·专题:ChatGPT 与人工智能技术应用·

GPT-4 对多模态大模型在多模态理解、生成、交互上的启发

刘 静[1,2*] 郭龙腾[1,2]

1. 中国科学院 自动化研究所,北京 100190
2. 中国科学院大学 人工智能学院,北京 100190

[摘 要] 对话式聊天机器人 ChatGPT 以近乎摧枯拉朽的气势席卷社会,拨开了通用人工智能的曙光。ChatGPT 的升级版 GPT-4 是个多模态大模型,它从单调的文本交互,升级为可以接受文本与图像组合的多模态输入,相比传统的单模态大模型,多模态大模型更加符合人类的多渠道感认知方式,能够应对更加复杂丰富的环境、场景和任务。GPT-4 表明在多模态大模型中引入基于人类知识的自然语言理解与生成能力能够带来模型在多模态理解、生成、交互能力上的巨大提升。本文将介绍多模态大模型的概念、关键技术、近期进展和应用场景、GPT-4 的技术特性,并重点探讨以 GPT-4 为代表的大语言模型对构建多模态大模型的几点启发。具体而言,将讨论如何充分利用大语言模型的语言能力,在多模态大模型的构建中,借助语言的帮助更好地感知理解世界、创作生成内容、与人和环境交互。

[关键词] GPT-4;多模态大模型;多模态理解;多模态生成;多模态交互

1 多模态大模型技术概述

过去十多年内,深度学习技术大致经过了三次重大的研究范式转变,经历了从“监督学习+各自为政”到“预训练模型+任务微调”,再到如今的“预训练大模型+提示生成”的发展历程。传统人工智能模型往往依赖大量有标签数据的监督训练,而且一个模型一般只能解决一个任务,仅适用于单一场景,这使得人工智能的研发和应用成本高,场景适应能力弱,难以规模化应用。而大模型通常预先在海量数据上进行大规模预训练,然后通过微调、上下文学习、零样本学习等方式以适应一系列下游任务。

预训练大模型技术率先在文本领域取得了突破,产生了像 ChatGPT 这样的代表性成果,能够将不同类型的文本任务统一放到一个模型下解决。不同于文本任务有着相对统一的任务形态,在多模态任务中,涉及到多种模态信息(包括文本、语音、视觉、红外、3D 点云等等),这些不同模态的各种输入、输出组合产生了丰富的任务形态。常见的多模态任务大致可以分为两类:(1) 多模态理解任务,如视频分类、视觉问答、跨模态检索、指代表达等;(2) 多模态生成任务[1],如以文生图和视频、歌词生成音乐、基于对话的图片编辑等。

基于多模态、多领域信息构建的预训练多模态大模型[2,3],能够从海量数据中学习不同模态之间的联合表征与转换生成关系,由此训练的模型具有良

刘静 中国科学院自动化研究所研究员,博士生导师。主要研究方向为多模态分析理解、多模态预训练大模型等。曾获中国电子学会自然科学奖一等奖、中国图象图形学学会科学技术奖二等奖、世界人工智能大会“卓越人工智能引领者奖 SAIL”等奖项。在相关领域的国际学术竞赛中荣获冠军 10 余次。已发表高水平学术论文 150 余篇。主持国家自然科学基金优秀青年科学基金项目、面上项目等。

收稿日期:2023-06-23;修回日期:2023-09-28

* 通信作者,Email: jliu@nlpr.ia.ac.cn

本文受到科技创新 2030“新一代人工智能”重大项目(2022ZD0118801)和国家自然科学基金项目(U21B2043)的资助。

好的任务泛化性，在多模态理解和多模态生成等相关任务中取得了明显优势。相比于视觉、语言等单模态大模型主要建模特定模态的数据，多模态大模型则在预训练过程中整合多个模态的信息并构建跨模态关联，这使得其能够同时处理和理解来自不同感知通道的信息，并以多种模态表达自己的输出，实现更全面的感知、更丰富的生成，从而支持更灵活的交互能力。

多模态大模型的关键技术大致包括以下四部分（图 1）：大规模预训练数据、模型架构设计、自监督学习任务设计和下游任务适配。其中，与小模型使用小型人工标注数据不同，大模型的预训练数据通常是从互联网上收集、清洗的千万/亿级别的弱关联图文、视频—文本、视频—音频等多模态数据，样本数量多但噪声大是其显著特点。多模态预训练模型架构需要足够通用，以轻松兼容各类下游任务。其核心构成通常包括一个多模态层级编码网络，即先以单模态编码器实现对各个模态的单独编码，再以跨模态编码器实现多模态之间的交互理解，以及一个跨模态解码器用以根据所编码的多模态信息解码生成某一模态的内容。自监督学习任务的设计需要使得大模型能充分利用海量弱关联数据来挖掘模态之间的关联，常见的自监督学习任务包括模态对比学习、条件语言建模等。其中，模态对比学习将不同模态映射到统一语义空间中进行度量学习，拉近语义匹配的多模态数据的表征，有助于泛化到下游跨模态检索和开放词表分类等任务；条件语言建模则根据多模态上下文来预测后续文本序列，这有助于实现其他模态信息与语言的对齐和融合，同时带来跨模态生成能力。下游任务适配通常有零样本学习，上下文学习和模型微调等方法。其中零样本学习旨在利用大模型在预训练阶段中习得的强大泛化能力来解决相关任务，上下文学习进一步给出了相关任务的实例作为提示来进行推理，而微调则在特定任务的数据上进一步训练模型，使得模型参数偏向于拟合特定任务。

