·专题:ChatGPT 与人工智能技术应用·

跨模态语言大模型:进展及展望

陈 露[1,2] 张思拓[1,2] 俞 凯[1,2*]

1. 上海交通大学 计算机科学与工程系跨媒体语言智能实验室,上海 200240
2. 上海交通大学 人工智能教育部重点实验室,上海 200240

[摘 要] 以 ChatGPT 为代表的对话式语言大模型通过使用超大规模模型参数和海量训练数据,涌现出很强的上下文学习能力和思维链推理能力,在各种自然语言处理任务上取得了显著的进步,被视为颠覆性通用人工智能技术。在纯文本语言大模型突破的基础上,近期显现的重要技术发展趋势是向能够理解和生成语音、图像、图形等其他模态数据的跨模态语言大模型的转变。随着大模型技术的快速发展,跨模态语言大模型逐步拥有了较强的多模态感知以及初步的跨模态认知能力。本文将从多模态感知大模型、跨模态认知大模型、以及分布式智能体系统三种范式综述跨模态语言大模型技术体系的演进过程,并总结相关的评测基准,最后讨论跨模态语言大模型面临的技术挑战及潜在重要研究方向。

[关键词] 语言大模型;多模态感知;跨模态认知;分布式智能体

生成式语言大模型(Chat Generative Pre-trained Transformer, ChatGPT)[①]凭借其出色的语言生成能力和类人思考能力,引发学术和产业界广泛关注,被视为颠覆性的“通用人工智能”技术里程碑。ChatGPT 的核心是基于大规模无标签数据的预训练大语言模型,在指令微调、基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)等技术加持下,它展现出卓越的任务通用泛化能力。此外,ChatGPT 还能实现复杂问题的分解和基于思维链的逐步推理,并提供深入的问题解答和建议。这种认知决策能力使其可被视为一种对话式的通用认知大模型。

尽管 ChatGPT 在纯文本自然语言处理任务上表现出色,但它无法直接实现对复杂多模态物理世界的认知以及通过交互对其产生影响。因此,为通用认知大模型引入多种模态的信息处理能力,无疑是通用人工智能技术发展的必然趋势。

纵观大模型的技术演进历程,可以清楚地看到多种模态的信息处理能力正逐步融入预训练大模型体系中。大规模自监督预训练早期起源于自然语言

俞凯 上海交通大学计算机科学与工程系特聘教授,上海交通大学苏州人工智能研究院执行院长,思必驰公司首席科学家。中国人工智能产业发展联盟学术和知识产权组组长,CCF 语音对话及听觉专委会副主任,中文信息学会理事。长期从事智能语音及语言处理的研究和产业化工作。发表国际期刊和会议论文 200 余篇,研究成果获国际期刊和会议最佳论文奖 6 次以及中国人工智能学会吴文俊人工智能自然科学奖一等奖等。

陈露 上海交通大学计算机科学与工程系助理研究员。主要研究兴趣包括智能人机对话系统、对话式大语言模型、自然语言处理等。目前已在 *IEEE Transactions on Pattern Analysis and Machine Intelligence*、*Neural Information Processing Systems*、*Annual Meeting of the Association for Computational Linguistics* 等国际会议和期刊上发表论文 40 余篇,获最佳论文奖或提名 2 次。作为项目或子课题负责人承担国家自然科学基金青年科学基金项目、重大研究计划(重点项目)等。其部分研究成果通过产学研合作获得大规模推广应用,并获第二十三届中国专利奖优秀奖。

收稿日期:2023-06-30;修回日期:2023-10-17

* 通信作者,Email: kai.yu@sjtu.edu.cn

本文受到国家自然科学基金项目(62120106006,62106142)和上海市市级科技重大专项(2021SHZDZX0102)的资助。

① chat.openai.com

处理领域，近年逐渐扩展到了图像、音频、视频等多模态数据处理任务中。例如，Meta AI 所训练的无监督语音模型 wav2vec[1]，Google 发布的文本到图像扩散模型 Imagen[2]，以及 OpenAI 提出的（Contrastive Language-Image Pretraining，CLIP）[3]文本图像匹配模型等等。随着 ChatGPT 等认知大模型的出现，研究焦点从面向特定任务的多模态感知，逐渐转变为更高层次的跨模态通用认知，整体技术演进呈现出如图 1 所示的三个范式转变：

首先是多模态感知大模型。此范式的研究焦点是特定任务的多模态数据感知和分析。大模型从视听文等不同模态的数据中，独立平等地提取各模态通道的信号，然后再进行对齐和融合。模态对齐和融合的操作通常比较简单，往往针对特定任务进行独立设计，也即“分别采集—通道融合—分别输出”的模式。在此框架中，各模态总体平等，任务目标以识别、匹配等感知任务为主，即使涉及到“理解”相关的任务，也往往采用针对特定任务的单独优化架构。

