REVIEW

National Science Review
XX: nwae403, 2024 doi: https://doi.org/10.1093/nsr/nwae403

COMPUTER SCIENCE

关于多模态大型语言模型的一项调查

Shukang Yin[1,†], Chaoyou Fu[2,3,*,†], Sirui Zhao[1,*,†], Ke Li[4], Xing Sun[4], Tong Xu[1] and Enhong Chen[1,*]

摘要

近年来，以GPT-4V为代表的多模态大语言模型（MLLM）已成为新的研究热点，它利用强大的大语言模型（LLMs）作为大脑来执行多模态任务。MLLM令人惊讶的新兴能力，如基于图像编写故事和无光学字符识别（OCR）的数学推理，在传统多模态方法中很少见，这为实现通用人工智能指明了潜在的路径。为此，学术界和工业界都在努力开发能够与GPT-4V竞争甚至超越它的MLLM，以惊人的速度推动研究的极限。在本文中，我们旨在追踪和总结MLLM的最新进展。首先，我们介绍了MLLM的基本公式，并勾勒出了其相关概念，包括架构、训练策略和数据，以及评估。然后，我们介绍了关于如何将MLLM扩展以支持更多粒度、模态、语言和场景的研究课题。我们接着讨论了多模态幻觉和扩展技术，包括多模态自适应学习（M-ICL）、多模态自适应推理（M-CoT）和LLM辅助的视觉推理（LAVR）。在本文结尾，我们讨论了现存挑战，并指出了有前景的研究方向。

关键词：多模态大型语言模型、视觉语言模型、大型语言模型

[1] 中国科学技术大学，合肥 230026，中国；

[2] 新型软件技术国家重点实验室，南京大学，南京 210023，中国；

[3] 南京大学智能科学与技术学院，苏州 215163，中国；

[4] 腾讯优图实验室，中国上海 200233

[*] **通讯作者**。电子邮件：bradyfu24@gmail.com；司瑞@mail.ustc.edu.cn；陈赫@ustc.edu.cn。[†]Equally 对本研究有贡献。

收稿日期： XXXX 年；
修订日期： XXXX 年；
录用日期： XXXX 年

引言

近年来，大型语言模型（LLMs）取得了显著的进展[1,2]。通过扩大数据规模和模型规模，这些大型语言模型展现出了非凡的涌现能力，通常包括指令遵循[3]、上下文学习（ICL）[4]和思维链（CoT）[5]。尽管大型语言模型在大多数自然语言处理（NLP）任务[6]甚至复杂的实际应用中展示了令人惊讶的零/少量推理性能[7-9]，但它们本质上对视觉是“盲”的，因为它们只能理解离散的文本。与此同时，大型视觉模型（LVMs）能够清晰地看到[10,11]，但通常在推理方面存在滞后。

鉴于这种互补性，语言模型（LLM）和视觉模型（LVM）相互融合，催生了多模态大语言模型（MLLM）这一新领域。正式来说，它指的是基于语言模型（LLM）的模型，能够接收、推理和输出多模态信息。在多模态大语言模型（MLLM）出现之前，已经有很多致力于多模态的工作，可以分为判别式[12,13]和生成式[14,15]范例。CLIP[12]是前者的代表，它将视觉和文本信息投影到一个统一的表示空间，为下游的多模态任务搭建了桥梁。相比之下，OFA[14]是后者的代表，它以序列到序列的方式统一了多模态任务。根据序列操作，多模态大语言模型（MLLM）可以被归类为后者，但它与传统同类模型相比表现出了两个不同的特点：（1）多模态大语言模型（MLLM）基于具有数十亿规模参数的语言模型（LLM），这是之前模型所没有的。（2）多模态大语言模型（MLLM）采用新的训练范例来释放其全部潜力，例如使用多模态指令调整[16]来鼓励模型遵循新的指令。具备这两种特性，MLLM 展现出了新的能力，例如基于图片编写网站代码[17]、理解表情包的深层含义[18]以及无需光学字符识别（OCR）的数学推理[19]。

自从 GPT-4 发布以来[20]，由于其令人惊叹的多模态示例，关于多语言语言模型（MLLMs）的研究出现了热潮。快速的发展是由效率推动的

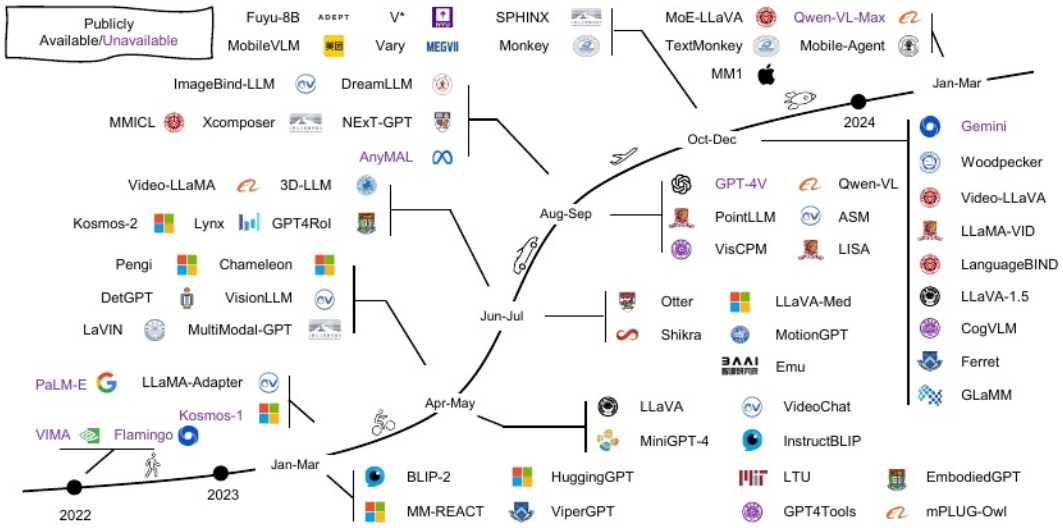

图 1 。具有代表性的多语言学习的代表性时间线。我们看到这个领域的快速发展。更多的工作可以在我们每日更新的发布的 GitHub 页面上找到。

来自学术界和工业界的堡垒。关于多模态语言模型（MLLMs）的初步研究集中在基于文本提示的文本内容生成上，以及图像[16]、视频[21,22]、音频[23]。后续的工作扩展了其能力或使用场景，包括：（1）更好的粒度支持。开发了对用户提示的更精细控制，以支持通过框[24]或点击[25]指定区域或特定对象。（2）对输入和输出模态（如图像、视频、音频和点云）的增强支持[26,27]。（3）改进的语言支持。努力将多模态语言模型的成功扩展到其他语言（例如中文），这些语言的训练语料库相对有限[28]。（4）扩展到更多的领域和使用场景。一些研究将多模态语言模型的强大能力转移到其他领域，如医学图像理解[29]和文档解析[30]。此外，还开发了多模态代理来协助现实世界的交互，*例如*具身代理[31]和图形用户界面代理[32]。多模态语言模型的时间线如图1所示。

鉴于该领域的快速发展和有前景的结果，我们撰写这篇综述，旨在为研究人员提供对多模态语言模型（MLLMs）的基本思想、主要方法和当前进展的掌握。请注意，我们主要关注视觉和语言模态，但也包括涉及其他模态（如视频和音频）的工作。具体而言，我们涵盖了多模态语言模型的最重要方面，并提供了相应的总结，同时开设了一个 GitHub 页面，以便实时更新。据我们所知，这是关于多模态语言模型的首次综述。

以下部分是本次调查的结构安排：本次调查首先对多语言学习的关键方面进行了全面审查，包括（1）主流架构；

（2）完整的训练策略及数据配方；

(3) 绩效评估的常见做法。然后，我们深入探讨了一些关于多模态语言模型的重要话题，每个话题都聚焦于一个主要问题：(1) 哪些方面可以进一步改进或扩展？(2) 如何缓解多模态幻觉问题？调查继续介绍三种关键技术，每种技术都适用于特定的场景：M-ICL是一种在推理阶段常用于提高少样本性能的有效技术。另一种重要技术是M-CoT，通常用于复杂的推理任务。之后，我们概述了开发基于大型语言模型的系统以解决复合推理任务或处理常见用户查询的一般思路。最后，我们以总结和潜在的研究方向结束我们的调查。

建筑学

一个典型的多模态语言模型可以被抽象为三个模块，*即*一个预训练的模态编码器、一个预训练的大型语言模型（LLM）以及一个连接它们的模态接口。打个比方，模态编码器，如图像/音频编码器，就像人类的眼睛/耳朵，接收并预处理光学/声学信号，而LLM则像人类的大脑，理解和推理处理后的信号。在中间，模态接口用于对齐不同的模态。一些多模态语言模型还包括一个生成器，用于输出除文本之外的其他模态。该架构的示意图如图2所示。在本节中，我们

依次介绍每个模块。

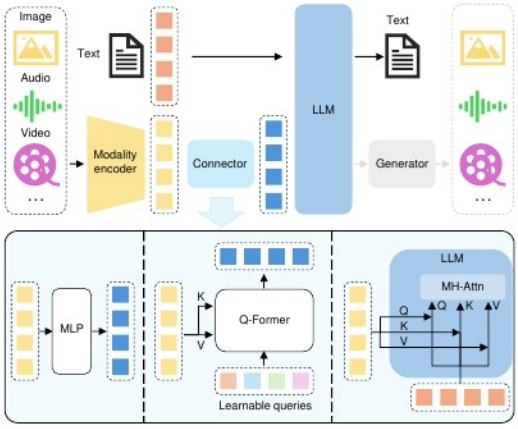

图 2。典型的 MLLM 架构的图示。它

它包括一个编码器、一个连接器和一个语言模型（LLM）。一个可选的生成器可以连接到语言模型上，以生成除文本之外的其他模态。编码器接收图像、音频或视频，并输出特征，由连接器进行处理，以便语言模型能够更好地理解。连接器大致有三种类型：基于投影的、基于查询的和基于融合的。前两种类型采用标记级融合，将特征处理为标记，与文本标记一起发送，而最后一种类型在语言模型内部实现特征级融合。

模态编码器

编码器将原始信息（如图像或音频）压缩成更紧凑的形式。一种常见的方法不是从头开始训练，而是使用预先训练好的编码器，该编码器已与其他模态对齐。例如，CLIP [12] 通过在图像 - 文本对上进行大规模的预训练，将视觉编码器与文本进行语义对齐。因此，利用此类预先对齐的编码器通过对齐预训练与大型语言模型（LLM）进行对齐会更实用。