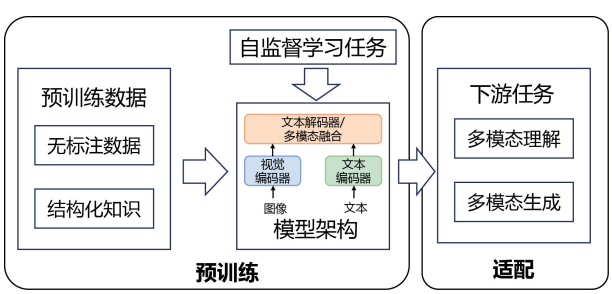

图 1 多模态大模型的整体技术框架

近年来，多模态大模型在多模态理解和多模态生成上都取得了显著的进展：

（1）在多模态理解上，主要有以下几个关键进展：一是模型体量和训练数据量持续增大，如 Flamingo[4] 模型达到了 80 B 的参数量，GIT[5] 模型使用了 10 B 级的预训练数据；二是模态更丰富，如 VALOR[6] 联合建模了图像、视频、文本与音频四个模态，ImageBind[7] 通过以视觉为中心进行对比学习来连接文本、音频、视觉、红外、惯性测量单元（IMU）信号等六种模态；三是所支持的任务更多样，如 OFA[8] 联合建模了多种理解与生成任务（视觉问答、图像生成、图像描述、文本任务、目标检测等）；四是强化面向开放世界的细粒度感知，如 GLIPv2[9] 可以实现开放词表的目标检测；五是对接大语言模型并增加指令学习，如 LLaVA[10] 构造指令微调数据集来训练大模型遵循人类指令的能力。

（2）在多模态生成上，主要进展总结为以下四个方面：首先，近期涌现了许多大型预训练生成模型，如 Stable Diffusion[11] 和 Google 的 Imagen[12] 等，它们能够基于简单的语言描述生成几乎以假乱真的图像，实现了前所未有的生成质量和泛化能力。其次，跨模态表示学习逐渐成为多模态生成领域的研究重点之一，其中，CoDi[13] 采用桥接对齐的方式同时学习文本、图像、视频和音频等多种模态的表示，为模型提供了更强大的多模态理解能力。第三，同时融合多种生成任务能力于一体的模型开始涌现，例如，UniDiffuser[14] 基于扩散模型同时建模文本生成图像、图像生成文本和图像文本对生成等多项任务，NUWA[15] 基于序列生成模型完成文本生成图像、文本生成视频以及图像生成视频等多种生成任务的建模。最后，多模态生成大模型的应用领域也不断扩展到图像视频编辑、分子图生成、三维图像生成以及数据增强等领域。这些进展为多模态生成技术在解决各种现实世界问题中的应用提供了新的机遇和潜力。

多模态大模型在多模态理解与生成上的进展进一步支撑起了多模态交互技术的广泛应用，能够与人类或外部环境等对象进行基于多模态输入、输出的多轮互动交互，包括交互式多模态问答对话、交互式内容编辑、多模态环境下的交互式决策等。得益于在这些多模态理解、生成和交互任务上展现的强大能力和突破性进展，多模态大模型能够支撑非常广泛的应用场景，以下示例展示了几个典型的应用场景和多模态大模型能够为这些场景带来的优势

特性：

（1）情感分析：多模态大模型可以结合语音、面部表情、肢体语言等多种模态数据来分析人的情感状态，从而更准确地理解和响应用户的情感需求，支持更具“情商”的智能机器人、智能座舱等应用。

（2）问答对话：多模态大模型可以接受图像、视频、音频等输入，并结合文字描述进行多模态问答对话，回答用户提出的与多模态信息相关的问题，这在盲人辅助、智能助理、聊天客服、智慧教育等领域具有重要应用价值。

（3）图像、音频、视频的生成与编辑：多模态大模型可以根据文字、音频或视觉信息来生成图像、音频、视频甚至有声视频等模态内容。此外，通过将人类的多模态编辑指令作为模型的控制条件，还能实现可控的内容编辑功能。

（4）自动驾驶：多模态大模型可以融合视觉、雷达、红外等多种传感器数据，实现对道路、车辆和行人的感知和理解，从而增强自动驾驶能力。

（5）医学诊断与监测：多模态大模型可以结合医学图像、声音等多种模态数据，帮助医生进行影像判读、“望闻问切”疾病综合诊断、监测患者病情，并提供个性化的医疗建议。

（6）增强现实（Augmented Reality，AR）与虚拟现实（Virtual Reality，VR）：多模态大模型可以结合视觉和声音信息，通过与数字人等技术结合，为增强现实和虚拟现实应用提供更丰富的体验和交互方式。

（7）智能辅助设备：多模态大模型可以结合语音、图像等模态数据，为智能助理、智能家居等设备提供更自然、智能的人机交互方式，提升用户体验。

2　GPT-4 的技术特性

从 GPT-1 开始，OpenAI 通过逐步扩大模型和数据规模，不断优化其 GPT 系列的语言模型，推动了人工智能技术的演进和升级（图 2）。GPT-1 首先仅利用 Transformer 的解码器并通过大量无标注数据进行自监督预训练，然后通过有监督微调来解决不同的下游任务，大大减少了任务之间的迁移困难。GPT-2 进一步扩大了模型和参数规模，尽管表现出一定的通用性，但仍受限于其性能边界。随后，GPT-3 的问世突破了之前版本的限制，通过使用 1 750 亿参数规模的模型和 45 TB 的数据量，即使无需任何微调，也能够仅通过提示或少数样例完成多种任务，展示出惊人的通用性。OpenAI 进一步提出了 InstructGPT[16] 模型，通过结合有监督的指令微调和人类反馈的强化学习方式，实现了模型的自我优化和更新，更好地遵循用户的意图和需求。而 ChatGPT 则是基于 GPT-3.5，结合了 InstructGPT 的训练方式，并加入更丰富的数据类型（如代码和思维链数据）进行训练，从而拥有更强的逻辑推理和多轮对话能力。

