448 (如InternViT-1.2 (Chen et al., 2023e))不等。受限于视觉编码器的分辨率，常见的图像处理方案是：不论输入图片的原始分辨率是多少，统一缩放到与视觉编码器一致的分辨率，如LLaVA-1.5 (Liu et al., 2023c)、MiniGPT-4 (Zhu et al., 2023a)、EMU2 (Sun et al., 2023a)、InternLM-XComposer (Zhang et al., 2023a)等。显然，这种方案对信息的压缩程度过大，往往因缺失图像细节而降低模型的跨模态理解能力 (Li et al., 2023a; Ye et al., 2023a)，甚至导致幻觉 (Yu et al., 2023b)。

为了提高多模态模型的视觉编码能力，Monkey (Li et al., 2023j)等模型根据视觉编码器的分辨率大小将图像分片，每一片都由视觉编码器编码，然后拼接到一起作为图像的高分辨率特征。除了分片后的局部视图，通常还会有一个全局视图，即将原始图像放缩到视觉编码器对应

分辨率获得的低分辨率特征。全局低分辨率特征与局部高分辨率特征进行拼接，组成图片的完整特征表示。这就是多模态大模型的**多视图视觉表征**。这样的视觉表征方式虽然保留了更多的细粒度图像信息，但是拼接后的视觉序列过长，不但压缩了文本的表示能力，而且会造成训练困难 (Xu et al., 2024)。为了更有效率的进行图像特征表示，往往需要对图像特征序列进行压缩 (Ye et al., 2023a; Ye et al., 2023c; Yu et al., 2024)。最近的工作在图像切片的基础上衍生出了三种高分辨率的视觉编码方案，分别是：

1. **基于Resampler的视觉特征压缩**：得益于多模态预训练模型在早期对模型架构的探索，由Flamingo (Alayrac et al., 2022)所提出的Perceiver Resampler和BLIP (Li et al., 2022a)所提出的Q-Former架构天然地适用于视觉表征压缩。因此，以UReader (Ye et al., 2023a)、mPLUG-DocOwl 1.5 (Hu et al., 2024a)、InternLM-Xcomposer (Dong et al., 2024b)和TextHawk (Yu et al., 2024)为代表的模型采用一个与Q-Former结构相似的Resampler，在图像切片经过视觉编码器编码并拼接后进行特征压缩。Resampler通常是一个解码器结构，采用一组可学习的查询表示与视觉特征进行交叉注意力计算，以抽取高度聚合的视觉表征。训练往往冻结视觉编码器，只对Resampler进行微调。除了采用Resampler，以Monkey为代表的模型 (Li et al., 2023j; Liu et al., 2024d)在训练阶段还使用了低秩自适应的参数高效微调(LoRA)技术对视觉编码器进行参数更新，使得视觉编码器能够独立的建模不同图像切片的特征。

2. **基于双视觉编码器的多尺度特征融合**：利用多尺度图像信息允许模型捕获较小尺度中存在的细粒度细节和较大尺度中可用的全局上下文。CogAgent (Hong et al., 2023)、Mini-Gemini (Li et al., 2024)和DeepSeek-VL (Lu et al., 2024)都采用了这种方案，他们使用一个高分辨率视觉编码器用于高分辨率的全局视图处理，另一个低分辨率编码器用于低分辨率的局部视图处理。其中，Mini-Gemini (Li et al., 2024)提出了一个新颖的补丁信息挖掘(patch info mining)策略，使用低分辨率的视觉特征作为查询，通过交叉注意力从高分辨率的图像特征中检索相关的视觉线索。研究表明 (Shi et al., 2024)，一个多尺度较小模型的能力可以与一个较大模型相当，验证了多尺度方案的合理性。

3. **直接训练高分辨率视觉编码器**：以InternVL (Chen et al., 2023e)为代表的模型认为，多模态大模型受限的跨模态表征能力来源于不同模态之间参数的不平衡。因此，InternVL (Chen et al., 2023e; Chen et al., 2024b)系列模型将视觉编码器的参数量提高到6B，并采用MLP连接件来保留更多的视觉特征。InternViT-6B(Chen et al., 2023e)是拥有和大语言模型相同参数尺度的视觉编码器，通过对比学习直接和大语言模型进行对齐，来弥合视觉编码器和大语言模型之间参数尺度和特征表示能力的巨大差距。在后续的版本迭代中，InternViT-6B将图像切片的基础分辨率提高到448x448，并且能将任意分辨率的图片动态匹配到一个图像切片模版进行处理。值得一提的是，与此前的大部分模型不同，InternVL-1.5在预训练的第一阶段就解冻InternViT的参数，并持续使用高质量的图像文本数据对InternViT进行二阶段预训练，来增强模型对视觉信息的理解处理能力。在目前的开源多模态大模型中，InternVL-1.5展现了优异的性能，在多个指标上与代表性闭源大模型GPT4V (OpenAI, 2023b)相当，展现了该方案的潜力。

4 多模态大模型训练

4.1 预训练阶段

由于视觉编码器的输出空间与大语言模型的语义空间存在差异，需要利用多模态数据完成视觉文本特征之间的对齐。在预训练过程中，各个多模态大模型除了使用多种多样的多模态数据，还采取了不同的预训练策略，包含不同的预训练参数、预训练任务以及多预训练阶段。

4.1.1 预训练数据

多模态大模型的预训练数据根据数据的形式可以分为图片文本对、Grounded图片文本对、图文交错序列、表格/OCR数据、视频文本对、视频图文交错序列和仅文本语料。表 1 列出了不同的数据形式下常见的公开数据集的大小。在这些训练数据中，Taisu (Liu et al., 2022a)、Wukong Captions (Gu et al., 2022)和Youku-mPLUG (Xu et al., 2023a)的文本是中文

数据形式	数据集	图片/视频数	数据条数
图片文本对	MS-COCO (Lin et al., 2014)	110K	550K
	CC3M (Sharma et al., 2018)	2.9M	2.9M
	CC12M (Changpinyo et al., 2021)	11.1M	11.1M
	LAION-400M (Schuhmann et al., 2021)	400M	400M
	SBU Captions (Ordonez et al., 2011)	860K	860K
	VG Captions (Krishna et al., 2017)	100K	100K
	COYO-700M	580M	747M
	LAION-COCO	600M	600M
	CapsFusion-120M (Yu et al., 2023a)	120M	120M
	ShareGPT4V-Caption (Chen et al., 2023c)	1.3M	1.3M
	ALLaVA-Caption-4V (Chen et al., 2024a)	-	715K
	Wukong Captions (Gu et al., 2022)	100M	100M
	Taisu (Liu et al., 2022a)	166M	219M
Grounded图片文本对	GRIT (Peng et al., 2023)	90M	-
	Flickr30k Entities (Young et al., 2014)	30K	-
图文交错序列	MMC4 (Zhu et al., 2024)	571M	101.2M
	Wikihow (Yang et al., 2021)	772K	772K
	Wikipedia	-	-
表格、OCR数据	Char2Text (Kantharaj et al., 2022)	44K	44K
	UniChart (Masry et al., 2023)	611K	611K
	Paper2Fig100K (Rodriguez et al., 2023)	102K	102K
	Widget Captioning (Li et al., 2020d)	21K	162K
	Screen2Words (Wang et al., 2021a)	22K	112K
	TextOCR (Singh et al., 2021)	28K	903K
	COCO-Text (Veit et al., 2016)	63K	145K
视频文本对	InternVid (Wang et al., 2023f)	7.1M	234M
	Youku-mPLUG (Xu et al., 2023a)	10M	10M
	MSR-VTT (Xu et al., 2016)	7.2K	10K
	Webvid10M (Bain et al., 2021)	10.7M	10.7M
视频图文交错序列	YT-Storyboard-1B	20M	-
仅文本语料	Pile (Gao et al., 2020)	-	-