常用的图像编码器总结于表1中。除了基本的CLIP图像编码器[12]外，一些研究还探索了使用其他变体。例如，MiniGPT-4[17]采用了EVA-CLIP[36]（ViT-G/14）编码器，该编码器使用改进的训练技术进行训练。Osprey[25]引入了一种基于卷积的ConvNext-L编码器[33]，以利用更高的分辨率和多级特征。一些研究还探索了无编码器架构。例如，Fuyu-8b[37]的图像块在发送给语言模型之前直接进行投影。通过这种设计，模型自然支持灵活的输入图像分辨率。

在选择编码器时，人们通常会考虑分辨率、参数大小和预训练语料库等因素。值得注意的是，许多研究已经通过实验验证，使用更高的分辨率可以实现显著的性能提升[28, 38]。提高输入分辨率的方法可以分为直接缩放和分块方法。直接缩放的方法是将更高分辨率的图像输入到编码器中，这通常需要进一步调整编码器[28]或使用更高分辨率的预训练编码器[39]。同样，CogAgent[32]使用双编码器机制，其中两个编码器分别处理高分辨率和低分辨率图像。高分辨率特征通过交叉注意力注入到低分辨率分支中。分块方法将高分辨率图像分割成小块，并重新使用低分辨率编码器。例如，Monkey[38]和SPHINX[40]将大图像分割成较小的块，并将子图像与下采样后的高分辨率图像一起发送到图像编码器，其中子图像和低分辨率图像分别捕获局部和全局特征。相比之下，实证研究[41]发现，与输入分辨率相比，参数大小和训练数据的构成重要性较低。

对于其他模态，也有类似的编码器可用。例如，Pengi [23] 使用 CLAP [42] 模型作为音频编码器。ImageBind-LLM [26] 使用 ImageBind [43] 编码器，该编码器支持对图像、文本、音频、深度、热成像和惯性测量单元（IMU）数据的编码。配备强大的编码器，ImageBind-LLM 能够对各种模态的输入做出响应。

预训练的大型语言模型

与其从头训练一个大型语言模型（LLM），从预训练的模型开始使用更加高效且实用。通过对网络语料库进行大量的预训练，大型语言模型已被嵌入丰富的世界知识，并展现出强大的泛化和推理能力。

我们在表2中总结了常用的公开可用的大型语言模型。值得注意的是，大多数大型语言模型都属于因果解码器类别，遵循GPT-3[4]。其中，Flan-T5[44]系列是相对早期的大型语言模型，用于BLIP-2[50]和InstructBLIP[51]等工作中。LLaMA系列[45]和Vicuna家族[46]是具有代表性的开源大型语言模型，吸引了大量的学术关注。由于这两种大型语言模型主要是在英语语料库上进行预训练的，它们在多语言支持方面存在局限性，比如中文。相比之下，Qwen [48]是一个支持中文和英文的双语大型语言模型。

应当指出的是，扩大参数规模

表 1. 常用图像编码器的概述

变体	预训练语料库	分辨率	样本（B）	参数大小（兆）
OpenCLIP-ConvNext-L [33]	LAION-2B	320	29	197.4
CLIP-ViT-L/14 [12]	OpenAI 的 WIT	224/336	13	304.0
EVA-CLIP-ViT-G/14 [34]	LAION-2B，COYO-700M	224	11	1000.0
OpenCLIP-ViT-G/14 [33]	LAION-2B	224	34	1012.7
OpenCLIP-ViT-bigG/14 [33]	LAION-2B	224	34	1844.9
InternViT-6B [35]	多个数据集	448	-	5540.0

表 2 。常用开源的大型语言模型（LLM）的概述。en、zh、fr 和 de 分别代表英语、汉语、法语和德语。

模型	发布日期	预训练数据规模	参数大小（字节）	语言支持	架构
弗拉恩 T5 超大/特大号 [44]	2022 年 10 月	-	3/ 11	英语、法语、德语	编码器 - 解码器
拉玛 [45]	2023年2月	1.4 个代币	7/ 13/ 33/ 65	中文	因果解码器
羊驼[46]	2023年3月	1.4 个代币	7/ 13/ 33	中文	因果解码器
拉玛 - 2 [47]	2023 年 7 月	2 个 T 代币	7/ 13/ 70 的中文翻译为 7/ 13/ 70	中文	因果解码器
Qwen [48]	2023年9月	3T 代币	1.8 / 7 / 14 / 72	英语，中文	因果解码器
拉玛-3 [49]	2024年4月	15 个 T 代币	8 / 70 / 405	英语、法语、德语等等	因果解码器

大型语言模型（LLMs）的规模大小也会带来额外的收益，类似于提高输入分辨率的情况。具体来说，刘等人[39,52]发现，仅仅将大型语言模型从 70 亿个参数增加到 130 亿个参数，就在各种基准测试中带来了全面的改进。此外，当使用 340 亿个参数的大型语言模型时，该模型显示出零样本中文能力，尽管在训练期间只使用了英语多模态数据。卢等人[53]通过将大型语言模型从 130 亿个参数增加到 350 亿个参数以及 650 亿/700 亿个参数观察到了类似的现象，其中更大的模型规模在专门为多语言大型语言模型设计的基准测试中带来了持续的收益。有些工作则使用较小规模的大型语言模型，以方便在移动设备上部署。例如，MobileVLM 系列[54]使用缩小规模的 LLaMA [45]，以便在移动处理器上实现高效推理。

最近，针对大型语言模型（LLMs）的专家混合（MoE）架构的探索引起了越来越多的关注[55]。与密集模型相比，这种稀疏架构通过有选择地激活参数，能够在不增加计算成本的情况下扩大总参数规模。从经验来看，MM1 [41]和 MoE-LLaVA [56]发现，在几乎所有基准测试中，MoE 的实现都比密集模型表现更好。

模态界面

由于大型语言模型只能感知文本，有必要弥合自然语言与其他模态之间的差距。然而，以端到端的方式从头开始训练一个大型多模态模型成本高昂。一种更实用的方法是在预训练的视觉编码器和大型语言模型之间引入一个可学习的连接器。另一种方法是借助专家模型将图像转换为语言，然后将语言发送给大型语言模型。

可学习连接器。它负责弥合不同模态之间的差距。具体而言，该模块将信息投影到大型语言模型能够高效理解的空间中。基于多模态信息如何融合，对于不同的模态，大致有两种实现此类接口的方式，*即*标记级和特征级融合。

对于令牌级别的融合，编码器输出的特征被转换为令牌，并在被送入大型语言模型（LLMs）之前与文本令牌连接起来。一种常见的解决方案是利用一组可学习的查询令牌以基于查询的方式提取信息[57]，这种方式最初在 BLIP-2 中实现[50]，随后被各种工作继承[22,51]。这种 Q-Former 式方法将视觉令牌压缩为数量更少的表示向量。相比之下，一些方法只是使用基于多层感知机（MLP）的接口来弥合模态差距[16]。例如，LLaVA 系列采用多层感知机[16,39]来投影视觉令牌，并将特征维度与词嵌入对齐。BLIVA [58]采用基于多层感知机和基于 Q-Former 的连接器的集成来提高在文本丰富的场景中的性能。

作为另一条路线，特征级融合插入额外的模块，以实现文本特征和视觉特征之间的深度交互和融合。例如，Flamingo [59]在大型语言模型的冻结Transformer层之间插入额外的交叉注意力层，从而用外部视觉线索扩充语言特征。同样，CogVLM [60]在每个Transformer层中插入视觉专家模块，以实现视觉和语言特征之间的双重交互和融合。为了获得更好的性能，引入模块的QKV权重矩阵

是从预训练的大型语言模型初始化的。同样，LLaMA-Adapter [61] 将可学习的提示引入到 Transformer 层中。这些提示首先嵌入视觉知识，然后作为前缀与文本特征连接起来。

顺便说一下，MM1 [41] 对连接器的设计选择进行了消融实验，发现对于标记级融合，模态适配器的类型远不如视觉标记的数量和输入分辨率重要。然而，Zeng 等人[62]比较了标记级和特征级融合的性能，并经验性地表明，在视觉问答基准方面，标记级融合变体表现更好。关于性能差距，作者认为，交叉注意力模型可能需要更复杂的超参数搜索过程才能实现相当的性能。

就参数规模而言，与编码器和大型语言模型相比，可学习接口通常只占很小一部分。以 Qwen-VL [28] 为例，Q-Former 的参数规模约为 0.08B，占整个参数的比例不到 1%，而编码器和大型语言模型分别占约 19.8%（1.9B）和 80.2%（7.7B）。

专家模型。除了可学习的接口外，使用专家模型，如图像描述模型，也是一种可行的方式来弥合模态差距[63]。其基本思想是在不进行训练的情况下将多模态输入转换为语言。通过这种方式，大型语言模型可以通过转换后的语言来理解多模态。例如，VideoChat-Text[21]使用预训练的视觉模型提取诸如动作等视觉信息，并使用语音识别模型丰富描述。虽然使用专家模型简单直接，但它可能不如采用可学习的接口灵活。将外部模态转换为文本会导致信息损失。例如，将视频转换为文本描述会扭曲时空关系[21]。

训练策略与数据

一个成熟的 MLLM 会经历三个训练阶段，*即*预训练、指令调整和对齐调整。每个训练阶段都需要不同类型的数据，并实现不同的目标。在本节中，我们将讨论训练目标以及每个训练阶段的数据收集和特征。

Input: <image>
Response: {caption}

表 3 。一个用于构建标题的简化模板

data. {<图像>} 是视觉标记的占位符，{标题} 是图像的标题。请注意，只有标记为红色的部分用于损失计算。

“预训练”

训练详情

作为第一个训练阶段，预训练主要旨在协调不同的模态，并学习多模态的世界知识。预训练阶段通常需要大规模的文本配对数据，*例如*标题数据。通常，标题对以自然语言描述图像/音频/视频。

在这里，我们考虑了一个常见的场景，即训练多模态语言模型（MLLMs）以实现视觉与文本的对齐。如表3所示，给定一张图像，模型被训练以自回归的方式预测图像的标题，遵循标准的交叉熵损失。一种常见的预训练方法是冻结预训练的模块（例如视觉编码器和大型语言模型），并训练一个可学习的接口[16]。其思想是在不损失预训练知识的情况下对齐不同模态。一些方法[28]还解锁了更多的模块（例如视觉编码器），以允许更多的可训练参数进行对齐。需要注意的是，训练方案与数据质量密切相关。对于简短且嘈杂的标题数据，使用较低分辨率（例如224）可以加快训练过程，而对于较长且更清洁的数据，最好使用更高分辨率（例如448或更高）以减轻幻觉。此外，ShareGPT4V[64]发现，在预训练阶段使用高质量的标题数据，解锁视觉编码器可以促进更好的对齐。

数据

预训练数据主要有两个用途，*即*（1）对齐不同的模态，（2）提供世界知识。根据粒度，预训练语料库可以分为粗粒度和细粒度数据，我们将依次介绍。我们在表4中总结了常用的预训练数据集。