2023 年，GPT-4 的发布标志着多模态大模型技术的一个新里程碑。下面将深入探讨 GPT-4 的五大突出特性，以及它如何改变我们理解和应用 AI 技术的方式。

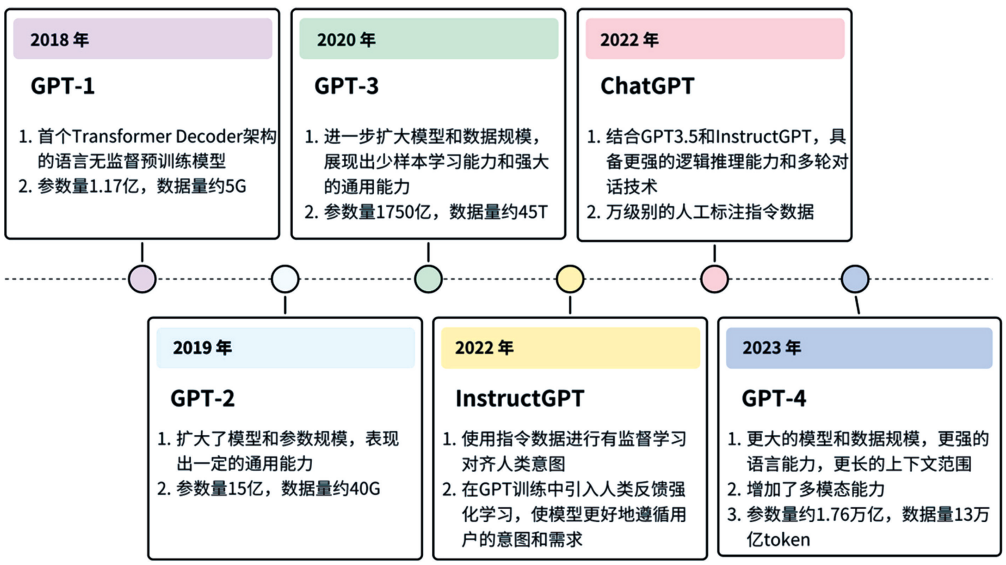

图 2　GPT 系列模型的发展脉络

(1) GPT-4 具备处理多模态输入的能力。它不仅仅可以理解和生成文字,还具备强大的图像理解能力。这种全新的维度不仅仅限于单一的文字或图像模态,而是能够理解各类图文混合的内容。令人惊讶的是,在 4 个场景中,GPT-4 甚至能够在零样本状态下超越微调后的专家模型(表 1)。这在很多应用场景中都非常有价值,比如视觉问答、图像描述生成、解释图表等。

(2) GPT-4 引入了更长的上下文窗口支持。能处理更长的上下文对于复杂问题或长篇文章的分析尤为重要,这意味着 GPT-4 能够更好地持续理解和回答与较长文本相关的问题。有两个版本的 GPT-4 分别支持 8 K 和 32 K 的上下文长度,这分别是 ChatGPT 支持上下文长度的 2 倍和 8 倍。尽管这带来了更高的计算成本(分别为 ChatGPT 的 3 倍和 7 倍),但它为更深层次、更全面的对话和文本分析提供了可能。

(3) GPT-4 在处理复杂任务方面有了显著进步。它在更多复杂和细微的任务处理中表现得更加可靠和有创意。这一点在多项考试和测验中得到了验证。具体来说,GPT-4 在多种不同年龄段和类别的考试中均名列前茅,如在律师职业资格考试和生物学奥赛中分别位列人类考生成绩中的前 10% 和前 1%。同时,在 MMLU 基准测试中,它也明显优于其他大模型。

(4) 在改善幻觉和安全性方面,GPT-4 也做出了重大努力。GPT-4 在各类任务上显著减轻了幻觉问题,安全能力比 GPT-3.5 模型提高了 40%,这确保了 GPT-4 能够提供更加准确和可靠的输出。

(5) GPT-4 还为我们带来了预测模型扩展性的创新。GPT-4 通过在 1/1 000 的计算量上预测模型扩展性,极大地提升了大模型训练效率,降低了大模型训练成本。这在大模型不适合广泛调参的情况下尤为重要,因为它可以使用较小的模型预先预测训练行为和损失,从而增强了大模型训练的可控性。

表 1 GPT-4 的零样本多模态能力

测试基准	GPT-4	专用模型最优性能
VQAv2[17]	77.2%	84.3% (PaLI-17B[18])
TextVQA[19]	78.0%	71.8% (PaLI-17B[18])
AI2 Diagram (AI2D)[20]	78.2%	42.1% (Pix2Struct Large[21])
DocVQA[22]	88.4%	88.4% (ERNIE-Layout 2.0[23])
Infographic VQA[24]	75.1%	61.2% (Applica.ai TILT[25])
TVQA[26]	87.3%	86.5% (MERLOT Reserve Large[27])
LSMDC[28]	45.7%	52.9% (MERLOT[29])

综上所述,GPT-4 是一个真正的里程碑,它在多方面都实现了显著的改进和进步。无论是其多模态理解能力、更长的上下文窗口,还是其在处理复杂任务方面的强大能力,GPT-4 都为我们展示了 AI 技术的巨大潜力和未来方向。通过不断改善其局限性和引入创新的特性,GPT-4 无疑将为 AI 领域开辟新的可能性和应用前景。