Table 1: 多模态大模型预训练数据

的。还有一些图片文本对格式的数据，利用了ChatGPT等大语言模型对文本进行了重写，如LAION-COCO、CapsFusion-120M (Yu et al., 2023a)、ShareGPT4V-Caption (Chen et al., 2023c)、ALLaVA-Caption-4V (Chen et al., 2024a)。除了多模态的训练数据外，还有模型使用仅文本语料进行预训练，防止大语言模型出现灾难性遗忘的情况。

4.1.2 预训练方法

多模态大模型的预训练阶段旨在对齐预训练模态编码器和预训练大语言模型的语义空间，我们从预训练参数、预训练任务和多预训练阶段三个角度进行介绍。

预训练参数 一部分多模态大模型预训练阶段只训练连接模块，如BLIP2 (Li et al., 2023d)、LLaVA (Liu et al., 2023d)、MiniGPT4 (Zhu et al., 2023a)等。DeepSeek-VL (Lu et al., 2024)训练连接模块和大语言模型。还有一部分模型预训练阶段会对全部参数进行训练，以Qwen-VL (Bai et al., 2023b)、Intern-VL (Chen et al., 2023e)为代表。

预训练任务 大部分模型在预训练阶段直接采用自回归语言建模任务对齐不同模态，如LLaVA (Liu et al., 2023d)、MiniGPT4 (Zhu et al., 2023a)等。为了更好的学习视觉表征中与文本有关的信息，BLIP2 (Li et al., 2023d)额外引入了图文对比学习、图文匹配任务，Intern-VL (Chen et al., 2023e)引入了对比学习。

多任务预训练阶段 Qwen-VL (Bai et al., 2023b)和MiniGPT-v2 (Chen et al., 2023b)根据训练数据的质量和数据形式将预训练阶段分为预训练和多任务预训练；Intern-VL (Chen et al., 2023e)根据预训练任务和数据质量将预训练阶段分为对比预训练和生成预训练；DeepSeek-VL (Lu et al., 2024)则根据训练参数将预训练阶段分为适配器预训练和联合预训练。

4.2 指令微调数据和微调方法

指令微调指的是用一系列带有指令和遵循指令的回答构成的文本对，对模型进行调优的一种方法。指令微调的目的是训练模型更好地理解用户的指令并遵循这个指令完成要求的任务。指令微调技术在大语言模型上的应用说明这样的方法能够帮助模型对齐人类的需求（以语言的形式描述任务），并泛化到新的场景和指令上 (Ouyang et al., 2022b; Chung et al., 2024; Chiang et al., 2023)。我们在这节中将介绍如何在多模态领域进行指令微调。

4.2.1 指令微调数据集的构建

由于指令微调数据在格式和任务的表述上有着很强的灵活性和多样性，使得收集这些数据往往比传统的有监督学习数据加困难且成本更高。在这一节里，我们将介绍三种主要的大规模收集指令微调数据的策略。

重构任务导向数据 目前存在着大量具有高质量标注的多模态数据集，但这些数据集大多用于特定的任务无法直接使用。因此很多研究工作 (Dai et al., 2023; Wang et al., 2024; Chen et al., 2023a; Xu et al., 2022; Zhang et al., 2023b; Gao et al., 2023a; Zhao et al., 2023b; Luo et al., 2023a)重构现有的高质量数据集为指令微调的形式。以经典的视觉问答任务（VQA）为例，原始的数据样本为一个输入输出对，其中输入包含一张图片和一个自然语言的问题，输出则是对应的文本回答。此类数据集的输入输出可以自然的构成指令微调阶段的输入输出。除了直接利用数据集中原始的输入外，指令微调的输入指令还可以通过手工设计或者GPT辅助生成。部分工作 (Dai et al., 2023; Xu et al., 2022; Zhu et al., 2023a; Chen et al., 2023a; Li et al., 2023c; Li et al., 2023f)针对同一个特定任务采用手工构建很多语义近似的指令组成候选池，然后在训练的时候随机选择池子中的一个作为该条数据的输入。其他一些工作 (Wang et al., 2024; Li et al., 2023e; Gong et al., 2023)通过手工设计一些种子指令，并使用这些种子指令来提示GPT生成更多的指令。现有的视觉问答（VQA）和图像描述（Image Captioning）数据集的答案通常较为简洁，直接使用这些数据集进行指令微调可能会让训练得到模型倾向于简短的输出。ChatBridge (Zhao et al., 2023b)中提到如果任务导向的原始数据回答很简短，在指令需要添加类似“in short” 这样的指示模型用简短回答的指令，而对于那些用一句话描述粗略答案则应该加上“a sentence” 等表示用一句话说完的语句。第二种方法扩展了原始回答的长度，例如M^{3}IT (Li et al., 2023f)提出通过使用原始问题、答案以及图片的背景信息（例如标题和OCR）来提示ChatGPT，从而扩展原始问题的答案。

闭源模型辅助 尽管现有的任务导向数据集能够提供丰富的数据资源，但它们通常无法很好地满足现实场景中的人类需求，如多轮对话等。部分研究通过更强大的闭源模型辅助来收集样本。这类方法利用现在市面上更强大的闭源LLMs（GPT4、Gemeni等），依据少量手工标注的样本，生成遵循指令的文本数据。具体而言，这些方法首先人工标注一些遵循指令的样例作为示范，随后引导GPT-4等闭源模型根据这些示范样例生成更多的指令样本。LLaVA (Liu et al., 2024c)将此方法扩展到多模态领域，通过将图像内容转换为标题和带有标识物体位置的检测框的文本，然后提示GPT-4生成新数据。基于此思路，后续研究如MiniGPT-4 (Zhu et al., 2023a)、ChatBridge (Zhao et al., 2023b)、GPT4Tools (Yang et al., 2023)及DetGPT (Pi et al., 2023)开发了适应不同需求的各类多模态指令微调数据集。而随着更为强大的多模态模型GPT-4V的发布，许多研究直接采用GPT-4V生成更高质量的数据，例如LVIS-Instruct4V (Wang et al., 2023a)和ALLaVA (Chen et al., 2024a)等。