其次是跨模态认知大模型。语言大模型的出现促使研究重点从多模态感知向跨模态认知范式转变。这里的“跨模态”不同于各通道相对独立平等的“多模态”概念，其涵义是指各种模态信息被内生性的同时处理，不同模态信息在统一编码框架中自由交叉混合。模型架构中往往没有独立的模态融合模块，而是以大语言认知模型为核心，构建统一的语义空间，进行内在的联合处理，包括通用性理解、推理决策和语义生成，不同模态的输入、输出可能存在自由交叉。

最后是以认知大模型为核心的分布式智能体系统。在此范式中，多模态感知能力和其它专业技能与黑箱认知大模型解耦，整个架构将语言大模型作为核心控制器，以自然语言或者形式化语言为接口实现大模型与外部多模态感知模型以及专业工具的信息交换，形成“1＋N”的分布式智能体系统。与端到端的黑箱认知大模型相比，分布式智能体系统拥有更好的可扩展性和可解释性。进一步，这些系统可以通过与外部环境的互动，根据外部反馈和内部记忆实现持续学习，进而构建出不断进化的分布式智能体系统。

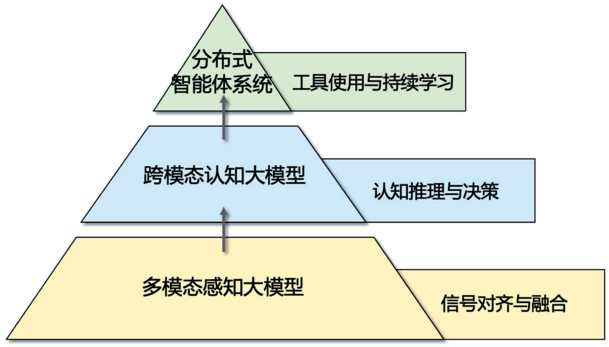

图 1　跨模态语言大模型三种范式概念关系图

鉴于多种模态信息处理的重要性，近些年众多的综述[4-6]对多模态数据的表示、对齐及融合策略相关的研究进行了深入总结，并构建了相对完善的多模态表示学习框架。然而，它们未涉及大语言模型引发的多模态研究范式的变革。Gan 等人[7]从典型多模态任务的视角进行综述，着重介绍了图像视觉、“图像—文本”及“视频—文本”任务，并突出了这些特定任务的多模态预训练方法与模型。但其重点仅限于视觉和文本的结合，对其他模态的讨论不够全面。他们集中关注传统的感知任务和如 VQA[8]这样的基础认知任务，对更高阶的复杂认知任务和通用智能探讨较少。Yin 等人[9]从四个维度探讨了跨模态语言大模型的进展，包括：多模态指令微调、上下文学习、思维链和文本语言大模型辅助视觉推理。但他们更多是对当前算法的细致总结，未对多模态模型的整体技术演进历史、理论框架和趋势进行深入分析。相对于上述工作，本文站在以语言符号协议为核心的通用人工智能统一框架的视角，从多模态感知大模型、跨模态认知大模型以及分布式智能体系统三大范式展开，系统性地回顾了跨模态语言大模型技术的发展历程，同时深入探索了技术的挑战与前沿，并总结了跨模态语言大模型评测基准的最新进展，为该领域的研究提供了一个全新的视角。

1　多模态感知大模型

多模态感知大模型的范式如图 2 所示：模型对每个模态的信号进行初步独立处理，提取各模态关键特征，进行信息融合后应用于特定的下游任务。

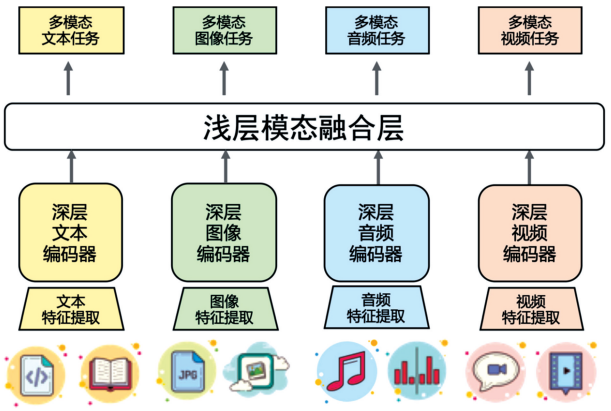

图 2　多模态感知大模型示意图

各个模态处理时的地位总体平等，融合方法和模型结构设计百花齐放，往往与特定下游任务相关。依据多模态特征提取和融合交互方式的不同，通常可以分为双编码器、融合编码器、统一骨干网络三种主流架构。