粗粒度的标题数据有一些典型的共同特征：（1）由于样本通常来自互联网，数据量很大。（2）由于网络爬取的性质，标题通常简短且嘈杂，因为它们源自网络图像的替代文本。这些数据可以通过自动工具进行清理和筛选，例如，使用

CLIP [12] 模型用于过滤掉相似度低于预先设定阈值的图像 - 文本对。接下来，我们将介绍一些具有代表性的粗粒度数据集。

CC. CC-3M [65] 是一个包含 330 万对图像-标题的网络规模字幕数据集，其中原始描述源自与图像相关的替代文本。作者设计了一个复杂的数据清理流程：（1）对于图像，过滤掉内容不当或纵横比不合适的图像。（2）对于文本，使用自然语言处理工具获取文本注释，并根据设计的启发式方法对样本进行筛选。（3）对于图像-文本对，通过分类器为图像分配标签。如果文本注释与图像标签不重叠，则丢弃相应的样本。

CC-12M [66] 是 CC-3M 的后续工作，包含 1240 万对图像-标题。与之前的工作相比，CC-12M 放宽并简化了数据收集流程，从而收集了更多的数据。

SBU 标题[67]。这是一个带有标题的照片数据集，包含100万对图像-文本对，图像和描述来自Flickr网站。具体来说，通过使用大量查询词对Flickr网站进行查询，获取初始的图像集。因此，图像的描述作为标题。然后，为了确保描述与图像相关，保留的图像满足以下要求：（1）图像的描述长度令人满意，通过观察来决定。（2）标题应该至少包含两个预定义术语列表中的单词和一个表示空间关系的介词（例如“在”、“下”）。

LAION。该系列是大型网络规模数据集，其中图像是从互联网上抓取的图片，并附有相关的替代文本作为标题。为了筛选图像-文本对，执行以下步骤：（1）长度过短的文本或大小过小或过大的图像被丢弃。（2）基于 URL 进行图像去重。（3）提取 CLIP [12] 嵌入用于图像和文本，并使用嵌入来丢弃可能非法的内容以及嵌入之间的余弦相似度低的图像-文本对。在此，我们对一些典型的变体进行简要总结：

- LAION-5B [68]：这是一个用于研究目的的数据集，包含 58.5 亿对图像 - 文本。该数据集是多语言的，其中包含一个 20 亿的英语子集。
- LAION-COCO [69]：它包含从 LAION-5B 的英语子集中提取的 6 亿张图像。这些标题是合成的，使用 BLIP [70] 生成各种图像标题，并使用 CLIP [12] 选择最合适的。

COYO-700M [71]。它包含 7.47 亿个图像 - 文本。

表 4。用于预训练的常见数据集。

数据集	样本	日期
粗粒度的图像 - 文本		
CC-3M [65]	3.3米	2018
CC-12M [66]	12.4兆	2020
SBU 标题 [67]	1米	2011
LAION-5B [68]	5.9B	2022 年 3 月
LAION-2B [68]	2.3B	2022 年 3 月
拉伊恩 - 科科 [69]	6亿	2022年9月
COYO-700M [71]	747M	2022年8月
细粒度的图像 - 文本		
共享 GPT4V-PT [64]	1.2米	2023年11月
LVIS-Instruct4V [72]	11.1万	2023年11月
阿拉瓦 [73]	709,000	2024年2月
视频 - 文本		
MSR-VTT [74]	20 万	2016
“音频 - 文本”		
WavCaps [75]	24K	2023年3月

对从 CommonCrawl 中提取的图像-文本对进行数据过滤。在数据过滤方面，作者设计了以下策略来过滤掉数据样本：（1）对于图像，那些大小、内容、格式或纵横比不合适的图像会被过滤掉。此外，基于 pHash 值来过滤掉与诸如 ImageNet 和 MS-COCO 等公共数据集重叠的图像。（2）对于文本，只有长度、名词形式和用词恰当的英语文本才会被保存。句子前后的空格会被删除，连续的空格字符会被替换为一个空格。此外，出现超过 10 次（例如“image for”）的文本会被丢弃。（3）对于图像-文本对，基于（图像 pHash，文本）元组来删除重复的样本。

最近，更多的研究[64,73]探索了通过提示强大的多模态语言模型（例如GPT-4V）来生成高质量的细粒度数据。与粗粒度数据相比，这些数据通常包含对图像更长、更准确的描述，从而能够在图像和文本模态之间实现更精细的对齐。然而，由于该方法通常需要调用商业用途的多模态语言模型，成本较高，数据量较小。值得注意的是，ShareGPT4V[64]通过首先使用GPT-4V生成的10万数据训练一个描述生成器，然后使用预训练的描述生成器将数据量扩展到120万，从而实现了平衡。

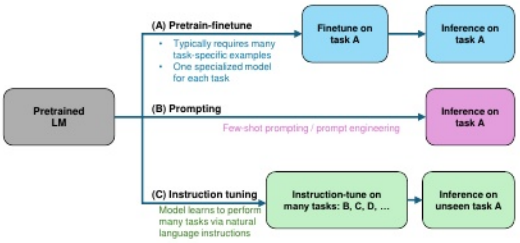

图 3。三种典型学习范例的比较，改编自[76]。

“指令调整”

引言

指令是指对任务的描述。直观地说，指令调整旨在教导模型更好地理解用户的指令并完成所需的任务。通过这种方式进行调整，大型语言模型可以通过遵循新的指令来泛化到未见过的任务，从而提高零样本性能。这个简单而有效的想法引发了后续自然语言处理工作的成功，如 ChatGPT [77]、InstructGPT [78]。

指令调整与相关典型学习范例之间的比较如图 3 所示。监督式微调方法通常需要大量特定任务的数据来训练特定任务模型。提示方法减少了对大规模数据的依赖，并且可以通过提示工程来完成特定任务。在这种情况下，尽管少样本性能有所提高，但零样本性能仍然相当一般[4]。不同之处在于，指令调整学习的是如何泛化到未见过的任务，而不是像这两个对应方法那样适应特定任务。此外，指令调整与多任务提示[79]和学习[80]高度相关。

在本节中，我们描述了指令样本的格式、训练目标、收集指令数据的典型方式以及相应的常用数据集。

训练详情

多模态指令样本通常包括一个可选的指令和一个输入输出对。指令通常是描述任务的自然语言句子，例如，“*详细描述图像*”。输入可以是像视觉问答任务[82]那样的图像-文本对，也可以是像图像描述任务[83]那样的仅图像。输出是针对输入的指令的回答。指令模板是灵活的，并受到人工设计的约束[21]，如表5所示。请注意，指令模板也可以推广到多轮人类-代理对话的情况[16,81]。

形式上，多模态指令样本可以以三元组的形式表示，*即*（I，M，R），其中 I、M、R 分别表示指令、多模态输入和真实响应。多模态语言模型（MLLM）根据指令和多模态输入预测一个答案：

$$\mathcal{A} = f(\mathcal{I}, \mathcal{M}; \theta) \tag{1}$$

在此，A 表示预测的答案，$_{\theta}$are 表示模型的参数。训练目标通常是用于训练大型语言模型（LLM）的原始自回归目标[16]，基于此，鼓励多语言语言模型（MLLM）依次预测响应的下一个标记：

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(\mathcal{R}_i | \mathcal{I}, \mathcal{R}_{<i}; \theta) \tag{2}$$

其中 $_{N}$is 为真实长度的长度 。

数据收集

由于指令数据在格式上更具灵活性，在任务表述上更加多样，因此收集数据样本通常更棘手且成本更高。在本节中，我们总结了三种大规模获取指令数据的典型方式，*即*数据适配、自我指导和数据混合。

数据适应。特定任务的数据集是高质量数据的丰富来源。因此，大量研究[51,84]利用现有的高质量数据集来构建指令格式的数据集。以视觉问答（VQA）数据集的转换为例：原始样本是一个输入输出对，其中输入包括图像和自然语言问题，输出是针对图像对问题的文本回答。这些数据集的输入输出对可以自然地包含指令样本的多模态输入和响应。指令，*即*任务的描述，可以来自人工

Below is an instruction that describes a task. Write a response that appropriately completes the request

Instruction: <instruction>
Input: {<image>, <text>}
Response: <output>

表 5. 用于构建多内容的简化模板

模态指令数据。<指令>是对任务的文本描述。{<图像>、<文本>}和 <输出> 是数据样本的输入和输出。请注意，在某些数据集中，输入中的<文本>可能会缺失，例如图像描述数据集仅仅有<图像>。该示例改编自[81]。

- <Image> {Question}
- <Image> Question: {Question}
- <Image> {Question} A short answer to the question is
- <Image> Q: {Question} A:
- <Image> Question: {Question} Short answer:
- <Image> Given the image, answer the following question with no more than three words. {Question}
- <Image> Based on the image, respond to this question with a short answer: {Question}. Answer:
- <Image> Use the provided image to answer the question: {Question} Provide your answer as short as possible:
- <Image> What is the answer to the following question? "{Question}"
- <Image> The question "{Question}" can be answered using the image. A short answer is

表 6. 视觉问答（VQA）数据集的指令模板，引自[51]。<图像>和{问题}分别是原始 VQA 数据集中的图像和问题。

设计或从由 GPT 辅助的半自动生成中生成。具体而言，一些工作[17]在训练期间手动制作一组候选指令并从中抽取一条。我们给出了一个关于视觉问答（VQA）数据集的指令模板示例，如表 6 所示。其他工作手动设计一些种子指令，并使用这些种子指令提示 GPT 生成更多[21]。

请注意，由于现有的视觉问答（VQA）和图像描述数据集的答案通常很简洁，直接使用这些数据集进行指令调整可能会限制多语言语言模型（MLLM）的输出长度。解决这个问题有两种常见策略。第一种是在指令中明确指定相应的要求。例如，ChatBridge [85] 明确声明*短*回答数据应*简*短明了。第二种是扩展现有答案的长度[86]。例如，M^3IT [86] 提出通过向 ChatGPT 提示原始问题、答案以及图像的上下文信息（例如通过光学字符识别提取的图像描述和文本）来重新表述原始答案。