文本指令数据 除了多模态指令微调数据之外，加入纯文本的指令数据同样能够用于增强对话技巧及指令遵循能力 (Gao et al., 2023b; Ye et al., 2023b; Gong et al., 2023; Luo et al., 2023a)。

4.2.2 指令微调方法

指令微调参数 相较于预训练阶段，指令微调阶段所引入的信息量更多，为了增大模型的容量，即可学习参数，除了少部分工作 (Li et al., 2023d; Dai et al., 2023; Zhu et al., 2023b)仅训练连接模块以外，多数模型会进一步训练语言模型基座，其中也会通过LoRA，Adapter等手段进行小参数的微调 (Zeng et al., 2023b; Chen et al., 2023b; Gong et al., 2023)。LLaVA-1.5则通过对比试验发现固定住语言基座可能会让模型无法适应新的指令和形式 (Liu et al., 2023c)。对

于视觉编码器部分，多数工作都将其冻结，也有部分工作训练视觉编码器来适应新的视觉建模方式或编码新的视觉信息 (Lu et al., 2024; Li et al., 2023j; Chen et al., 2023e)。

指令微调训练任务 与预训练阶段类似，指令微调阶段同样依靠自回归的语言建模任务来训练模型，不同于预训练阶段，多数模型以如下的形式来讲指令数据构造为完整的文本序列：

$X_{\text{system-message}}$
Human: X^1_{instruct}
Assistant: X^1_{response}
...
Human: X^n_{instruct}
Assistant: X^n_{response}

其中$X_{\text{system-message}}$为系统信息，用于给出对话的设定，$X^i_{\text{instruct}}$和$X^i_{\text{response}}$分别是第$i$轮对话的指令和回复，一共有$n$轮对话，具体的形式，包括对话的角色在不同的模型上设置不同，通常会和所使用的语言基座设定一致。其中只有红色部分的文本会用于计算损失。

4.3 基于人类反馈强化学习的方法

现有大语言模型基于交叉熵损失函数进行训练，虽然能够在一定程度上提升模型的泛化能力，但是没有显式引入人类偏好，与之进行对齐。因此，人类反馈强化学习（RLHF）(Stiennon et al., 2020; Ouyang et al., 2022a; Bai et al., 2022)引入人类反馈信号来对模型进行进一步优化。人类反馈学习的工作流程通常包含三个阶段：对预训练模型进行监督微调；创建奖励模型并使用人类标注数据训练；以及使用近端策略优化（PPO）(Schulman et al., 2017)通过奖励模型的奖励来优化策略模型。人类反馈强化学习作为大语言模型训练中一种强大并且可扩展的策略，也在最近被应用到了多模态大模型的领域中。

与纯文本领域不同，LLaVA-RLHF(Sun et al., 2023c)中策略模型与奖励模型都接受图片和文本作为输入。奖励模型初始化自基本的LLaVA模型(Liu et al., 2024c)，将最后一个词元的嵌入输出被线性投影为一个标量值作为输出的总体奖励。人类标注者通过比较两个由相同提示产生的回复选择出具有更少幻觉的答案作为人类偏好答案。该多模态大模型被训练以最大化由奖励模型模拟的人类奖励。除此之外，该文还提出了"事实增强RLHF"算法，即通过增加图像描述等额外的事实信息来校准奖励信号，从而在一定程度上缓解了可能出现的奖励作弊问题。ViGoR (Yan et al., 2024)设计了一个细粒度奖励模型，用于更新预训练的多模态语言模型，目标是改进视觉定位并减少幻觉现象。该模型结合了人类偏好和自动指标，通过众包收集细粒度的句子级反馈来获取用于训练奖励模型的人类判断和偏好。同时，它还利用先进的视觉感知模型来评估生成文本的定位和真实性，并在强化学习优化过程中结合成单一的奖励分数。

直接偏好优化（DPO）(Rafailov et al., 2024)作为一种新的优化方法已成为人类反馈强化学习的替代方案，这种方法直接优化策略模型以符合人类偏好，不需要创建奖励模型或使用强化学习进行优化。给定一个关于模型响应的人类偏好数据集，就可以通过直接偏好优化使用简单的二元交叉熵目标来优化策略。RLHF-V(Yu et al., 2023b)基于直接偏好优化方法提出了密集直接偏好优化（DDPO），直接根据密集和细粒度的段级偏好来优化策略模型；并且在数据层面上提供了细粒度的段级纠正形式的人类反馈数据集，以符合清晰密集且更为细粒度的人类偏好。HalDetect(Gunjal et al., 2024)提出用于检测幻觉内容的MHalDetect 数据集。该数据集涵盖了各种幻觉类型，包括不存在的物体、不真实的描述和不准确的关系。基于此数据集，HalDetect 训练了一个多模态奖励模型，并提出了细粒度直接偏好优化（FDPO）。细粒度直接偏好优化利用个别示例的细粒度偏好来增强模型区分准确描述的能力。

5 多模态大模型评测方法

多模态大模型的评测是推动该领域发展的关键环节。评测不仅为多模态大模型的持续优化提供了宝贵的反馈，还能帮助比较不同技术路径模型的性能差异。与传统多模态模型的评测方法相比，多模态大模型的评测呈现出几个新的特点：（1）由于多模态大模型具有强大的泛化能力，评估其综合性能变得尤为重要；（2）由于多模态大模型的输出形式为自由化文本，需要为这种灵活的输出设置稳定的自动化评估流程；（3）"幻觉"是大语言模型的常见问题，因此评估多模态大模型的幻觉问题也是评测过程中需要考虑的问题。现有的评测基准主要包含：以任务

为导向的基准、综合能力的基准以及评估幻觉的数据集准。根据任务类型的不同，也通常使用多样化的评估方法评测模型在相应数据集上的性能

5.1 评测数据集

多模态大模型评估数据集用于测试和比较不同的多模态大模型在各种任务上的性能。这些数据集，如VQAv2 (Goyal et al., 2017)和VSR (Liu et al., 2023d)，旨在模拟真实世界的视觉语言处理场景，涵盖视觉问答、场景文本识别、视觉推理和空间理解等多种任务。其次，鉴于大模型出色的综合表现能力，衡量综合能力的全面的评测数据集被提出，如MME (Fu et al., 2023)和MMBench (Liu et al., 2023e)。另外，大模型会过度依赖训练数据中的一些模式，从而导致严重的幻觉问题，许多评测幻觉的基准进而被提出，如CCEval (Zhai et al., 2023a)和CHAIR (Dai et al., 2022)。在这一节，我们整理了42个流行的基准，并将这些基准分为三类：面向任务的基准，综合的基准和幻觉的基准，如表 2所示。