3 GPT-4 对多模态大模型的启发

GPT-4 为多模态大模型的发展启发了新的方向:一方面,GPT-4 使机器对语言的理解和表达能力达到接近人类的水平,借助这类语言能力,多模态大模型有望显著提升自身在语言理解、语言生成、逻辑推理、隐式知识等能力上的表现,而这些能力上的跃升将使得之前无法实现的多模态理解、生成和交互应用成为可能;另一方面,GPT-4 展现了强大的多模态理解和交互能力,这指明将多模态大模型与语言能力相结合有望实现更加高效的理解和交互方式。

3.1 以语言和多模态结合的方式感知理解世界

GPT-4 的强大图像理解能力启发了多模态大模型通过结合语言和多模态信息来获取关于周围世界的更全面立体的感知理解,并因此在近期掀起了构建类似 GPT-4 的“多模态大语言模型”的研究热潮。

多模态的感知理解涉及到多个模态之间的对齐,而采取类似 GPT-4 路线,通过大语言模型为主导界面来实现多模态的对齐、融合、交互成为了目前被广泛认可的一种多模态大模型结构范式。这是由于文本有高效的表达效率、能够通过语义描述的方式与其他模态建立直接的联系。此外,大语言模型在预训练过程中学习到了非常多的世界知识,有潜在理解多模态信息的能力。受启发于 GPT-4 在结合语言进行图像理解与推理能力上的巨大成功,训练结合多模态大模型与大语言模型的“多模态大语言模型”近期收到众多研究者的关注。多模态大语言模型在结构方面常由单模态编码器、连接器与大语言模型三部分组成,其中单模态编码器和大语言模型的参数可以冻结或部分冻结以减少计算量、提高训练效率;连接器常见的有简单的线性映射层,或者特殊设计的网络模块如 BLIP-2[30] 中的 Q-former 结构等。多模态大语言模型通常涉及到两个阶段的

训练过程。在第一阶段，训练各个模态到大语言模型的语义对齐，通常利用大规模弱关联的跨模态数据（如图像—文本、视频—文本或音频—文本数据），基于条件语言建模任务进行训练。在第二阶段进行类似 InstructGPT 的指令微调以提升零样本多模态能力，此阶段的核心是构造面向多模态任务的指令微调数据，目前常见的多模态指令微调数据类型有多模态对话、多模态详细描述与多模态推理问答等。

类 GPT-4 的多模态大语言模型还存在着诸多问题与挑战有待解决。其一，模型的幻觉问题，由于指令微调阶段中的任务鼓励模型生成详细描述，这可能导致模型生成不存在于源数据本身的元素，进而影响模型实际应用的效果，在多模态大语言模型中这一问题甚至比在大语言模型中还要严重。其二，模型内部知识与外部知识库的协同作用机制尚未成熟，尽管大模型本身学习到了丰富的世界知识，但针对某些困难的问题或者既定领域的问题，引入外部知识库是非常重要的。因此需要构建庞大的多模态外部知识库，并且设计机制来实现多模态外部知识的引入判断与融合输出。其三，更多模态的细粒度对齐，尽管目前的多模态大语言模型对于图像具有了一定的细粒度理解能力，但是对于其他模态如视频、音频等，仍然停留在全局粗粒度的对齐与理解上，而实现更多模态的细粒度理解有助于实现更充分的模态交互与协同推理。

3.2　以语言和多模态结合的方式创作生成内容

除了理解世界知识的能力之外，多模态大模型所需具备的另一大能力是生成多模态内容，即基于给定的多模态输入信息，输出多模态生成结果，这可以创作出丰富多彩的内容。GPT-4 的强大语言能力让多模态内容生成能够以语言为通用接口，让多模态大模型的生成能力更为精细、可控和易用，实现自然直观的内容生成和编辑体验。

在语言与多模态生成的结合领域，已经涌现出许多成功的研究，包括基于语言生成图片、视频、音频、3D、分子结构等多种模态内容的方法，以及基于语言对话实现图像、视频的编辑等应用。在众多生成模型中，目前多模态生成领域最为成功和引人瞩目的模型主要包括序列生成模型和扩散生成模型等。主流方法的发展脉络如图 3 所示。DALL-E[31] 是典型的图文多模态序列生成模型，其采用自回归生成范式，在大规模数据（2.5 亿个图文对）上进行文本到图像生成的训练，在以文生图任务上取得了突破性的生成质量和泛化能力。CogView[32] 进一步探索了多模态生成模型在下游任务上精调后的泛化能力，在基于文本控制的样式学习、服装设计和图像超分等任务上均取得出色的效果。考虑到自回归生成范式在计算成本和误差积累方面存在问题，NUWA[15] 提出了一种通用的 3D 视觉编码—解码器网络结构，统一了文生图和文生视频等多种多模态生成任务。区别于序列生成模型，扩散模型首先采用文本预训练模型作为文本编码器，将给定的文本描述映射到语义特征空间中，再基于文本特征逐步给一幅噪音图像去噪得到生成图像，取得了超越上述自回归方法的生成质量。LDM[11] 先压缩图像的像素信息，获取与图像一一对应的隐特征表达，再采用扩散模型来建模图像隐特征分布。Stable Diffusion 拓展 LDM 至开放领域的文本至图像生成，采用图文大规模预训练模型 CLIP[33] 来提取文本特征，是当前开源模型的代表方法。DALL-E2[34] 通过训练一个独立的映射模型将 CLIP 模型的文本特征映射到图像特征空间，极大提升了生成图像与输入文本的匹配程度。Imagen[12] 发现基于通用大语言模型 T5[35] 提取的文本特征生成的图像比基于 CLIP 模型的图像细节准确度更高。