自我指导。尽管现有的多任务数据集可以提供丰富的数据来源，但在现实场景中，如多轮对话中，它们通常不能很好地满足人类的需求。为了解决这个问题，一些工作通过自我指导来收集样本[89]，利用大型语言模型（LLMs）使用少量手动标注的样本生成遵循文本指令的数据。具体来说，一些遵循指令的样本被手工制作为演示，然后提示 ChatGPT/GPT-4以这些演示为指导生成更多的指令样本。LLaVA[16]将这种方法扩展到多模态领域，通过将图像转换为标题和边界框的文本，并提示纯文本的GPT-4在需求和演示的指导下生成新数据。通过这种方式，构建了一个多模态指令数据集，称为LLaVA-Instruct-150k。遵循这一想法，后续的工作如MiniGPT-4[17]和GPT4Tools[90]开发了不同的数据集以满足不同的需求。最近，随着更强大的多模态模型GPT-4V的发布，许多研究采用了GPT-4V来生成更高质量的数据，例如LVIS-Instruct4V[72]和ALLaVA[73]。我们在表7中总结了通过自指导生成的流行数据集。值得注意的是，这种范例高度依赖于先进但闭源的模型，这可能对数据扩展来说成本高昂。这种方法可能是由于早期模型的能力有限。未来的研究可以探索利用开源模型来生成高质量的指令数据。

数据混合。除了多模态的指令数据外，纯语言的用户辅助对话数据也可用于提高对话能力和指令遵循能力[91]。LaVIN [91]通过从纯语言和多模态数据中随机采样直接构建小批量数据。MultiInstruct [84]探索了使用单模态和多模态数据融合进行训练的不同策略，包括混合指令调整（同时使用两种类型的数据并随机打乱）和顺序指令调整（先使用文本数据，然后使用多模态数据）。

数据质量

近期研究表明，指令调整样本的数据质量与数量同等重要。Lynx [62] 发现，在大规模但有噪声的图像上进行预训练的模型

表 7。一份由自我指导生成的热门数据集的总结。对于输入/输出模态，I：图像，T：文本，V：视频，A：音频。对于数据构成，M-T 和 S-T 分别表示多轮和单轮。

数据集	示例	模态	来源	合成
LLaVA 指令 [16]	158,000	我 + 时间 → 时间	“MS-COCO”	23000 条标题 + 58000 个机器翻译问答 + 77000 条推理
低视力辅助系统指导版 [72]	220,000	我 + 时间 → 时间	低视力辅助器具	11 万条字幕 + 11 万次机器翻译问答
阿拉瓦 [73]	1.4米	我 + 时间 → 时间	VFlan，拉尼翁	709K 条字幕 + 709K 条 S-T 质量保证
视频聊天 GPT [87]	十万	V + T → T	活动网	7K 分辨率描述 + 4K 移动电视质量保证
视频聊天 [21]	11000	V + T → T	网络视频	描述 + 总结 + 创造
克洛索 - 细节 [88]	3900	A + T → T	克洛索	标题

文本对的表现不如使用较小但更干净的预训练数据集所训练的模型。*同样，魏*等人[92]发现，质量更高的少量指令调整数据可以实现更好的性能。对于数据过滤，该工作提出了一些评估数据质量的指标，并相应地提出了一种自动过滤掉劣质视觉-语言数据的方法。在此，我们讨论数据质量的两个重要方面。

提示的多样性。人们发现，指令的多样性对于模型的性能至关重要。Lynx [62] 通过实证验证，多样化的提示有助于提高模型的性能和泛化能力。

任务覆盖率。*就训练数据所涉及的任务而言，Du* 等人[93]进行了一项实证研究，发现视觉推理任务在提升模型性能方面优于描述和问答任务。此外，该研究表明，更复杂的指令优于增加任务多样性和纳入细粒度的空间注释。

校准调优

引言

对齐调整更常用于模型需要与特定人类偏好相符的场景，*例如*响应中减少幻觉。目前，强化学习结合人类反馈（RLHF）和直接偏好优化（DPO）是两种主要的对齐调整技术。在本节中，我们将依次介绍这两种技术的主要思想，并提供一些它们在解决实际问题中的应用示例，最后，给出相关数据集的汇总。

训练详情

强化学习人类反馈（Reinforcement Learning from Human Feedback，RLHF）[94,95]。该技术旨在利用强化学习算法使大型语言模型（LLMs）符合人类偏好，在训练循环中，以人类标注作为监督。正如 InstructGPT [78] 所举例说明的，强化学习人类反馈包含三个关键步骤：

(1) *监督微调*。此步骤旨在对预训练模型进行微调，以呈现初步期望的输出行为。在 RLHF 环境中，经过微调的模型被称为*策略模型*。请注意，由于监督策略模型 π^{SFT} 可以从指令调整的模型初始化，因此此步骤可能会被跳过。

(2) *奖励建模*。在本步骤中，使用偏好对训练*奖励模型*。给定一个多模态提示（例如图像和文本）$_X$and 一个响应对（Y W，$_{YL}$）），奖励模型 $R_\theta$learns 给偏好的响应 Y_W 更高的奖励，反之亦然，目标是：

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log(\sigma(r_\theta(x,y_w) - r_\theta(x,y_l))\Big] \tag{3}$$

其中 D = {(X，Y_W，Y_L)} 是由人类标注员标注的比较数据集。在实践中，奖励模型 $R_\theta$shares 具有与策略模型类似的结构。

（3）*强化*学习。在此步骤中，采用近端策略优化（PPO）算法来优化强化学习策略模型 π^{RL}。每个标记的 KL 惩罚它经常被添加到训练目标中，以避免偏离原始策略太远[78]，从而得出该目标：

$$\mathcal{L}(\phi) = -\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\phi^{RL}(y|x)}\Big[r_\theta(x,y) - \beta\cdot\mathbb{D}_{KL}\left(\pi_\phi^{RL}(y|x)||\pi^{REF}(y|x)\right)\Big] \tag{4}$$

其中 $\beta$is 是 KL 惩罚项的系数。通常，强化学习策略 π^{RL} and 和参考模型 π^{REF} 均从监督模型 π^{SFT} 中初始化。通过这一调整过程，期望获得的强化学习策略模型能符合人类的偏好。

研究人员已经探索了使用 RLHF 技术以实现更好的多模态对齐。对于

表 8。用于校准调优的数据集摘要。对于输入/输出模态，I：图像，T：文本。

数据集	示例	模态	来源
拉瓦 - 右心心力衰竭 [96]	一万	我 + 时间 → 时间人类	
右心室肥厚型心肌病 - V 型[96]	5千	我 + 时间 → 时间人类	
虚拟现实反馈 [99]	38 万	我 + 时间 → 时间GPT-4V	

例如，LLaVA-RLHF [96] 收集人类偏好数据，并基于 LLaVA [16] 对模型进行调优，使其产生更少的幻觉。

DPO [97]。它利用简单的二元分类损失从人类偏好标签中学习。与基于 PPO 的 RLHF 算法相比，DPO 无需学习明确的奖励模型，从而将整个流程简化为两个步骤，*即*人类偏好数据收集和偏好学习。该算法的学习目标如下：

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\phi^{\mathrm{RL}}(y_w|x)}{\pi^{\mathrm{REF}}(y_w|x)} - \beta\log\frac{\pi_\phi^{\mathrm{RL}}(y_l|x)}{\pi^{\mathrm{REF}}(y_l|x)}\right)\right] \tag{5}$$

RLHF-V [98] 通过纠正模型响应中的幻觉来收集细粒度（片段级别）的偏好数据对，并使用获得的数据执行密集的 DPO。Silkie [99] 则通过提示 GPT-4V 来收集偏好数据，并通过 DPO 将偏好监督提炼为一个指令调整的模型。

数据

校准调优的数据收集的要点是收集模型响应的反馈，*即*决定哪种响应更好。收集此类数据通常成本更高，而且用于这一阶段的数据量通常甚至少于前几个阶段。在本部分，我们介绍一些数据集，并在表8中对其进行了总结。

LLaVA-RLHF [96]。它包含从人类关于诚实和乐于助人的反馈中收集到的 10,000 对偏好对。该数据集主要用于减少幻觉。

RLHF-V [98]。它拥有通过执行片段级别的幻觉校正所收集的 5700 条精细的人类反馈数据。

VLFeedback [99]。它利用人工智能就模型响应提供反馈。该数据集包含超过 38 万对由 GPT-4V 根据有用性、忠实性和伦理问题进行评分的对比。

评估

评估是开发多模态语言模型（MLLMs）的重要组成部分，因为它为模型优化提供了反馈，并有助于比较不同模型的性能。与传统多模态模型的评估方法相比，多模态语言模型的评估具有几个新特点：（1）由于多模态语言模型通常是多用途的，因此对其进行全面评估非常重要。（2）多模态语言模型表现出许多需要特别关注的新兴能力（例如，无OCR的数学推理），因此需要新的评估方案。根据问题类型，多模态语言模型的评估可以分为两类，包括封闭集和开放集。封闭集评估通常涉及特定任务基准和专门为多模态语言模型设计的更全面的基准，其中答案仅限于预定义的集合。开放集评估通常包括人工评分、GPT评分和案例研究。

封闭集

封闭式问题是指一种可能答案选项是预先定义的，并且仅限于有限集合的问题。评估通常在特定任务的数据集上进行。在这种情况下，响应可以通过基准指标自然地进行判断。例如，InstructBLIP [51]报告了在ScienceQA [100]上的准确性，以及NoCaps [102]上的CIDEr分数[101]。评估设置通常是零样本[51, 84]或微调[29,51]。第一种设置通常选择涵盖不同一般任务的广泛数据集，并将其分为保留数据集和保留外数据集。在前者上进行调优后，零样本性能在后者上使用未见过的数据集甚至未见过的任务进行评估。相比之下，第二种设置通常在特定任务评估中观察到。例如，LLaVA [16]报告了在ScienceQA [100]上的微调性能。LLaVA-Med [29]报告了在生物医学视觉问答[103]上的结果。

上述评估方法通常局限于一小部分选定的任务或数据集，缺乏全面的定量比较。为此，一些研究努力专门为多语言语言模型开发新的基准测试[104,105]。例如，傅等人*[104]构建*了一个全面的评估基准测试MME，其中包括总共14个感知和认知任务。MME中的所有指令-答案对都是手动的。

旨在避免数据泄露。MMbench [105] 是一个专门为评估模型能力多个维度而设计的基准测试，使用 ChatGPT 将开放性响应与预先定义的选项进行匹配。Video-ChatGPT [87] 和 Video-Bench [106] 专注于视频领域，并提出专门的基准测试以及评估工具。

“开放集”

与封闭式问题不同，对开放式问题的回答可以更加灵活，其中多语言语言模型通常会扮演聊天机器人的角色。由于聊天的内容可以是任意的，因此判断起来比封闭式输出更棘手。标准可以分为人工评分、GPT 评分和案例研究。人工评分需要人类评估生成的响应。这种方法通常涉及手工制作的问题，旨在评估特定的维度。例如，mPLUG-Owl [107] 收集了一个与视觉相关的评估集，以判断对自然图像、图表和流程图理解的等能力。同样，GPT4Tools [90] 分别构建了两组用于微调零样本性能，并从思想、行动、论点和整体方面对响应进行评估。