5.1.1 面向任务的基准

早期的基准专注于解决特定任务的问题，这些基准拥有面向任务的形式和数据，往往会设计最适合某种任务的评测方式和评测指标。下面将按任务类型对这些基准分别进行介绍。

场景文本识别 旨在对场景图像作光学字符识别。TextVQA (Singh et al., 2019)要求识别出图片中的文字，从这些文字、图像信息中预测答案。DocVQA (Mathew et al., 2021)专注于考察对文档内容的理解。OCR-VQA (Mishra et al., 2019)则要求通过阅读图像中的文字进行可视化答题。

视觉推理 要求模型能够针对图像信息进行推理。VQAv2 (Goyal et al., 2017)减少了数据集中的语言偏见，强化了视觉图像在推理中的地位。GQA (Hudson and Manning, 2019)进一步扩大了数据集规模并涉及多步推理的问题。Whoops (Bitton-Guetta et al., 2023)创建故意违背常识的图像组成，并要求解释图像异常的原因。OK-VQA (Marino et al., 2019)需要依据外部知识作出回答。ScienceQA (Lu et al., 2022)是从小学和高中科学课程中收集的多模态科学选择题。VizWiz (Gurari et al., 2018)回答盲人的视觉问题，反映用户真实需求。ViQuAE (Lerner et al., 2022)包括使用知识库回答基于视觉上下文的命名实体的问题。A-OKVQA (Schwenk et al., 2022)是OK-VQA的增强，需要广泛的世界知识和常识来回答。

空间理解 考察对物体相对空间关系的理解。VSR (Liu et al., 2023d)涉及了66种空间关系用于评测空间推理能力。CLEVR (Johnson et al., 2017)包含100K渲染的图像，询问图像中的简单的3D形状的相关信息。EmbSpatial-Bench (Du et al., 2024)通过对3D具身场景的自动化探索，构建了一个评估大模型对6种以自身为中心的空间关系理解能力。

图文关系推理 要求从图像文本的关系中进行推理。WikiHow (Koupaee and Wang, 2018)包含一系列日常工作的任务图文步骤和文本概要。Winoground (Thrush et al., 2022)用于测量视觉-语言的组合推理，要求模型必须正确配对两个图片和对应的两个描述，两个描述的词汇构成完全相同，但顺序不同。SNLI-VE (Xie et al., 2019)要求预测图像和文本之间的关系是蕴涵、中性亦或矛盾的。MOCHEG (Yao et al., 2023)支持端到端的多模态事实检验和解释生成，输入是一条声明和大量网络资源，目的是评估该声明的真实性。

视觉描述 衡量对视觉内容进行生成描述的能力。TextCaps (Sidorov et al., 2020)要求识别文本并联系视觉上下文，在多个文本标记和视觉实体之间进行空间、语义和视觉推理。NoCaps (Agrawal et al., 2019)是针对新颖物体描述的大规模数据集，包含400种左右的新颖物体。Flickr30K (Young et al., 2014)基于30K图像和150K描述性标题的大型语料库构建表意图，进而构造视觉描述数据集。

视觉对话 包含VisDial (Das et al., 2017)数据集，其中智能体必须与人类就视觉内容进行对话。

引用表达 包含RefCOCO (Yu et al., 2016)数据集，要求模型能够在图像中定位到文本表达中的实体或者对图像区域内实体进行正确的文本描述。

5.1.2 综合的基准

多模态大模型旨在解决绝大多数任务。为此，现有的针对多模态大模型的基准往往会评估不同任务的性能，以期望全面综合地反映一个多模态大模型的表现。这些基准往往覆盖广阔范围的任务种类和数据内容，同时评测方式适应于多模态大模型自由化文本输出的形式，从而达到自动化评估的目的。

LAMM (Yin et al., 2023)旨在建设开源的多模态指令微调及评测框架，包含高度优化的训练框架、全面的评测体系，并支持多种视觉模态。MME (Fu et al., 2023)将问题构造成判断题的形式，并在4种任务和14种子任务上对大模型进行评测。LVLM-eHub (Xu et al., 2023b)由定量能力评估和在线互动评测平台组成，一方面，在47个标准视觉语言基准上定量评估大模型的视觉感知、视觉知识获取、视觉推理、视觉常识、对象幻觉和具身智能6类多模态能力，另一方面，搭建在线互动评测平台提供用户层面的模型排名。MMBench (Liu et al., 2023e)拥有自上而下的能力维度设计，根据定义的能力维度构造评测数据集，另外引入ChatGPT (OpenAI, 2023a)，以及提出了CircularEval的评测方式，使得评测的结果更加稳定。MMMU (Yue et al., 2023)包括11.5K来自大学考试、测验和教科书的多模态问题，涵盖30个学科和183个子领域，包括30种高度异构的图像类型。MM-Vet (Yu et al., 2023c)定义了6种核心视觉语言能力，并人工构造了包含200张图像和218个问题的开放问答式的评测基准，并使用GPT4对模型性能进行评估。ReForm-Eval (Li et al., 2023i)重构了现有的61个任务导向的视觉语言数据集将其转化为统一的生成和选择任务形式，并系统化地评估了模型地不稳定性。LLaVA-Bench-in-the-Wild收集了一组24张图像包括室内和室外场景、表情包、绘画、素描等，并人工构造了60个问题包含对话（简单问答）、详细描述和复杂推理三种问题类型。Seed-Bench (Li et al., 2023b)由19k由人类标注的多项选择题组成。评测基准涵盖12个不同的能力维度，包括图像和视频的理解能力。

5.1.3 面向幻觉的基准

与评估一般多模态大模型能力的基准不同，面向幻觉的基准主要针对生成内容中的幻觉判别以及模型的非幻觉生成。POPE (Li et al., 2023h)、NOPE (Lovenia et al., 2023) 和CIEM (Hu et al., 2023a)等任务主要关注模型生成文本中的物体幻觉，基准主要采用准确率作为评估指标。具体地，上述评测基准通过询问图像中是否存在物体并将模型响应与真实答案进行比较以计算模型对于图像中物体产生幻觉的情况。相较于于幻觉判别，生成类任务不仅能够评估模型的物体幻觉，也可以评估图像中的属性和关系幻觉 (Lovenia et al., 2023; Jing et al., 2023)。M-HalDetect (Gunjal et al., 2024) 包含16k条VQA数据的细粒度注释，使用人类评分和奖励模型分数评估模型输出包含幻觉的程度；GAVIE (Liu et al., 2023b)收集了来自Visual Genome (Krishna et al., 2017), VisText (Tang et al., 2023a)和Visual News (Liu et al., 2020)的图像，构造生成式的幻觉评测任务，并使用GPT4从准确性和相关性两方面评测模型能力；AMBER (Wang et al., 2023b)融合了生成性和判别性任务,并使用了一系列分类和生成指标评估模型生成的幻觉情况。