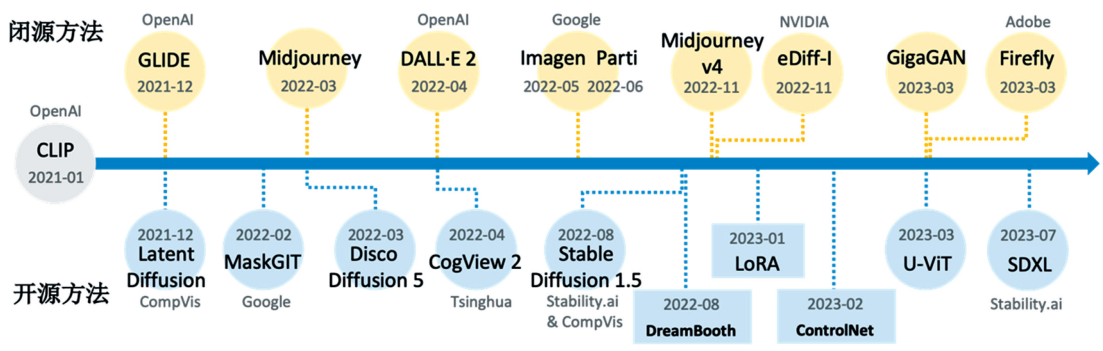

图 3　基于文本的视觉内容生成与编辑方法发展脉络

（蓝色圆圈：基于文本的视觉内容生成方法；蓝色方块：基于文本的视觉内容编辑方法）

GPT-4 对多模态生成大模型的启发可以分为提供更准确的文本表示以及多模态数据扩增两个方面。在文本表示上，尽管多模态生成领域已经取得了巨大的突破，但现有的多模态生成模型中的文本编码器存在显著的缺陷（图 4）。诸如 DALL-E 和 NUWA 等序列生成模型都直接采用了文本 tokenizer 输出的离散文本标记（Token）所对应的词嵌入向量（Word Embedding）来对输入文本进行编码。然而，这种编码方式并不包含足够的语义信息，因此其表达能力受到限制。其他主流方法的文本表示主要基于 BERT 结构的向量编码，通常采用 CLIP 或 T5 模型的文本编码器部分。然而，CLIP 文本编码器虽然关注图像的全局特征，但往往忽略了文本描述的细粒度细节。而 T5 文本编码器虽然能更好地刻画文本描述的各个目标，但整体样貌的准确性仍有改进空间[36]。这导致基于这两类文本编码器的多模态生成模型，在输出的图像上容易存在细节或全局语义上的缺陷。而 GPT-4 类大语言模型具有前所未有的语言理解能力，这使得它们有望为当前的生成模型提供更出色的文本表示，能够支持需要常识推理、隐式表达理解、复杂上下文理解等高级语言能力的生成任务，例如理解“帮我用梵高早期的画作风格画一副毕加索的代表作”这一图像生成指令。

在数据扩增上，考虑到大规模多模态数据的获取仍然是目前多模态生成领域的难题，GPT-4 类多模态大语言模型在视觉理解与语言表达上的强大能力为解决这一问题带来了可能。例如，在广泛使用的图文数据集 MSCOCO[37] 中，原始的图像文本描述通常较为简洁（平均长度仅为 11.8 个单词），对图像细节的描述不够充分；而使用 GPT-4 就能够通过理解图像生成更详尽、丰富的文本描述，从而填补了原始数据的信息缺失，使其更加有用和具体，而这还使得我们能够充分利用起海量缺乏对应文本描述的图片来进行训练。此外，GPT-4 在数据扩增上的应用潜力不仅限于图文数据，还可以扩展到视频—文本数据。当前，获取高质量的视频—文本数据相对困难，而 GPT-4 的视觉描述能力可以用来生成精确和详尽的视频描述，从而增加多模态数据资源中的视频-文本信息。这不仅有助于提高模型在视频分析和处理方面的性能，还有助于扩大多模态模型的应用领域，例如视频内容摘要、自动字幕生成等。因此，GPT-4 的视觉—语言整合能力为多模态数据的丰富化和扩展提供了新的途径。然而，GPT-4 在生成文本时的错误回答和幻觉现象（即生成的文本不遵循输入或者不符合事实）也是不容忽视的问题，这意味着使用 GPT-4 进行数据扩增时，生成的图像或视频对应描述中可能存在内容不一致、虚假、错误、偏见信息问题，会对数据集的质量造成负面影响，进而影响生成结果的质量和可靠性。因此，利用 GPT-4 等模型进行数据扩增过程中需要谨慎处理。可以采取以下策略来减轻负面影响：结合多模态信息进行过滤，将生成的文本与原图像或视频进行综合考虑，可以更好地判断描述是否准确，并避免单一模态的幻觉问题对数据增广的影响；人工数据审核，排除或修正其中的错误、虚假或有偏见的内容。

当前基于语言的多模态生成技术还面临不少挑战，例如：在内容的可控性上，如何更好地结合语言与其他精细控制方式（如鼠标、区域框等）来实现更加精细可控的多模态生成与编辑；在数据偏差与公平性上，如何克服在文本和图像等模态中存在的偏差和不平衡问题，更好地减轻模型的偏见。随