由于人工评估劳动强度大，一些研究人员探索使用 GPT 进行评分，即 GPT 评分。这种方法常用于评估多模态对话的表现。LLAVA [16] 提出通过纯文本的 GPT-4 从不同方面（如有用性和准确性）对响应进行评分。具体而言，从 COCO [108] 验证集中抽取 30 幅图像，每幅图像都与一个简短问题、一个详细问题和一个复杂推理问题相关联，这些问题通过在 GPT-4 上的自指导生成。将模型和 GPT-4 生成的答案都发送给 GPT-4 进行比较。后续的工作遵循这一思路，提示 ChatGPT 或 GPT-4 对结果进行评分[29]或判断哪一个更好[109]。

应用纯文本的 GPT-4 进行评估的一个主要问题是，判断仅仅基于翻译后的文本内容，例如标题或边界框坐标，而不访问图像[29]。因此，在这种情况下，将 GPT-4 设定为性能上限可能存在疑问。随着 GPT 视觉接口的发布，一些工作利用更先进的 GPT-4V 模型来评估多模态语言模型（MLLMs）的性能。例如，Woodpecker [63]采用 GPT-4V 模型来判断模型回答的响应质量。由于 GPT-4V 可以直接访问图像，这种评估预计会比使用纯文本的 GPT-4 更准确。

由于基准评估不够全面，一种补充方法是通过案例研究来比较多语言大型语言模型（MLLMs）的不同能力。例如，一些研究通过在各种领域和任务上制作一系列样本，从初步技能（如标题和物体计数）到需要世界知识和推理的复杂任务（如笑话理解和作为实体代理的室内导航），对GPT-4V进行了深入的定性分析。Wen等人[111]通过设计针对自动驾驶场景的样本对GPT-4V进行了更集中的评估。Fu等人[112]通过将Gemini-Pro与GPT-4V进行比较，对Gemini-Pro进行了全面的评估。结果表明，尽管GPT-4V和Gemini在响应风格上有所不同，但它们在视觉推理能力方面相当。

扩展

近期的研究在扩展多语言语言模型（MLLMs）的能力方面取得了重大进展，涵盖了从更强大的基础能力到更广泛的场景覆盖范围。在此方面，我们追溯了多语言语言模型的主要发展。

粒度支持。为了促进代理和用户之间的更好交互，研究人员开发了具有更精细粒度支持的MLLM。在输入方面，支持从用户提示中更精细控制的模型逐渐发展，从图像到区域[24]，甚至像素[25]。具体来说，Shikra[24]支持区域级别的输入和理解。用户可以通过参考自然语言形式的边界框中的特定区域更灵活地与助手交互。Ferret[113]更进一步，通过设计混合表示方案支持更灵活的引用。该模型支持不同形式的提示，包括点、框和草图。同样，Osprey[25]通过利用分割模型[10]支持点输入。在预训练分割模型的卓越能力的帮助下，Osprey能够通过一次点击指定单个实体或其一部分。在输出方面，随着输入支持的开发，接地能力也得到了改善。Shikra [24] 支持基于图像的响应，通过框标注来实现，从而提高了精度和精细度。

Ferring 经验。LISA [114] 进一步支持掩码级别的理解和推理，这使得像素级别的锚定成为可能。

模态支持。对模态的更多支持是 MLLM 研究的一个趋势。一方面，研究人员探索调整 MLLM 以支持更多多模态内容的输入，例如 3D 点云[115]。另一方面，MLLM 也被扩展以生成更多模态的响应，例如图像[116]、音频[117]和视频[118]。例如，NExT-GPT [119] 提出了一个框架，在扩散模型[120]的帮助下，支持混合模态的输入和输出，特别是文本、图像、音频和视频的组合，该框架应用了编码器 - 解码器架构，并将 LLM 作为理解和推理的枢纽。

语言支持。当前的模型主要是单语言的，这可能是由于高质量的非英语训练语料库稀缺。一些工作致力于开发多语言模型，以便覆盖更广泛的用户群体。Vis-CPM [121]通过设计一个多阶段训练方案将模型能力转移到多语言环境中。具体来说，该方案将英语作为关键语言，因为英语有大量的训练语料库。利用预训练的双语大型语言模型，通过在指令调整期间添加一些翻译样本，将多模态能力转移到中文。采用类似的方法，Qwen-VL [28]是从双语大型语言模型Qwen [48]开发的，支持中文和英文。在预训练期间，将中文数据混合到训练语料库中，以保留模型的双语能力，占整个数据量的22.7%。

场景/任务扩展。除了开发通用的通用助手之外，一些研究还聚焦于应考虑实际条件的更特定场景，而其他研究则将多语言语言模型扩展到具有特定专业知识的下游任务。

一个典型的趋势是将多模态语言模型适应于更具体的现实生活场景。例如，一些工作开发了与现实世界交互的代理，*例如*专门为图形用户界面（GUI）设计的用户友好型助手，如 CogAgent [32]、AppAgent [122] 和 Mobile-Agent [123] 所例证的。研究人员还开发了能够进行推理、导航和操作的具身代理[19,31]，促进了能够为人类执行任务的自动代理的发展。总的来说，这些助手在规划和执行每一步以完成用户指定的任务方面表现出色，充当着对人类有帮助的代理。

另一条路线是增强多语言语言模型（MLLMs）以具备特定技能，用于解决不同领域的任务，*例如*文档理解[30]和医疗领域[29]。对于文档理解，mPLUG-DocOwl [124] 利用各种形式的文档级数据进行调优，从而在无光学字符识别（OCR）的文档理解中增强模型。TextMonkey [30]结合了与文档理解相关的多个任务以提高模型性能。同样，多语言语言模型也可以训练以适应传统的视觉任务，如视觉接地[125,126]。与传统方法[13,127]相比，多语言语言模型统一了输入/输出格式，并简化了整个学习和推理过程。具体而言，在统一的语言建模目标下，将接地任务重新表述为条件化的框坐标预测任务是可行的[24,28,52]。该模型经过训练，以自然语言的形式预测指定对象的坐标。多语言语言模型还可以通过注入专业知识扩展到医疗领域。例如，LLaVA-Med [29]通过注入领域知识开发了专门从事医学图像理解和问答的助手。

高效的 MLLM。近来，使用轻量级的 MLLM 进行高效部署越来越受欢迎[128 - 130]。这些模型经过精心设计和优化，在不牺牲太多模型性能的情况下，更经济地利用资源或适应资源有限的场景。

从模型的角度来看，已经探索了各种技术来促进高效的训练和推理。例如，MobileVLM [54] 探索为资源有限的情况开发多模态语言模型的小型变体。一些设计和技术被用于在移动设备上的部署，例如较小规模的多模态语言模型和量化技术以加快计算速度。同样，MiniCPM-V [129] 为端侧计算构建高效的多模态语言模型。采用 Q-Former [28] 来减少图像每个补丁的视觉标记数量。

从数据的角度来看，Bunny [130] 全面研究了用于模型训练的高效数据选择和组合方案。所获得的模型在性能上与参数规模更大的多语言语言模型相当。

多模态幻觉

多模态幻觉是指由多模态语言模型生成的响应与图像内容不一致的现象[63]。这一根本问题受到了越来越多的关注。在本节中，我们简要介绍相关概念和研究进展。

预备工作

多模态幻觉可分为三种类型[131]:

1. *存在幻觉*是一种常见的类型，这意味着模型错误地判定物体的存在。
2. *属性幻觉*指的是错误地描述某些物体的属性，*例如*无法辨别一只狗的颜色。
3. *关系幻觉*是一种更复杂类型的幻觉。它指的是对物体之间关系（如相对位置）的错误描述。

接下来，我们首先介绍评估方法，这些方法有助于衡量减轻幻觉的方法的性能。然后，我们讨论不同类型方法的减轻方法。

评估方法

CHAIR [132] 是一个早期用于评估开放式标题中幻觉水平的度量标准。该度量标准测量具有幻觉对象的句子比例或所有提及对象中幻觉对象的比例。相比之下，POPE [133] 是一种评估封闭集选择的方法。具体而言，制定了具有二元选择的多个提示，每个提示都询问特定对象是否存在于图像中。采用类似的评估方法，MME [104] 提供了更全面的评估，涵盖了存在性、数量、位置和颜色等方面，如[63]中所举例说明的。

与之前使用匹配机制来检测和判定幻觉的方法不同，一些研究探索通过模型自动评估文本响应。例如，HaELM [134] 提出使用大型语言模型（LLM）作为评判者，根据参考标题来决定多语言模型（MLLM）的标题是否正确。鉴于纯文本的大型语言模型只能访问有限的图像上下文并且需要参考注释，Woodpecker [63] 使用 GPT-4V 直接评估基于图像的模型响应。

缓解方法

根据减轻幻觉的高级思路，当前的方法大致可以分为三类：前校正、过程中校正和后校正。

预校正。对于幻觉的一个直观解决方案是收集专门的数据（例如负面数据）并用这些数据进行微调，从而实现产生幻觉更少的模型。

LRV-Instruction [135] 引入了一个视觉指令调整数据集以鼓励忠实生成。同样，LLaVA-RLHF [96] 收集人类偏好对，并使用强化学习技术对模型进行微调。

在生成过程中的校正。另一条路线是改进建筑设计或特征表示。这些工作试图探究幻觉产生的原因，并在生成过程中设计缓解措施。例如，HallE-Switch [131] 引入了一个连续的控制因素，用于在推理过程中控制模型输出中的想象程度。

后校正。与之前的范例不同，后校正以一种补救后的方式减轻幻觉。例如，Woodpecker [63] 是一个用于幻觉校正的无需训练的框架。具体而言，该方法纳入专家模型以补充图像的上下文信息，并构建了一个逐步校正幻觉的管道。

扩展技术

多模态的上下文学习

ICL 是大型语言模型（LLMs）的重要新兴能力之一。该技术的本质是以少数几个例子作为提示引导模型，使模型更容易回答查询。ICL 有两个良好的特点：（1）ICL 的关键在于从类比中学习[136]，从而大大减少了对数据样本的需求。（2）ICL 通常在无需训练的方式中实现[136]，并且在推理时可以灵活地集成到各种框架中。

在多模态语言建模（MLLM）的背景下，指令微调学习（ICL）已被扩展到更多的模态，从而产生了多模态指令微调学习（M-ICL）。在推理时，M-ICL 可以通过向原始样本添加演示集（*即*一组上下文样本）来实现。在这种情况下，模板可以如表 9 所示进行扩展。

集成计算实验室（ICL）能力的提升

近来，越来越多的工作集中于在各种情况下提高 ICL 的性能。

<BOS> Below are some examples and an instruction that describes a task. Write a response that appropriately completes the request