5.2 评测方法

现有的评测基准主要包括文本生成和选项选择这两种任务形式，而这两种任务形式分别对应不同的任务评估方法。

5.2.1 文本生成

文本生成类任务是评估多模态大模型性能最直接的形式，最适配大模型灵活的输出形式。一般包括OCR类任务、图片描述任务和视觉问答任务。

OCR类任务 主要评估多模态大模型识别图像中文本的能力，要求模型生成与图片中完全一致的目标词。在自动化评估的过程中，主要使用词语准确率（word accuracy） (Xu et al., 2023b; Liu et al., 2023g)进行评估：

$$\text{Accuracy}_{word} = \frac{\#\{\text{ predictions include target word }\}}{\#\{\text{ all predictions }\}}.$$

图像描述任务 主要评估多模态大模型对于图像内容的理解能力。图像描述任务的输出也可以用以评估多模态大模型产生幻觉的情况。具体地，针对一般的图像描述任务，主要

基准类别	任务类型	基准	图片来源	评估形式	评估指标
面向任务	场景文本识别	TextVQA (Singh et al., 2019)	OpenImages	生成	准确率
		DocVQA (Mathew et al., 2021)	网络	生成	准确率
		OCR-VQA (Mishra et al., 2019)	Amazon	生成	准确率
	视觉推理	VQAv2 (Goyal et al., 2017)	COCO	生成	准确率
		GQA (Hudson and Manning, 2019)	COCO, Flickr	生成	准确率
		Whoops (Bitton-Guetta et al., 2023)	自制	生成	匹配度，生成类指标
		OK-VQA (Marino et al., 2019)	COCO	生成	准确率
		ScienceQA (Lu et al., 2022)	网络	选择	准确率
		VizWiz (Gurari et al., 2018)	6个公开数据集	生成	准确率，生成类指标
		ViQuAE (Lerner et al., 2022)	网络	选择+生成	精确率，命中率，F1，精准匹配
		A-OKVQA (Schwenk et al., 2022)	COCO	选择+生成	准确率
	空间理解	VSR (Liu et al., 2023d)	COCO	选择	准确率
		CLEVR (Johnson et al., 2017)	自制	生成	准确率
	图文关系推理	WikiHow (Koupaee and Wang, 2018)	网络	生成	生成类指标
		Winoground (Thrush et al., 2022)	Getty Images	选择	文本分数，图像分数，组分数
		SNLI-VE (Xie et al., 2019)	Flickr30K	选择	准确率
		MOCHEG (Yao et al., 2023)	网络	选择+生成	精确率，召回率，F1，生成类指标
	视觉描述	TextCaps (Sidorov et al., 2020)	OpenImages	选择	生成类指标
		NoCaps (Agrawal et al., 2019)	OpenImages	生成	生成类指标
		Flickr30K (Young et al., 2014)	Flickr	生成	生成类指标
	视觉对话	VisDial (Das et al., 2017)	COCO	选择	召回率
	引用表达	RefCOCO (Yu et al., 2016)	COCO	生成	准确率，生成类指标
综合	综合	LAMM (Yin et al., 2023)	14个公开数据集	选择+生成	准确率，生成类指标
		MME (Fu et al., 2023)	7个公开数据集，自制	选择	准确率
		LVLM-eHub (Xu et al., 2023b)	47个公开数据集	选择	准确率
		MMBench (Liu et al., 2023e)	11个公开数据集，网络	选择	准确率
		MMMU (Yue et al., 2023)	教科书和网络	选择+生成	准确率
		MMVET (Yu et al., 2023c)	网络	生成	GPT-4分数
		ReForm-Eval (Li et al., 2023i)	现有的61个数据集	选择+生成	准确率，CIDEr
		LLaVA-Bench-in-the-Wild	网络	生成	GPT-4分数
		SEED-Bench (Li et al., 2023b)	CC3M	选择	准确率
幻觉	幻觉	CHAIR (Dai et al., 2022)	COCO, Open Images V4	生成	CHAIR
		CCEval (Zhai et al., 2023a)	VisualGenome	生成	CHAIR
		FAITHSCORE (Jing et al., 2023)	COCO	生成	FAITH
		GAVIE (Liu et al., 2023b)	Visual Genome, VisText, Visual News	生成	GPT4评分
		MMHal-Bench (Sun et al., 2023c)	OpenImages	生成	GPT4评分
		M-HalDetect (Gunjal et al., 2024)	COCO	生成	奖励模型分数+人工评分
		HaELM (Wang et al., 2023c)	COCO	生成	准确率
		POPE (Li et al., 2023h)	COCO	选择	准确率
		CIEM (Hu et al., 2023a)	COCO	选择	准确率
		NOPE (Lovenia et al., 2023)	10个VQA数据集	选择	准确率，METEOR
		AMBER (Wang et al., 2023b)	COCO, UnSplash	选择+生成	CHAIR, Cover ,Hal, Cog

Table 2: 多模态大模型评测数据集。生成类指标包括BLEU, ROUGE, METEOR, CIDEr等。

使用传统的图像描述任务的指标如CIDEr (Vedantam et al., 2015)、BLEU (Papineni et al., 2002)、METEOR (Banerjee and Lavie, 2005)和ROUGE-L(Lin, 2004)。针对评测多模态大模型的幻觉，一般使用CHAIR (Rohrbach et al., 2018)指标：

$$\text{CHAIR}_i = \frac{\#\{\text{ hallucinated objects }\}}{\#\{\text{ all objects in prediction }\}},$$

$$\text{CHAIR}_s = \frac{\#\{\text{ hallucinated sentences }\}}{\#\{\text{ all sentences }\}}.$$

除了CHAIR指标，AMBER (Wang et al., 2023b)进一步其他的幻觉评估指标：Cover，Hal，Cog以及AMBER分数。其中Cover分数主要衡量模型输出的物体类别覆盖真实物体类别的比例，理想的输出应该为最小化幻觉内容而不显著降低覆盖率。其公式为：

$$\text{Cover} = \frac{\#\{\text{objects in response} \bigcap \text{true objects }\}}{\#\{\text{ true objects }\}}.$$

Hal分数表示出现幻觉的反应比例。对于模型的输出，如果$\text{CHAIR}_i \neq 0$，则认为模型出现幻觉。Hal分数的公式为：

$$\text{Hal} = \begin{cases} 1 & \text{if CHAIR}_i \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Cog分数主要用于评估多模态大模型中的幻觉是否与人类认知中的幻觉相似。在给定人类幻觉目标物体集合（hallucinatory target objects）的情况下，Cog分数的公式为：

$$\text{Cog} = \frac{\#\{\text{objects in response} \bigcap \text{hallucinatory target objects }\}}{\#\{\text{ objects in response }\}}.$$