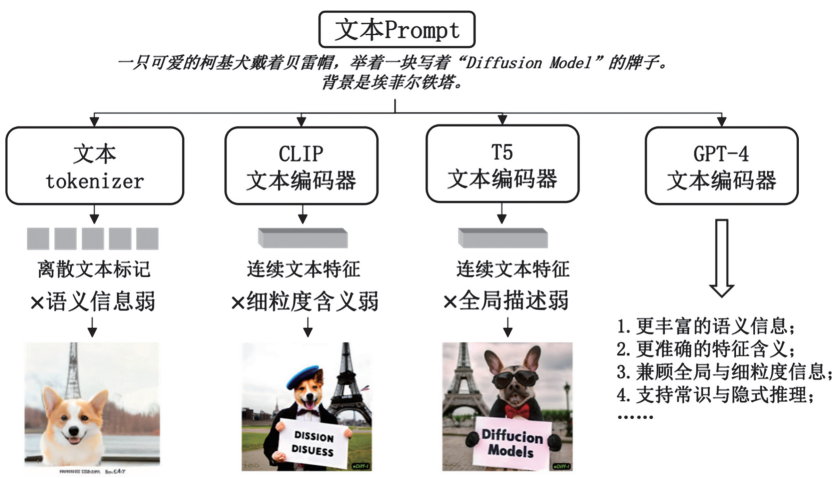

图 4 多模态生成模型中常用文本编码器的对应缺陷及使用 GPT-4 编码器的潜在优势

着基于语言的多模态生成技术的不断进步，我们已经看到并可以期待更多基于语言的多模态生成技术在生产和生活中的应用。例如：在艺术创作中，多模态生成可以实现以语言控制的跨媒介创作，加速艺术家的创作流程，丰富艺术表达的形式与内容；在游戏开发中，多模态生成可以实现以语言定义的动态剧情和角色形象，营造更加丰富的游戏体验；在商业营销中，可以根据商品卖点生成产品文案、广告视频；在电子商务中，实现根据商品介绍生成商品营销图、虚拟模特、虚拟试衣等效果。

3.3　以语言和多模态结合的方式与人和环境交互

GPT-4 所展现的多模态理解和基于多模态对话的人机交互方式，也启发了多模态大模型将多模态与语言相结合来更好地与人类用户、外部环境等对象进行交互。

自 GPT-4 问世以来，交互式自然语言处理的研究工作在多维度与不同对象进行了一系列复杂而高级的交互探索。首先，在与人类用户的互动中，这类模型能够通过个性化的交流方式，更准确地识别和满足用户的特定需求，从而有效地提升用户体验。其次，在与环境元素的交互中，模型具备根据语言指令执行具体具身任务的能力。进一步地，与外部知识库的深度集成增强了模型对背景知识的综合理解，以生成更准确和信息量更大的回应。最后，通过与各类专业工具的交互，模型能够将一项复杂任务分解为多个子任务，并借助更专业化的工具进行高效处理。这一整套交互模式不仅扩展了自然语言处理的应用范围，也在多方面增强了其功能和性能。

这种创新性的交互框架为多模态大模型提供了新的启示，展示了如何更有效地结合多模态信息和语言，实现与不同对象（包括人类、环境、外部知识、外部工具等）进行更多维度的交互。在与人类用户交互的过程中，GPT-4 等大语言模型的介入为多模态大模型开拓了新的可能性。当多模态大模型与用户进行对话时，它可以通过自然语言处理技术来解析理解用户的文本或者语音输入，同时利用视觉技术来分析用户的面部表情和姿态，以更好地理解用户的意图和情感，而在回复用户时，它可以同时生成文字回复和语音合成，还可以通过图像、视频的检索、生成等技术，以多种方式表达自己的输出。这种多模态的输入输出方式可以更好地满足用户的需求，提供更自然丰富和具有沉浸感的交互体验。例如，在在线医疗咨询中，多模态模型可以通过与病人的语言沟通过程获得更加个性化的病情描述等资料，并且除了通过自然语言沟通获取病情描述外，模型还可以分析由病人提供的声音样本、医学图像或者其他生物标志物数据，以做出更为精准和个性化的诊断。在具身智能的应用场景中，家居机器人可以在执行复杂的家居任务时利用视觉观测和自然语言与人类用户进行高效交流对话。

在与外部环境交互的过程中，GPT-4 的引入为多模态大模型提供了显著的优势。一方面，GPT-4 自身拥有庞大的内在知识库和高级的逻辑推理能力，这些能力为多模态大模型提供了更为全面和准确的认知和决策支持。例如，在导航任务中，模型可以整合自然语言指令、视觉地图、声音信号等多模态信息，通过 GPT-4 的高级逻辑推理进行路径规划和策略优化。另一方面，GPT-4 在任务执行过程中的反馈调整机制也更为先进和灵活。一旦模型在执行任务时遇到未预见的难题或者复杂情境，它可以快速地进行自我调整或者请求人类干预，以适应不断变化的外部环境。这种动态的调整和优化过程不仅提高了任务执行的成功率，也大大提升了模型的自适应能力和鲁棒性。

在与外部知识库进行交互时，多模态大模型不仅能通过文本查询与数据库对话，还能通过视觉和声音等其他模态来获取和整合信息。例如，当模型需要对一些专业性要求较高的任务进行深入解析时，它可能会通过语言模式检索文献资料，同时也能通过图像识别技术来分析与该任务相关的视觉素材。通过这种方式，模型能够从多个维度获取信息，并将这些信息整合在一起，生成一个更为全面和准确的回应。