Instruction: {instruction}
Image: <image>
Response: {response}

Image: <image>
Response: {response}

Image: <image>
Response: <EOS>

表 9. 模板转换为结构的简化示例

将一个 M-ICL 查询改编自[81]。例如，我们列出了两个上下文示例和一个查询，用虚线分隔。{指令}和{响应}是数据样本中的文本。<图像>是代表多模态输入（在此例中为图像）的占位符。<BOS>和<EOS>是分别表示输入到 LLM 的起始和结束的标记。

情景。在本节中，我们追溯该领域的发展，并总结相关工作。

MIMIC-IT [137] 通过构建一个以多模态上下文格式呈现的指令数据集，将上下文学习与指令调整相结合。其他一些工作则探索在特定设置下提高少样本学习的性能。例如，Link-context 学习[138]关注演示和查询之间的因果关系，并通过制定正负图像描述对来制定对比训练方案。*同样，Yang* 等人[139,140]探索了不同的策略来优化演示配置（上下文样本的选择或排序），以实现更好的少样本性能。

应用程序

在多模态应用方面，M-ICL 主要应用于两种场景：（1）解决各种视觉推理任务[141]；（2）教导大型语言模型（LLM）使用外部工具[142,143]。前者涉及从几个特定任务的示例中学习，并泛化到新的但类似的问题上。相比之下，工具使用的示例更加精细，通常包括一系列步骤来完成任务。

多模态思维链

CoT 是“一系列中间推理步骤”[5]。该技术已被证明在复杂的推理任务中是有效的。其主要思想是促使大型语言模型不仅输出最终答案，还要输出得出答案的推理过程，类似于人类的认知过程。

受自然语言处理领域成功的启发，多项工作[144,145]提议将该技术扩展到多模态对话式教学（M-CoT）。我们首先介绍了获取多模态对话式教学能力的不同范例。然后，我们详细描述了多模态对话式教学更具体的方面，包括链结构和模式。

学习范例

获取 M-CoT 能力大致有三种方式，*即*通过微调以及无需训练的少/零样本学习。

直观地说，微调方法通常涉及为 M-CoT 学习整理特定的数据集。例如，Lu 等人[100]构建了一个带有讲座和解释的科学问答数据集 ScienceQA，它可以作为学习 CoT 推理的来源。

与微调相比，少样本/零样本学习在计算效率方面更具优势。少样本学习方法通常需要手工制作上下文示例来逐步教授推理步骤。相比之下，零样本学习方法直接使用设计的指令进行提示[144]。

链路配置

结构和长度是推理链的两个关键方面。就结构而言，当前的方法可以分为单链[100]和树形方法[146]。链长可分为自适应和预定义的形式。前者配置要求大型语言模型决定何时停止推理链[100]，而后者设置则以预定义的长度停止链[147]。

生成模式

我们将相关工作总结为（1）基于填补的模式和（2）基于预测的模式。具体而言，基于填补的模式需要推导周围上下文（前后的步骤）之间的步骤以填补逻辑空白[144]。相比之下，基于预测的模式则需要根据诸如指令和先前的推理历史等条件来扩展推理链[142]。

法律硕士辅助的视觉推理

引言

受到工具增强型大型语言模型（LLM）成功的启发[148]，一些研究人员进行了探索

调用外部工具或视觉基础模型以执行视觉推理任务的可能性。这些工作将大型语言模型作为具有不同角色的助手，构建特定任务或通用的视觉推理系统。

与传统的视觉推理模型相比，这些工作表现出几个良好的特点：（1）强大的泛化能力。这些系统配备了从大规模预训练中学习到的丰富的开放世界知识，能够轻松地泛化到未见过的对象或概念，具有出色的零/少样本性能[149]。（2）涌现能力。在大规模语言模型强大的推理能力的帮助下，这些系统能够执行复杂的任务，*例如*理解图像的深层含义[18]。（3）更好的交互性和控制性。传统模型通常只允许有限的控制机制，而基于大规模语言模型的系统则允许在用户友好界面（例如点击和自然语言查询）中进行更精细的控制[150]。

在这部分，我们先介绍在构建 LLM 辅助的视觉推理系统中所采用的不同训练范例。然后，我们深入探讨 LLM 在这些系统中所发挥的主要作用。

训练范例

根据训练范例，大型语言模型辅助的视觉推理系统可分为两类，*即*无训练型和微调型。

无需训练。由于预先训练的大型语言模型中存储了大量先验知识，一种直观且简单的方法是冻结预先训练的模型，并直接提示大型语言模型满足各种需求。根据设定，推理系统可进一步分为少样本模型[142]和零样本模型[150]。

微调。一些工作采用进一步的微调来提高关于工具使用的规划能力[90]，或者提高系统的定位能力[114]。例如，GPT4Tools [90]收集了一个与工具相关的指令数据集来对模型进行微调。

函数

关于法律硕士（LLM）在法律硕士辅助的视觉推理系统中究竟扮演何种角色，现有的相关工作分为三类:

- 法学硕士担任财务主管
- 法学硕士作为决策者
- 法学硕士作为语义的完善者

在接下来的部分中，我们将阐述大型语言模型如何发挥这些作用。

作为控制器的 LLM。在这种情况下，LLM 充当中央控制器，其作用是（1）将复杂任务分解为更简单的子任务/步骤，（2）将这些任务分配给适当的工具/模块。具体而言，LLM 被明确提示输出任务规划[151]，或者更直接地输出要调用的模块[90,142,143]。例如，VisProg [143]提示 GPT-3 输出一个可视化程序，其中每一行程序调用一个模块来执行一个子任务。

作为决策者的法律硕士。在这种情况下，复杂的任务往往以多轮的方式解决，通常是迭代的方式[152]。决策者通常（1）总结背景以决定是否完成任务，（2）以用户友好型的方式组织答案。

大型语言模型作为语义细化器。当大型语言模型被用作语义细化器时，研究人员主要利用其丰富的语言和语义知识。具体而言，大型语言模型经常被指示将信息整合到流畅的自然语言句子中[153]，或者根据不同的特定需求生成文本[149,150,154]。

挑战与未来方向

多模态语言模型（MLLMs）的发展仍处于初级阶段，因此有很大的改进空间，我们总结如下:

- 当前的 MLLM 在处理长上下文的多模态信息方面存在局限性。这限制了具有更多多模态标记的先进模型的发展，*例如*长视频理解和长文档中图像和文本的交错。
- 多语言大型语言模型（MLLMs）应当进行升级，以遵循更复杂的指令。例如，由于其先进的遵循指令的能力，生成高质量问答对数据的主流方法仍然是提示闭源的 GPT-4V，而其他模型通常无法实现此类目标。
- 在诸如 M-ICL 和 M-CoT 这样的技术方面仍有很大改进空间。目前对这两种技术的研究仍处于初级阶段，多语言语言模型（MLLMs）的相关能力仍然薄弱。因此，对潜在机制和改进的探索很有前景。
- 基于多模态语言模型（MLLMs）开发具身智能体是一个热门话题。开发能够与现实世界交互的此类智能体是有意义的。此类努力需要具备关键能力的模型，包括感知、推理、规划和执行。

- 安全问题。与大型语言模型类似，多语言大型语言模型也容易受到精心设计的攻击。换句话说，多语言大型语言模型可能会被误导而输出有偏见或不期望的响应。因此，提高模型安全性将是一个重要的研究课题。
- 跨学科研究。鉴于多语言模型强大的泛化能力和丰富的预训练知识，一个有前景的研究方向可能是利用多语言模型来促进自然科学领域的研究，*例如*利用多语言模型对医学图像或遥感图像进行分析。为了实现这一目标，向多语言模型注入特定领域的多模态知识可能是必要的。

结论

在本文中，我们对现有的多语言学习文献进行了综述，并对其主要方向进行了全面介绍，包括基本方法和相关扩展。此外，我们还强调了当前需要填补的研究空白，并指出了一些有前景的研究方向。我们希望这篇综述能为读者提供多语言学习的最新进展的清晰图景，并激发更多相关研究。鉴于多语言学习的时代才刚刚开始，我们将不断更新这篇综述，并希望它能激发更多的研究。一个收集最新论文的相关GitHub链接可在此处找到。

资金

这项工作部分得到了中国国家自然科学基金（62222213、62406264、U22B2059、U23A20319、62072423 和 61727809）和四川省自然科学基金委青年科学家基金（2023NSFSC1402）的支持。

作者贡献

C.F. 是项目负责人。S.Y.、C.F. 和 S.Z. 进行了文献综述。K.L.、X.S.、T.X. 和 E.C. 提供了相关领域的见解。S.Y.、C.F. 和 S.Z. 撰写了文章。所有作者都对手稿进行了讨论和建议。

*利益冲突声明*无声明。

参考文献

1. Zhao WX, Zhou K, Li J *et al.* A survey of large language models. *arXiv:2303.18223* ; . 1
2. Xu B and Poo Mm. Large language models and brain-inspired general intelligence. *Natl Sci Rev* 2023; 10: nwad267. 1
3. Peng B, Li C, He P *et al.* Instruction tuning with gpt-4. *arXiv:2304.03277* ; . 1
4. Brown T, Mann B, Ryder N *et al.* Language models are few-shot learners. *Conference on Neural Information Pro-cessing Systems*, volume 33 (2020) 1877–1901. 1, 3, 7
5. Wei J, Wang X, Schuurmans D *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Conference on Neural Information Processing Systems*, volume 35 (2022) 24824–24837. 1, 14
6. Li H. Deep learning for natural language processing: advantages and challenges. *Natl Sci Rev* 2018; 5: 24–26. 1
7. Zhao W. A panel discussion on ai for science: the opportunities, challenges and reflections. *Natl Sci Rev* 2024; nwae119. 1
8. Xie WJ and Warshel A. Harnessing generative ai to decode enzyme catalysis and evolution for enhanced engineering. *Natl Sci Rev* 2023; 10: nwad331. 1
9. Gong P, Guo H, Chen B *et al.* iearth: an interdisciplinary framework in the era of big data and ai for sustainable development. *Natl Sci Rev* 2023; 10: nwad178. 1
10. Kirillov A, Mintun E, Ravi N *et al.* Segment anything. *IEEE/CVF International Conference on Computer Vision* (2023) 4015–4026. 1, 11
11. Shen Y, Fu C, Chen P *et al.* Aligning and prompting every-thing all at once for universal visual perception. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 13193–13203. 1
12. Radford A, Kim JW, Hallacy C *et al.* Learning transferable visual models from natural language supervision. *Inter-national Conference on Machine Learning* (2021) 8748– 8763. 1, 3, 4, 6
13. Li J, Selvaraju R, Gotmare A *et al.* Align before fuse: Vision and language representation learning with momentum distillation. *Conference on Neural Information Processing Systems*, volume 34 (2021) 9694–9705. 1, 12
14. Wang P, Yang A, Men R *et al.* Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning*, volume 162 (2022) 23318–23340. 1
15. Cho J, Lei J, Tan H *et al.* Unifying vision-and-language tasks via text generation. *International Conference on Ma-chine Learning* (2021) 1931–1942. 1
16. Liu H, Li C, Wu Q *et al.* Visual instruction tuning. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 1, 2, 4, 5, 7, 8, 9, 10, 11
17. Zhu D, Chen J, Shen X *et al.* Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592* ; . 1, 3, 8
18. Yang Z, Li L, Wang J *et al.* Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv:2303.11381* ; . 1, 15
19. Driess D, Xia F, Sajjadi MS *et al.* Palm-e: An embodied multimodal language model. *International Conference on Machine Learning*, volume 202 (2023) 8469–8488. 1, 12