为了全面评估各种多模态大模型在生成式和判别式任务下的表现，AMBER分数被提出，来整合生成任务上的CHAIR分数和判别任务上的F1分数。AMBER分数的公式为：

$$\text{AMBER Score} = \frac{1}{2} \times (1 - \text{CHAIR} + F1).$$

视觉问答任务 主要形式是给定图片和任意问题，利用模型生成自由文本来回答问题。这类任务能够反映模型的综合能力 (Yue et al., 2023; Yu et al., 2023c)。然而，由于模型输出的形式自由，且答案逻辑较为复杂，通常很难计算准确率等指标。因此，在评测过程中通常采用人工评分，或者使用大语言模型如GPT-4 (OpenAI, 2023b)，通过对比人工标注的答案和模型生成的结果来评分，从而衡量模型输出的质量。

5.2.2 选项选择

选项选择是指任务所有可能的答案选项是预先定义好的，并且限制在一个有限的集合内。使用有限选项的选择题的评估形式能够极大增强评估结果的稳定性和可重复性。选项选择问题的形式包含单选题、判断题等，检索任务可以看作一种特殊的选择题形式。例如，在视觉问答中通过选择题的方式提供选项，要求模型输出正确的选项标记。

选项选择任务可以通过计算选项准确率评估模型的性能，也会计算扰乱选项后输出选项的准确率排除选项顺序对于模型输出的影响 (Fu et al., 2023; Liu et al., 2023e)。

由于模型输出格式的问题，因此在获取多模态模型的选择的选项过程中，一般有如下技术：（1）使用上下文样本来引导模型按期望的格式输出选项，再进行字符串匹配；（2) 计算模型在给定图片文本下对于不同选项的生成概率，选择最高概率的选项作为模型选择题的答案 (Li et al., 2023i)；（3）针对图像生成类问题，通过计算生成图像和真实图像的CLIP相似分数，来获取模型选择题的答案 (Li et al., 2023b)；（4）通过计算模型对于各个选项的困惑度，选择困惑度最低的选项作为模型选择的答案 (Li et al., 2023b)。

6 多模态大模型的新趋势

6.1 解决多模态大模型的幻觉问题

“幻觉”被定义为生成无意义的内容或偏离其来源的内容(Ji et al., 2023)，而多模态大模型中的“幻觉”是指模型的视觉输入和文本输出之间的矛盾。从视觉语言任务的角度来看，当多模态大模型对用户的询问或陈述的响应与实际视觉数据不一致，即判断或描述具有缺陷时，则出现了幻觉(Liu et al., 2024a)。多模态大模型的幻觉来源于多种因素，包括来自数据的幻觉，即现有训练数据分布不平衡导致的数据偏差和不准确的标签或是注释；来自视觉编码器的幻觉，即有限的视觉分辨率和缺乏细粒度的视觉语义；来自多模态对其的幻觉，即连接模块结构过于简单和受限的标记约束；来自大语言模型的幻觉，即上下文注意力不足，随机采样解码以及能力错配等。

为解决多模态大模型的幻觉问题，可以从导致幻觉的因素出发。首先，减少幻觉的直接有效的方法是优化训练数据。具体的方法是引入负面数据和反事实数据，以增加数据集的多样性或是减少现有数据集中的噪声和错误，通过重写文本标注来提高数据集质量。对于数据偏差，CIEM(Hu et al., 2023a)利用现有的大语言模型从带有标注的图像文本数据集中生成对比的正负问答对，并用于对比指令微调。LRV-Instruction(Liu et al., 2023a)提出一个多样化大规模数据集，包含正确的视觉指令以及分别在三个不同的语义层面上对应的负指令。对于不准确的注释，构建丰富注释的数据集精确提取视觉内容和全面对齐模态也能够减轻多模态大模型的幻觉(Gunjal et al., 2024; You et al., 2023; Lu et al., 2023b)。

从视觉编码器改善多模态大模型幻觉的角度而言，利用支持更高分辨率的视觉编码器已经被广泛证明能够提升模型的视觉感知能力(Liu et al., 2023c; Bai et al., 2023b; Li et al., 2023j; Lu et al., 2024)。InternVL(Chen et al., 2023e)则进一步增大视觉编码器的参数规模，扩展至60亿参数。然而，现有的大多数多模态大模型使用ViT(Dosovitskiy et al., 2021)作为视觉编码器关注显著对象却忽略了一些关键的视觉线索。(Jain et al., 2023; Zhao et al., 2023a)引入额外的空间信息来引导多模态大模型处理用户查询，进一步增加了模型的对象级感知能力和空间感知能力。

对于视觉编码器和大语言模型之间的连接模块，连接模块的参数或结构影响了视觉文本之间的模态对齐，导致了幻觉的产生。最近一些工作开发了更强大的视觉语言连接模块：例

如，LLaVA-1.5将连接模块由单个线性层升级为多层感知机，实现了在各类多模态大模型评测数据集上的指标提升；InternVL(Chen et al., 2023e)利用LLaMA(Touvron et al., 2023a)构建了QLLaMA，由于QLLaMA使用预训练权重初始化，并且具有80 亿参数，因此在视觉语言对齐方面显著优于Q-Former(Li et al., 2023d)。

从大语言模型的角度来看，改进训练目标或是训练方法可以减少幻觉。(Chen et al., 2023f)引入辅助监督使用额外的标注信息来辅助监督模型关注图像内容。(Jiang et al., 2023b)通过对比损失减少文本和视觉样本之间的分布差距。基于人类反馈学习的方法能够进一步对齐人类偏好：LLaVA-RLHF(Sun et al., 2023c)通过引入人类反馈来减轻幻觉现象，将人类反馈学习范式从文本领域扩展到视觉语言对齐任务中；ViGoR(Yan et al., 2024)则设计了一个细粒度的奖励模型来更新策略模型，用于改进视觉定位并减少幻觉现象；RLHF-V(Yu et al., 2023b)基于直接偏好优化(Rafailov et al., 2024)提出密集直接偏好优化（DDPO），直接根据密集和细粒度的段级偏好来优化策略模型；HalDetect(Gunjal et al., 2024)提出了细粒度直接偏好优化（FDPO），使用来自个别示例的细粒度偏好来直接减少生成文本中的幻觉。除此之外，HallE-Switch(Zhai et al., 2023a)通过控制大语言模型中的参数知识来减少幻觉；OPERA(Huang et al., 2023)则提出基于过度信任惩罚和回顾分配策略的多模态大模型解码方法，无需额外的数据、知识或训练即可缓解幻觉问题。

6.2 任意模态输出的扩展

多数多模态大模型构建在大语言模型的基础上，也以文本作为主要的输出形式，然而正如GPT-4V (OpenAI, 2023b)以及GPT-4o (OpenAI, 2024)展现出的能力，用户的需求往往需要模型以多模态的输出来满足，对于开源模型来说，输出端的模态扩展也是研究的热门趋势。