近期，已经有一些工作探索将大语言模型和多模态感知结合，来打造能在现实世界中与人和外部环境交互的具身智能体。谷歌在 2023 年 3 月公布了 PaLM-E[38] 具身多模态语言模型，其参数量高达 5 620 亿（GPT-3 的参数量为 1 750 亿），是全球已知的最大视觉语言模型。PaLM-E 将现实世界的多种模态信息（如图像、状态估计或其他传感器信号）注入到预训练大语言模型的嵌入空间中，使得大语言模型能够以处理文本的方式处理多模态信息，从而建立语言和感知之间的联系。由此训练的 PaLM-E 模型能够遵从语言指令，面向真实环境进行多模态感知、推理和任务规划，进而用生成的行动计划调度机器人完成特定任务。谷歌 DeepMind 进一步推出

了“视觉—语言—行动”模型 RT-2[39]，它能够基于用户的语言指令和机器人的环境感知，在大语言模型的帮助下直接输出机器人的可执行行动指令，实现机器人的端到端闭环控制。

GPT-4 的出现为多模态大模型在交互层面带来了新的启示和机会，也带来了新的挑战性问题。首先，GPT-4 展示了模型能在某种程度上“理解”人类语言和需求，但另一方面，当前的多模态模型仍然大多是响应式而非主动式的。这意味着这些模型通常是被动地等待用户的输入然后做出响应，或者在满足某些预设条件的情况下才会询问用户意图。这些询问通常是预先编程的、类型固定的，而不是根据上下文动态生成的。因此，模型在主动性方面，比如主动提出多样的问题或者针对特定情境提供个性化建议，还有待进一步优化。此外，与大语言模型结合的多模态大模型可能在受控的虚拟环境中表现出色，但如何将这些模型的成果有效地迁移到更加复杂和不确定的现实世界环境依然是一个巨大的挑战。现实世界的不确定性、复杂性以及动态性极大地增加了模型需要处理的变量和因素。这些因素往往超出了模型在训练阶段所暴露的情境，因此要求模型具有更强的泛化能力和适应性。

4 总 结

GPT-4 这类大语言模型的出色语言能力为多模态大模型的发展提供了新的方向。借助大语言模型强大的语言理解和生成能力，通过将其与视觉、听觉、触觉等真实世界的多模态信号结合，多模态大模型能够实现以语言赋能的多模态理解、多模态生成和多模态交互。这将帮助多模态大模型在感知世界、创作内容、与外部对象交互等能力上产生飞跃，推动实现通用的多模态人工智能。然而，在语言赋能多模态大模型的研究上仍然存在一些挑战需要克服。例如，如何更好地利用克服多模态大语言模型中的幻觉问题；如何结合语言和多模态指令实现更加精细可控的多模态生成与编辑；如何在多模态环境感知、人类语言和物理行动之间建立映射关系，从而增强机器人等智能体的自主交互能力等。通过不断攻克这些挑战，并促进语言大模型和多模态大模型的发展和融合，多模态人工智能将在更智能化和多样化的现实应用场景中发挥更大的价值。

参 考 文 献

[1] 胡铭菲，左信，刘建伟. 深度生成模型综述. 自动化学报，2022，48(1)：40—74.

[2] 刘华峰，陈静静，李亮，等. 跨模态表征与生成技术. 中国图象图形学报，2023，28(6)：1608—1629.

[3] 王惠茹，李秀红，李哲，等. 多模态预训练模型综述. 计算机应用，2023，43(4)：991—1004.

[4] Alayrac JB，Donahue J，Luc P，et al. Flamingo：a visual language model for few-shot learning. Advances in Neural Information Processing Systems，2022：23716—23736.

[5] Wang JF，Yang ZY，Hu XW，et al. GIT：a generative image-to-text transformer for vision and language.（2022-05-27）/[2023-06-22]. https://arxiv. org/pdf/2205. 14100. pdf.

[6] Chen SH，He XJ，Guo LT，et al. VALOR：vision-audio-language omni-perception pretraining model and dataset.（2023-04-17）/[2023-06-22]. https://arxiv. org/pdf/2304. 08345. pdf.

[7] Girdhar R，El-Nouby A，Liu Z，et al. Imagebind one embedding space to bind them all// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2023：15180—15190.

[8] Wang P，Yang A，Men R，et al. OFA：unifying architectures，tasks，and modalities through a simple sequence-to-sequence learning framework.（2022-02-07）/[2023-06-22]. https://arxiv. org/pdf/2202. 03052. pdf.

[9] Zhang HT，Zhang PC，Hu XW，et al. Glipv2：Unifying localization and vision-language understanding. Advances in Neural Information Processing Systems，2022：36067—36080.

[10] Liu HT，Li CY，Wu QY，et al. Visual instruction tuning.（2023-04-17）/[2023-06-22]. https://arxiv. org/pdf/2304. 08485. pdf.

[11] Rombach R，Blattmann A，Lorenz D，et al. High-resolution image synthesis with latent diffusion models// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2022：10674—10685.

[12] Saharia C，Chan W，Saxena S，et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems，2022：36479—36494.

[13] Tang ZN，Yang ZY，Zhu CG，et al. Any-to-any generation via composable diffusion.（2023-05-19）/[2023-06-22]. https://arxiv. org/pdf/2305. 11846. pdf.

[14] Bao F, Nie S, Xue KW, et al. One transformer fits all distributions in multi-modal diffusion at scale. (2023-03-12)/[2023-06-22]. https://arxiv.org/pdf/2303.06555.pdf.

[15] Wu CF, Liang J, Ji L, et al. NÜWA: visual synthesis pre-training for neural visual world creation// Proceedings of the 2022 European Conference on Computer Vision. Cham: Springer, 2022: 720—736.