20. OpenAI. Gpt-4 technical report. *arXiv:2303.08774* ; . 1

21. Li K, He Y, Wang Y *et al.* Videochat: Chat-centric video understanding. *arXiv:2305.06355* ; . 2, 5, 7, 8, 9

22. Zhang H, Li X and Bing L. Video-llama: An instruction-tuned audio-visual language model for video understand-ing. *Conference on Empirical Methods in Natural Lan-guage Processing* (2023) . 2, 4

23. Deshmukh S, Elizalde B, Singh R *et al.* Pengi: An audio language model for audio tasks. *Conference on Neural Information Processing Systems* 2023; 36: 18090–18108. 2, 3

24. Chen K, Zhang Z, Zeng W *et al.* Shikra: Unleashing multi-modal llm' s referential dialogue magic. *arXiv:2306.15195* ; . 2, 11, 12

25. Yuan Y, Li W, Liu J *et al.* Osprey: Pixel understanding with visual instruction tuning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 28202– 28211. 2, 3, 11

26. Han J, Zhang R, Shao W *et al.* Imagebind-llm: Multi-modality instruction tuning. *arXiv:2309.03905* ; . 2, 3

27. Moon S, Madotto A, Lin Z *et al.* Anymal: An efficient and scalable any-modality augmented language model. *arXiv:2309.16058* ; . 2

28. Bai J, Bai S, Yang S *et al.* Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv: 2308.12966* ; . 2, 3, 5, 12

29. Li C, Wong C, Zhang S *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 2, 10, 11, 12

30. Liu Y, Yang B, Liu Q *et al.* Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv:2403.04473* ; . 2, 12

31. Huang J, Yong S, Ma X *et al.* An embodied generalist agent in 3d world. *International Conference on Machine Learning* (2024) . 2, 12

32. Hong W, Wang W, Lv Q *et al.* Cogagent: A visual language model for gui agents. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 14281– 14290. 2, 3, 12

33. Cherti M, Beaumont R, Wightman R *et al.* Reproducible scaling laws for contrastive language-image learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 2818–2829. 3, 4

34. Sun Q, Fang Y, Wu L *et al.* Eva-clip: Improved training techniques for clip at scale. *arXiv:2303. 15389* ; . 4

35. Chen Z, Wang W, Tian H *et al.* How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404 16821* ; . 4

36. Fang Y, Wang W, Xie B *et al.* Eva: Exploring the limits of masked visual representation learning at scale. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 19358–19369. 3

37. Bavishi R, Elsen E, Hawthorne C *et al.* Introduc-ing our multimodal models. https://www.adept.ai/blog/fuyu-8b (17 October 2024, date last accessed). 3

38. Li Z, Yang B, Liu Q *et al.* Monkey: Image resolution and text label are important things for large multi-modal mod-els. *IEEE/CVF Conference on Computer Vision and Pat-tern Recognition* (2024) 26763–26773. 3

39. Liu H, Li C, Li Y *et al.* Improved baselines with visual in-struction tuning. *IEEE/CVF Conference on Computer Vi-sion and Pattern Recognition* (2024) 26296–26306. 3, 4

40. Lin Z, Liu C, Zhang R *et al.* Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv:2311.07575* ; . 3

41. McKinzie B, Gan Z, Fauconnier JP *et al.* Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv:2403.09611* ; . 3, 4, 5

42. Elizalde B, Deshmukh S, Al Ismail M *et al.* Clap learning audio concepts from natural language supervision. *IEEE International Conference on Acoustics, Speech and Sig-nal Processing* (2023) 1–5. 3

43. Girdhar R, El-Nouby A, Liu Z *et al.* Imagebind: One em-bedding space to bind them all. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 15180– 15190. 3

44. Chung HW, Hou L, Longpre S *et al.* Scaling instruction-finetuned language models. *J Mach Learn Res* 2024; 25: 1–53. 3, 4

45. Touvron H, Lavril T, Izacard G *et al.* Llama: Open and efficient foundation language models. *arXiv: 2302.13971* ; . 3, 4

46. Chiang WL, Li Z, Lin Z *et al.* Vicuna: An open-source chat-bot impressing gpt-4 with 90% chatgpt quality. https: //vicuna.lmsys.org (17 October 2024, date last ac-cessed). 3, 4

47. Touvron H, Martin L, Stone K *et al.* Llama 2: Open foun-dation and fine-tuned chat models. *arXiv:2307. 09288* ; . 4

48. Bai J, Bai S, Chu Y *et al.* Qwen technical report. *arXiv:2309.16609* ; . 3, 4, 12

49. meta. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3 (17 October 2024, date last accessed). 4

50. Li J, Li D, Savarese S *et al.* Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, volume 202 2023) 19730–19742. 3, 4

51. Dai W, Li J, Li D *et al.* Instructblip: Towards general-purpose vision-language models with instruction tuning. *Conference on Neural Information Processing Systems* (2023) . 3, 4, 7, 8, 10

52. Liu H, Li C, Li Y *et al.* Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next (17 October 2024, date last accessed). 4, 12

53. Lu Y, Li C, Liu H *et al.* An empirical study of scaling instruct-tuned large multimodal models. *arXiv:2309. 09958* ; . 4

54. Chu X, Qiao L, Lin X *et al.* Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv:2312.16886* ; . 4, 12

55. Shen S, Hou L, Zhou Y *et al.* Mixture-of-experts meets instruction tuning: A winning combination for large lan-guage models. *arXiv:2305.14705* ; . 4

56. Lin B, Tang Z, Ye Y *et al.* Moe-llava: Mixture of experts for large vision-language models. *arXiv:2401. 15947* ; . 4

57. Carion N, Massa F, Synnaeve G *et al.* End-to-end object detection with transformers. *European Conference on Computer Vision* (2020) 213–229. 4

58. Hu W, Xu Y, Li Y *et al.* Bliva: A simple multimodal llm for better handling of text-rich visual questions. *AAAI Con-ference on Artificial Intelligence*, volume 38 (2024) 2256– 2264. 4

59. Alayrac JB, Donahue J, Luc P *et al.* Flamingo: a visual language model for few-shot learning. *Conference on Neural Information Processing Systems*, volume 35 (2022) 23716–23736. 4

60. Wang W, Lv Q, Yu W *et al.* Cogvlm: Visual expert for pre-trained language models. *arXiv:2311 03079* ; . 4

61. Zhang R, Han J, Zhou A *et al.* Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *The Twelfth International Conference on Learning Rep-resentations* (2024) . 5

62. Zeng Y, Zhang H, Zheng J *et al.* What matters in training a gpt4-style language model with multimodal inputs? *Con-ference of the North American Chapter of the Associa-tion for Computational Linguistics*, volume 1 (2024) 7930–7957. 5, 8, 9

63. Yin S, Fu C, Zhao S *et al.* Woodpecker: Hallucination correction for multimodal large language models. *arXiv:2310.16045* ; . 5, 11, 13

64. Chen L, Li J, Dong X *et al.* Sharegpt4v: Improv-ing large multi-modal models with better captions. *arXiv: 2311.12793* ; . 5, 6

65. Sharma P, Ding N, Goodman S *et al.* Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Annual Meeting of the As-sociation for Computational Linguistics*, volume 1 (2018) 2556–2565. 6

66. Changpinyo S, Sharma P, Ding N *et al.* Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *IEEE/CVF Conference on Com-puter Vision and Pattern Recognition* (2021) 3558–3568. 6

67. Ordonez V, Kulkarni G and Berg T. Im2text: Describing im-ages using 1 million captioned photographs. *Conference on Neural Information Processing Systems* 2011; 24. 6

68. Schuhmann C, Beaumont R, Vencu R *et al.* Laion-5b: An open large-scale dataset for training next generation image-text models. *Conference on Neural Information Processing Systems*, volume 35 (2022) 25278–25294. 6

69. Schuhmann C, Köpf A, Vencu R *et al.* Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco (17 October 2024, date last accessed). 6

70. Li J, Li D, Xiong C *et al.* Blip: Bootstrapping language-image pre-training for unified vision-language understand-ing and generation. *International Conference on Machine Learning* (2022) 12888–12900. 6

71. Byeon M, Park B, Kim H *et al.* Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (17 October 2024, date last accessed). 6

72. Wang J, Meng L, Weng Z *et al.* To see is to be-lieve: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574* ; . 6, 8, 9

73. Chen GH, Chen S, Zhang R *et al.* Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684* ; . 6, 8, 9

74. Xu J, Mei T, Yao T *et al.* Msr-vtt: A large video descrip-tion dataset for bridging video and language. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016) 5288–5296. 6

75. Mei X, Meng C, Liu H *et al.* Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans Audio Speech Lang Process* 2024; 32: 3339–3354. 6

76. Wei J, Bosma M, Zhao VY *et al.* Finetuned language mod-els are zero-shot learners. *International Conference on Learning Representations* (2022) . 7

77. OpenAI. Introducing chatgpt. https:// www.openai.com/research/chatgpt (17 Octo-ber 2024, date last accessed). 7

78. Ouyang L, Wu J, Jiang X *et al.* Training language models to follow instructions with human feedback. *Conference on Neural Information Processing Systems*, volume 35 (2022) 27730–27744. 7, 9

79. Sanh V, Webson A, Raffel C *et al.* Multitask prompted training enables zero-shot task generalization. *Interna-tional Conference on Learning Representations* (2022) . 7

80. Zhang Y and Yang Q. An overview of multi-task learning. *Natl Sci Rev* 2018; 5: 30–43. 7

81. Gong T, Lyu C, Zhang S *et al.* Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790* ; . 7, 14

82. Antol S, Agrawal A, Lu J *et al.* Vqa: Visual question answering. *IEEE/CVF International Conference on Com-puter Vision* (2015) 2425–2433. 7

83. Karpathy A and Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. *IEEE/CVF Con-ference on Computer Vision and Pattern Recognition* (2015) 3128–3137. 7

84. Xu Z, Shen Y and Huang L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 11445–11465. 7, 8, 10

85. Zhao Z, Guo L, Yue T *et al.* Chatbridge: Bridging modal-ities with large language model as a language catalyst. *arXiv:2305.16103* ; . 8