首先被尝试扩展的模态是图像，应对文生图，图片编辑的任务需求。在大模型时代之前，StableDiffusion (SD) (Rombach et al., 2022)，InstructPix2Pix (Brooks et al., 2023)等模型已经在对应任务上展现了优异的性能，早期的Visual ChatGPT (Wu et al., 2023a)通过使用工具的方式引入了这样的能力，Mini DALLE3 (Lai et al., 2023)进一步探究了多个工具增强的大语言模型在交互式文生图场景下的表现。为了将图片生成能力内化到模型内部并且支持端到端的训练，GILL (Koh et al., 2023)首先提出用特殊字符在自回归生成过程力区分图片和文本的输出，并将输出图片位置的表示通过特殊的连接模块映射到SD模型输入，最终产生图片，这样的范式也在DreamLLM (Dong et al., 2024a)，MiniGPT-5 (Zheng et al., 2024)中被延用。除了来自SD的监督信号，原本SD使用的CLIP文本编码器也可以帮助对齐多模态大模型的输出和SD模型的输入 (Pan et al., 2024)。进一步，Emu (Sun et al., 2024b), Emu2 (Sun et al., 2024a), SEED (Ge et al., 2023)通过构建自编码器（AutoEncoder）的形式对齐了模型的输入输出，图片的输入表示可以用来恢复原本的图片并且监督多模态大模型自回归的训练。其中SEED以离散的形式表示中间的隐空间，Emu和Emu2则使用连续的隐空间。

除了图片，从任意模态到任意模态的生成则更具挑战，Next-GPT (Wu et al., 2023b)和Codi-2 (Tang et al., 2023b)对每种模态都构建了一个Diffusion模型，并用特殊字符区分不同的输出模态，将多模态大模型的输出表示提供给对应的模态生成器中。Any-GPT (Zhan et al., 2024)和Unified-IO2 (Lu et al., 2023a)则通过类似VQ-VAE的方法将多个模态的离散化表示（词表）增加到多模态模型输出的词表中，使得模型能够以原本自回归的方式输出多模态交错的序列。目前任意模态输出的方法主要以文本为支点，利用多个模态和文本的相关数据帮助训练，尽管目前也有方法来在GPT的帮助下构造部分多模态交错生成的数据 (Zhan et al., 2024; Wu et al., 2023b)，但是数据规模和形式相对有限，这也是限制当前模型的重要因素。

6.3 具身场景下的探索

得益于多模态大模型的多模态交互能力和强大推理能力，开发基于多模态大模型的具身智能体已成为具身智能领域的主要探索方向之一 (Zeng et al., 2023a)。其中一个主要挑战是如何将多模态大模型生成的文本与控制具身智能体的动作相结合，例如视觉语言导航任务 (Anderson et al., 2018b; Zhang et al., 2021a)。PaLM-E (Driess et al., 2023)通过接收语言、视觉和状态估计的多模态输入序列，输出低级别指令给下游的策略模块，以完成相应的具身任务。同时，它还是一个视觉语言通用模型，在传统的视觉语言任务（如VQA）中表现良好。与之相比，RT-2 (Brohan et al., 2023)可以将指令和视觉观察直接映射到机器人动作。具体来说，RT-2将低级别的机器人动作参数向量离散化为特殊的文本标记，加入模型的词表中。在实

际操作过程中，模型直接生成词表中的动作标记来控制机器人的动作。ManipLLM (Li et al., 2023g)则通过直接微调模型，使其输出具体的控制参数。具体而言，ManipLLM设计了物体、区域和姿势三个级别的微调任务，使模型能够逐步合理地预测以物体为中心的机器人操控姿势。NaviLLM (Zheng et al., 2023)和LLaRP (Szot et al., 2023)则利用额外的动作分类器，将特定标记对应的嵌入向量映射到合法动作，以完成动作控制。受Code-as-Policies (Liang et al., 2023)的影响，RoboCodeX (Mu et al., 2024)通过构造使用代码控制具身任务的预训练和指令微调数据，将代码生成与动作控制对齐，通过调用相应的API生成可执行代码，以控制下游具身任务的执行。目前，将多模态大模型的输出转化为可执行动作仍然是一个开放的问题。此外，现有的研究主要关注多模态大模型在特定具身任务上的应用。如何充分发挥其泛化能力，开发面向多样化具身任务的通用模型，也是一个亟待探索的问题。

7 总结

本文系统地回顾和探讨了多模态信息处理领域的研究进展，重点介绍了多模态预训练模型和多模态大模型的发展历程与技术细节。

首先，我们回顾了多模态预训练模型的早期研究，这些模型借鉴了文本预训练模型的成功经验，通过大量的视觉和文本数据进行自监督学习，取得了一定的成果。然而，预训练模型在泛化能力上存在不足，难以满足多样化应用场景的需求，很快被多模态大模型所取代。接下来，本文重点介绍了多模态大模型的出现及其架构设计。随着大语言模型的成功应用，多模态大模型通过扩展语言模型的能力，引入多模态编码器，实现了跨模态的高效对齐。我们详细讨论了多模态大模型的序列表示、基座模型和架构优化方案，特别是多视图视觉表征和特征压缩技术的应用。在训练方法方面，我们分析了多模态大模型的预训练阶段和指令微调方法，介绍了如何利用多模态数据完成视觉与文本特征的对齐，并通过指令微调和基于人类反馈的强化学习提高模型对自然语言指令的理解和执行能力。接着，我们分析了当前多模态大模型的评测方法，包含对已有基准数据集的简要介绍和对评测方法和归纳总结。最后，本文探讨了多模态大模型在解决幻觉问题、扩展任意模态输出以及具身场景探索方面的潜力和挑战。

本文的贡献在于全面、系统地总结了多模态信息处理的研究脉络。从早期的多模态预训练模型到当前的多模态大模型，本文详细分析了每个阶段的技术进展和应用场景，以及如何进行可靠的评测。通过对比不同方法的优缺点，我们揭示了各模型在处理跨模态信息时的优点和局限性。基于这些分析，我们提出了未来多模态研究可能的发展方向和潜在的创新点。希望本文能够为多模态技术的发展和应用提供有益的参考。

致谢

该论文得到了国家自然科学基金委(No. 62176058) 和科技部重点研发计划（2023YFF1204800）的项目经费支持。复旦大学CFFF平台为该项目提供了算力支持。

参考文献

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *NIPS*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390.*

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862.*

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609.*

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966.*

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, October.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022a. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc.

Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. 2022b. Vl-beit: Generative vision-language pre-training.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954.*

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *arXiv:2303.07274.*

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818.*

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, June.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297.*

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm. Dec.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tianlang Chen, Jiajun Deng, and Jiebo Luo. 2020a. Adaptive offline quintuplet loss for image-text matching. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 549–565, Berlin, Heidelberg. Springer-Verlag.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv:2305.04160.*

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023b. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478.*

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023c. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793.*

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023d. Pali: A jointly-scaled multilingual language-image model.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023e. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238.*

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023f. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479.*

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684.*

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821.*

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

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

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 18–24 Jul.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

HyungWon Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, ShixiangShane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, EdH. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, QuocV. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. Oct.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 958–979, January.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*, pages 326–335.

Mohammad Dehghani and Pavel Trojovskỳ. 2023. Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems. *Frontiers in Mechanical Engineering*, 8:1126450.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024a. Dreamllm: Synergistic multimodal comprehension and creation.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024b. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. 2022a. Coarse-to-fine vision-language pre-training with fusion in the backbone. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32942–32956. Curran Associates, Inc.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022b. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, June.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, March.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuan-Jing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6514–6524.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, Ruize Wang, Zejun Li, Haijun Shan, and Xuanjing Huang. 2021. Tcic: Theme concepts learning cross language and vision for image captioning. *arXiv preprint arXiv:2106.10936*.

Zhihao Fan, Zhongyu Wei, Zejun Li, Siyuan Wang, Haijun Shan, Xuanjing Huang, and Jianqing Fan. 2022. Constructing phrase-level semantic labels to form multi-grained supervision for image-text retrieval. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 137–145.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023a. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023b. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738.*

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790.*

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530.*

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914.*

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023a. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301.*

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023b. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038.*

Anwen Hu, Haiyang Xu, Jiabo Ye, Mingshi Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *ArXiv*, abs/2403.12895.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024b. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395.*

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12976–12985, June.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2023. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023b. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint arXiv:2312.06968*.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient multimodal large language models: A survey.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *ICCV*.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 18–24 Jul.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21487–21506. Curran Associates, Inc.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv:1810.09305*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73.

Zeqiang Lai, Xizhou Zhu, Jifeng Dai, Yu Qiao, and Wenhai Wang. 2023. Mini-dalle3: Interactive text to image by prompting large language models.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *45th ACM SIGIR*, pages 3108–3120.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344, Apr.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020b. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, November. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 121–137, Berlin, Heidelberg. Springer-Verlag.

Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020d. Widget captioning: Generating natural language description for mobile user interface elements. *arXiv preprint arXiv:2010.04295*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Zejun Li, Zhongyu Wei, Zhihao Fan, Haijun Shan, and Xuanjing Huang. 2021b. An unsupervised sampling approach for image-sentence matching using document-level structural information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13324–13332.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022c. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023a. Otterhd: A high-resolution multi-modality model. *ArXiv*, abs/2311.04219.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023c. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023e. Videochat: Chat-centric video understanding. *arXiv:2305.06355*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023f. M^{3}it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv:2306.04387*.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. 2023g. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023h. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, et al. 2023i. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023j. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022a. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. arxiv e-prints.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. 2023f. On the hidden mystery of ocr in large multimodal models. *ArXiv*, abs/2305.07895.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023g. On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024d. Textmonkey: An ocr-free large multimodal model for understanding document. *ArXiv*, abs/2403.04473.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023a. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023b. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv:2403.05525.*

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023a. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv:2305.15023.*

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023b. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207.*

Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. 2023. A benchmark for chinese-english scene text image super-resolution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19395–19404.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761.*

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611.*

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 ICDAR*, pages 947–952. IEEE.

MistralAITeam. 2023. Mixtral of experts a high quality sparse mixture-of-experts. [EB/OL]. https://mistral.ai/news/mixtral-of-experts/ Accessed December 11, 2023.

Yao Mu, Junting Chen, Qinglong Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, et al. 2024. Robocodex: Multimodal code generation for robotic behavior synthesis. *arXiv preprint arXiv:2402.16117.*

OpenAI. 2023a. Chatgpt (august 3 version).

OpenAI. 2023b. Gpt-4 technical report. *arXiv:2303.08774.*

OpenAI. 2024. Hello gpt-4o.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. Detgpt: Detect what you need via reasoning. *arXiv:2305.14167*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3689–3698.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. arxiv.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *ArXiv*, abs/2403.13043.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July. Association for Computational Linguistics.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023a. Generative multimodal models are in-context learners. *ArXiv*, abs/2312.13286.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023b. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023c. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Emu: Generative pretraining in multimodality.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. 2023. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November. Association for Computational Linguistics.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023a. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023b. Codi-2: In-context, interleaved, and interactive any-to-any generation.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288.*

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv:1601.07140.*

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021a. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021b. Ufo: A unified transformer for vision-language representation learning.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 17–23 Jul.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022c. Simvlm: Simple visual language model pretraining with weak supervision.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574.*

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397.*

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023c. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126.*

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023d. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079.*

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023e. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023f. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942.*

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671.*

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv:1901.06706.*

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul. PMLR.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online, August. Association for Computational Linguistics.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv:2212.10773.*

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. 2023a. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362.*

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv:2306.09265.*

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023c. Bridgetower: Building bridges between encoders in vision-language representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10637–10647, Jun.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *ArXiv*, abs/2403.11703.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118.*

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow. *arXiv preprint arXiv:2104.05845.*

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv:2305.18752.*

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *ACM SIGIR*, pages 2733–2743.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Mingshi Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Feiyan Huang. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *ArXiv*, abs/2310.05126.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178.*

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023c. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv:2306.06687.*

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704.*

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652.*

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216, May.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2023a. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550.*

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849.*

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *ArXiv*, abs/2404.09204.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502.*

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR, 17–23 Jul.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023a. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023b. What matters in training a gpt4-style language model with multimodal inputs? *arXiv:2307.02469*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023a. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling.

Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. 2018. S3d: Single shot multi-span detector via fully 3d convolutional networks.

Jiwen Zhang, Jianqing Fan, Jiajie Peng, et al. 2021a. Curriculum learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:13328–13339.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588.

Pan Zhang, Xiaoyi Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, Xinyu Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Y. Qiao, Da Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *ArXiv*, abs/2309.15112.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*.

Xinnong Zhang, Haoyu Kuang, Xinyi Mou, Hanjia Lyu, Kun Wu, Siming Chen, Jiebo Luo, Xuanjing Huang, and Zhongyu Wei. 2024. Somelvlm: A large vision language model for social media processing. *arXiv preprint arXiv:2402.13022*.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference.

Yongqiang Zhao, Zhenyu Li, Zhi Jin, Feng Zhang, Haiyan Zhao, Chengfeng Dou, Zhengwei Tao, Xinhai Xu, and Donghong Liu. 2023a. Enhancing the spatial awareness capability of multi-modal large language model. *arXiv preprint arXiv:2310.20357*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023b. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv:2305.16103*.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2023. Towards learning a generalist model for embodied navigation. *arXiv preprint arXiv:2312.02010*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2024. Minigpt-5: Interleaved vision-and-language generation via generative vokens.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NeurIPS*, 36.