[16] Ouyang L, Wu J, Xu J, et al. Training language models to follow instructions with human feedback. (2022-03-04)/[2023-06-22]. https://arxiv.org/pdf/2203.02155.pdf.

[17] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6325—6334.

[18] Chen X, Wang X, Changpinyo S, et al. PaLI: a jointly-scaled multilingual language-image model. (2022-09-14)/[2023-06-22]. https://arxiv.org/pdf/2209.06794.pdf.

[19] Singh A, Natarajan V, Shah M, et al. Towards VQA models that can read// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 8309—8318.

[20] Kembhavi A, Salvato M, Kolve E, et al. A diagram is worth a dozen images// Proceedings of the 2016 European Conference on Computer Vision. Cham: Springer, 2016: 235—251.

[21] Lee K, Joshi M, Turc I, et al. Pix2Struct: screenshot parsing as pretraining for visual language understanding. (2022-10-07)/[2023-06-22]. https://arxiv.org/pdf/2210.03347.pdf.

[22] Mathew M, Karatzas D, Jawahar CV. DocVQA: a dataset for VQA on document images// Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. New York: IEEE, 2021: 2199—2208.

[23] Peng QM, Pan YX, Wang WJ, et al. ERNIE-layout: layout knowledge enhanced pre-training for visually-rich document understanding. (2022-10-12)/[2023-06-22]. https://arxiv.org/pdf/2210.06155.pdf.

[24] Mathew M, Bagal V, Tito R, et al. InfographicVQA// Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2022: 2582—2591.

[25] Powalski R, Borchmann Ł, Jurkiewicz D, et al. Going full-TILT boogie on document understanding with text-image-layout transformer// Lladós J, Lopresti D, Uchida S. International Conference on Document Analysis and Recognition. Cham: Springer, 2021: 732—747.

[26] Lei J, Yu LC, Bansal M, et al. TVQA: localized, compositional video question answering. (2018-09-05)/[2023-06-22]. https://arxiv.org/pdf/1809.01696.pdf.

[27] Zellers R, Lu JS, Lu XM, et al. MERLOT RESERVE: neural script knowledge through vision and language and sound// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 16354—16366.

[28] Rohrbach A, Torabi A, Rohrbach M, et al. Movie description. International Journal of Computer Vision, 2017, 123(1): 94—120.

[29] Zellers R, Lu XM, Hessel J, et al. Merlot: multimodal neural script knowledge models. Advances in Neural Information Processing Systems, 2021, 34: 23634—23651.

[30] Li JN, Li DX, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. (2023-01-30)/[2023-06-22]. https://arxiv.org/pdf/2301.12597.pdf.

[31] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation. (2021-02-24)/[2023-06-22]. https://arxiv.org/pdf/2102.12092.pdf.

[32] Ding M, Yang ZY, Hong WY, et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 2021, 34: 19822—19835.

[33] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. (2021-02-26)/[2023-06-22]. https://arxiv.org/pdf/2103.00020.pdf.

[34] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents. (2022-04-13)/[2023-06-22]. https://arxiv.org/pdf/2204.06125.pdf.

[35] Colin R, Noam S, Adam R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 2020, 21(1): 5485—5551.

[36] Balaji Y, Nah S, Huang X, et al. eDiff-I: text-to-image diffusion models with an ensemble of expert denoisers. (2022-11-02)/[2023-06-22]. https://arxiv.org/pdf/2211.01324.pdf.

[37] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context// Proceedings of the 2014 European Conference on Computer Vision. Cham: Springer, 2014: 740—755.

[38] Driess D, Xia F, Sajjadi MSM, et al. PaLM-E: an embodied multimodal language model. (2023-03-06)/[2023-06-22]. https://arxiv.org/pdf/2303.03378.pdf.

[39] Brohan A, Brown N, Carbajal J, et al. RT-2: vision-language-action models transfer web knowledge to robotic control. (2023-07-28)/[2023-09-27]. https://arxiv.org/pdf/2307.15818.pdf.

Inspiration of GPT-4 on Multimodal Foundation Models in Multimodal Understanding, Generation, and Interaction

Jing Liu[1, 2*] Longteng Guo[1, 2]

1. *Institute of Automation, Chinese Academy of Sciences, Beijing* 100190

2. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100090

Abstract ChatGPT, a conversational chatbot, has swept across society with its almost unstoppable momentum, heralding the dawn of general artificial intelligence. Its upgraded version, GPT-4, is a multimodal large-scale model that goes beyond monotonous text interactions and can accept combinations of text and images as multimodal inputs. Compared to traditional unimodal foundation models, multimodal foundation models are more consistent with human cognitive processes that involve multiple channels, allowing them to adapt to more complex environments, scenes and tasks. GPT-4 demonstrates that incorporating natural language understanding and generation abilities into multimodal foundation models can greatly enhance the model's abilities in multimodal understanding, generation, and interaction. This article introduces the concept of multimodal foundation models, key technologies, recent advancements, and application scenarios. It also discusses the technical characteristics of GPT-4 and specifically explore several inspirations provided by large language models, such as GPT-4, for building multimodal foundation models. Specifically, it discusses how to fully leverage the language capabilities of large language models to better perceive and understand the world, generate creative content, and interact with humans and the environment in the construction of multimodal foundation models.

Keywords GPT-4; multimodal foundation models; multimodal understanding; multimodal generation; multimodal interaction

(责任编辑 崔国增 张 强)

* Corresponding Author, Email: jliu@nlpr.ia.ac.cn