86. Li L, Yin Y, Li S *et al.* M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv:2306.04387* ; . 8

87. Maaz M, Rasheed H, Khan S *et al.* Video-chatgpt: Towards detailed video understanding via large vision and language models. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 12585– 12602. 9, 11

88. Drossos K, Lipping S and Virtanen T. Clotho: An audio captioning dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing* (2020) 736– 740. 9

89. Wang Y, Kordi Y, Mishra S *et al.* Self-instruct: Aligning language model with self generated instructions. *Annual Meeting of the Association for Computational Linguistics*, volume 1 (2023) 13484–13508. 8

90. Yang R, Song L, Li Y *et al.* Gpt4tools: Teaching large lan-guage model to use tools via self-instruction. *Conference on Neural Information Processing Systems*, volume 36 (2023) . 8, 11, 15

91. Luo G, Zhou Y, Ren T *et al.* Cheap and quick: Efficient vision-language instruction tuning for large language models. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 8

92. Wei L, Jiang Z, Huang W *et al.* Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv: 2308.12067* ; . 9

93. Du Y, Guo H, Zhou K *et al.* What makes for good visual instructions? synthesizing complex visual reasoning in-structions for visual instruction tuning. *arXiv:2311.01487* ; .9

94. Ziegler DM, Stiennon N, Wu J *et al.* Fine-tuning language models from human preferences. *arXiv:1909. 08593* ; . 9

95. Stiennon N, Ouyang L, Wu J *et al.* Learning to summarize with human feedback. *Conference on Neural Information Processing Systems* 2020; 33: 3008–3021. 9

96. Sun Z, Shen S, Cao S *et al.* Aligning large multimodal models with factually augmented rlhf. *Findings of the As-sociation for Computational Linguistics* (2023) . 10, 13

97. Rafailov R, Sharma A, Mitchell E *et al.* Direct prefer-ence optimization: Your language model is secretly a re-ward model. *Conference on Neural Information Process-ing Systems* 2024; 36. 10

98. Yu T, Yao Y, Zhang H *et al.* Rlhf-v: Towards trust-worthy mllms via behavior alignment from fine-grained correctional human feedback. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 13807– 13816. 10

99. Li L, Xie Z, Li M *et al.* Silkie: Preference distillation for large visual language models. *arXiv:2312.10665* ; . 10

100. Lu P, Mishra S, Xia T *et al.* Learn to explain: Multimodal reasoning via thought chains for science question answer-ing. *Conference on Neural Information Processing Sys-tems*, volume 35 (2022) 2507–2521. 10, 14

101. Vedantam R, Lawrence Zitnick C and Parikh D. Cider: Consensus-based image description evaluation. *IEEE/ CVF Conference on Computer Vision and Pattern Recognition* (2015) 4566–4575. 10

102. Agrawal H, Desai K, Wang Y *et al.* Nocaps: Novel object captioning at scale. *IEEE/CVF International Conference on Computer Vision* (2019) 8948–8957. 10

103. He X, Zhang Y, Mou L *et al.* Pathvqa: 30000+ questions for medical visual question answering. *arXiv: 2003.10286* ; . 10

104. Fu C, Chen P, Shen Y *et al.* Mme: A comprehensive eval-uation benchmark for multimodal large language models. *arXiv:2306.13394* ; . 10, 13

105. Liu Y, Duan H, Zhang Y *et al.* Mmbench: Is your multi-modal model an all-around player? *European Conference on Computer Vision*, volume 15064 2024) 216–233. 10, 11

106. Ning M, Zhu B, Xie Y *et al.* Video-bench: A compre-hensive benchmark and toolkit for evaluating video-based large language models. *arXiv:2311.16103* ; . 11

107. Ye Q, Xu H, Xu G *et al.* mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178* ; . 11

108. Lin TY, Maire M, Belongie S *et al.* Microsoft coco: Com-mon objects in context. *European Conference on Com-puter Vision* 740–755. 11

109. Gao P, Han J, Zhang R *et al.* Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv: 2304.15010* ; . 11

110. Yang Z, Li L, Lin K *et al.* The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv: 2309.17421* ; . 11

111. Wen L, Yang X, Fu D *et al.* On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving ; . 11

112. Fu C, Zhang R, Lin H *et al.* A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv: 2312.12436* ; . 11

113. You H, Zhang H, Gan Z *et al.* Ferret: Refer and ground anything anywhere at any granularity. *International Con-ference on Learning Representations* (2024) . 11

114. Lai X, Tian Z, Chen Y *et al.* Lisa: Reasoning segmenta-tion via large language model. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 9579– 9589. 12, 15

115. Xu R, Wang X, Wang T *et al.* Pointllm: Empowering large language models to understand point clouds. *European Conference on Computer Vision* (2024) . 12

116. Sun Q, Yu Q, Cui Y *et al.* Generative pretraining in mul-timodality. *International Conference on Learning Repre-sentations* (2024) . 12

117. Zhang D, Li S, Zhang X *et al.* Speechgpt: Empowering large language models with intrinsic cross-modal conver-sational abilities. *arXiv:2305.11000* ; . 12

118. Wang X, Zhuang B and Wu Q. Modaverse: Efficiently transforming modalities with llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 26606–26616. 12

119. Wu S, Fei H, Qu L *et al.* Next-gpt: Any-to-any multimodal llm. *International Conference on Machine Learning* (2024) . 12

120. Ho J, Jain A and Abbeel P. Denoising diffusion probabilis-tic models. *Conference on Neural Information Processing Systems* 2020; 33: 6840–6851. 12

121. Hu J, Yao Y, Wang C *et al.* Large multilingual models pivot zero-shot multimodal learning across languages. *Interna-tional Conference on Learning Representation s* (2024) . 12

122. Yang Z, Liu J, Han Y *et al.* Appagent: Multimodal agents as smartphone users. *arXiv:2312.13771* ; . 12

123. Wang J, Xu H, Ye J *et al.* Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv:2401.16158* ; . 12

124. Ye J, Hu A, Xu H *et al.* mplug-docowl: Modularized multi-modal large language model for document understanding. *arXiv:2307.02499* ; . 12

125. Yu L, Poirson P, Yang S *et al.* Modeling context in referring expressions. *European Conference on Computer Vision*, volume 9906 (2016) 69–85. 12

126. Mao J, Huang J, Toshev A *et al.* Generation and compre-hension of unambiguous object descriptions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016) 11–20. 12

127. Zeng Y, Zhang X and Li H. Multi-grained vision language pre-training: Aligning texts with visual concepts. *Inter-national Conference on Machine Learning*, volume 162 (2022) 25994–26009. 12

128. OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (17 October 2024, date last accessed). 12

129. Yao Y, Yu T, Zhang A *et al.* Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408. 01800* ; . 12

130. He M, Liu Y, Wu B *et al.* Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530* ; . 12

131. Zhai B, Yang S, Zhao X *et al.* Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv:2310.01779* ; . 13

132. Rohrbach A, Hendricks LA, Burns K *et al.* Object hal-lucination in image captioning. *Conference on Empirical Methods in Natural Language Processing* (2018) 4035– 4045. 13

133. Li Y, Du Y, Zhou K *et al.* Evaluating object hallucination in large vision-language models. *2023 Conference on Em-pirical Methods in Natural Language Processing* (2023) 292–305. 13

134. Wang J, Zhou Y, Xu G *et al.* Evaluation and anal-ysis of hallucination in large vision-language models. *arXiv:2308.15126* ; . 13

135. Liu F, Lin K, Li L *et al.* Mitigating hallucination in large multi-modal models via robust instruction tuning. *Interna-tional Conference on Learning Representations* (2024) . 13

136. Dong Q, Li L, Dai D *et al.* A survey for in-context learning. *arXiv:2301.00234* ; . 13

137. Li B, Zhang Y, Chen L *et al.* Mimic-it: Multi-modal in-context instruction tuning. *arXiv:2306.05425* ; . 14

138. Tai Y, Fan W, Zhang Z *et al.* Link-context learning for mul-timodal llms. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024) 27176–27185. 14

139. Yang X, Wu Y, Yang M *et al.* Exploring diverse in-context configurations for image captioning. *Conference on Neu-ral Information Processing Systems*, volume 36 (2023) . 14

140. Yang X, Peng Y, Ma H *et al.* Lever lm: Configuring in-context sequence to lever large vision language models. *arXiv:2312.10104* ; . 14

141. Yang Z, Gan Z, Wang J *et al.* An empirical study of gpt-3 for few-shot knowledge-based vqa. *AAAI Conference on Artificial Intelligence*, volume 36 (2022) 3081–3089. 14

142. Lu P, Peng B, Cheng H *et al.* Chameleon: Plug-and-play compositional reasoning with large language models. *Conference on Neural Information Processing Systems*, volume 36 (2023) 43447 43478. 14, 15

143. Gupta T and Kembhavi A. Visual programming: Composi-tional visual reasoning without training. *IEEE/CVF Confer-ence on Computer Vision and Pattern Recognition* (2023) 14953–14962. 14, 15

144. Rose D, Himakunthala V, Ouyang A *et al.* Visual chain of thought: Bridging logical gaps with multimodal infillings. *arXiv:2305.02317* ; . 14

145. Zhang Z, Zhang A, Li M *et al.* Multimodal chain-of-thought reasoning in language models. *arXiv:2302.00923* 2023; . 14

146. Zheng G, Yang B, Tang J *et al.* Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Conference on Neural Information Pro-cessing Systems*, volume 36 (2023) 5168–5191. 14

147. Ge J, Luo H, Qian S *et al.* Chain of thought prompt tuning in vision language models. *arXiv:2304.07919* ; . 14

148. Parisi A, Zhao Y and Fiedel N. Talm: Tool augmented lan-guage models. *arXiv:2205.12255* ; . 14

149. Zhu X, Zhang R, He B *et al.* Pointclip v2: Prompt-ing clip and gpt for powerful 3d open-world learning. *IEEE/CVF International Conference on Computer Vision* (2023) 2639–2650. 15

150. Wang T, Zhang J, Fei J *et al.* Caption anything: Interac-tive image description with diverse multimodal controls. *arXiv:2305.02677* ; . 15

151. Shen Y, Song K, Tan X *et al.* Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Conference on Neural Information Processing Systems*, volume 36 (2024) . 15

152. You H, Sun R, Wang Z *et al.* Idealgpt: Iteratively decom-posing vision and language reasoning via large language models. *arXiv:2305.14985* ; . 15

153. Zeng A, Wong A, Welker S *et al.* Socratic models: Com-posing zero-shot multimodal reasoning with language. *International Conference on Learning Representations* (2023) . 15

154. Zhang R, Hu X, Li B *et al.* Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) 15211–15222. 15