1.1 双编码器

双编码器将不同模态的数据分别编码，然后仅使用简单的相似度匹配将模态表征映射到同一特征空间。一种典型代表是 OpenAI 的 CLIP 模型。通过在大量的“图像—文本对”数据上进行预训练，CLIP 实现了图像与文本之间的联合理解和推理。CLIP 的文本编码器通常采用 Transformer[10]，图像编码器可以选择（Vision Transformer，ViT）[11]或 CNN。训练过程中，CLIP 利用对比学习目标，提高正样本（即图像及其相应的文本描述）的特征相似性，同时降低负样本（即图像与非配对的文本描述）的特征相似性，以此将图像和文本映射到同一特征空间，以理解并对齐模态信息。经过预训练后的 CLIP 模型在各种任务中，如图像分类、图像描述等，都表现出卓越的性能和泛化能力，例如 CLIP 在 ImageNet 上取得了 76.2%的零样本分类准确率。

在 CLIP 的基础上，AudioCLIP[12]增加了音频编码器，使得模型能够处理三种模态的数据，并泛化到语音模态任务。各种类似模型，如 ALIGN[13]和 CLAP[14]模型也借鉴了 CLIP 的双编码器结构和对比学习预训练策略，进一步推动了多模态感知技术的发展。

1.2 融合编码器

尽管以 CLIP 为代表的双编码器模型在下游分类和检索任务上展现了出色的泛化能力，然而在复杂推理（如视觉推理和视觉问答）任务中的表现仍然不佳。这主要归因于在模态融合阶段，CLIP 仅使用了简单的相似度匹配方法，无法充分综合多模态推理所需的不同模态信息。为解决这个问题，融合编码器架构在编码过程中就进行模态特征融合，以提取更深层次的跨模态特征。

融合编码器早期通常依赖于在特定任务上预训练的模型来提取多模态特征。例如，ViLBERT[15]和 UNITER[16]使用预训练的目标检测模型（如 Faster-RCNN[17]）来提取图像特征，类似的，AV-HuBERT[18]在多模态语音识别任务中使用预训练的 ResNet[19]来提取图像特征。这种方法特征提取效率较低，多模态表征能力也因在特定领域上的训练而受限。随着 Transformer 在各个单模态任务上的广泛应用，它也逐渐成为融合编码器中通用特征提取架构的主流技术。ViLT[20]提出去除传统的目标检测器，采用 ViT 的方式将图像转化为离散的图像块嵌入，和文本嵌入拼接后输入 Transformer 编码器进行模态融合，以此来建模图像和文本之间的联系。ViLT 极大提升了多模态特征提取阶段的推理速度，并在下游任务上取得了和之前模态深度融合方法（UNITER）相当的性能。

ALBEF[21]使用 ViT 来编码图像的特征，并使用 BERT[22]模型对文本进行编码，然后通过交叉注意力融合图像和文本的表示。在图像文本检索任务上，尽管 ALBEF 使用的预训练图像数量更少，其性能仍然远超 CLIP，达到了最佳水平。类似地，Paraskevopoulos[23]等人在视觉语音多模态识别任务上也使用了相似的方法。

1.3 统一骨干网络

尽管基于双编码器和融合编码器架构的模型在特定多模态处理任务上表现优异，但由于它们往往是针对单一任务进行设计、训练或优化，在解决多个不同的下游任务时，都需要进行额外的训练数据、模型架构、目标函数等调整，效率较低，性能调优复杂。因此，统一骨干网络应运而生，所有模态的数据在同一骨干网络中处理，以进一步增强不同模态间的交互和融合。同时，统一架构也利于处理来自不同下游任务的多种输出。

SimVLM[24]和 SpeechT5[25]就是这样的统一框架。各种多模态任务被统一建模为生成式任务，通过编码器—解码器结构，将多模态编码序列转换为输出的文本序列，采用简单的多模态掩码语言建模目标进行优化，让模型学习到直接生成与多模态输入相关的文本描述。统一的生成式训练目标简化了模型设计和训练流程，同时提升了模型的泛化能力和性能，在各种多模态任务上取得出色的表现。这些基于序列到序列的生成式统一骨干网络架构缩小了多模态任务和自然语言处理任务之间的鸿沟，为跨模态认知大模型的发展奠定了良好的基础。

2 跨模态认知大模型

多模态感知范式下，虽然已经出现了多种融合方法和统一架构，但各个模态都被视为独立感知通道，总体被平等对待，以优化特定任务为主。随着文本语言大模型展现出复杂推理和决策规划等通用认知能力，逐渐出现了以语言为核心的认知型跨模态语言大模型的研究趋势。跨模态认知范式下，各通

道的感知从属于通用认知目标，语言大模型成为认知处理核心，借助统一的计算框架，在统一的语义空间内实现对所有模态信息的全面融合，进行全方位的理解、推理决策以及语义生成。以下首先介绍跨模态认知模型的基本框架，然后探讨如何通过多模态指令微调来增强大模型的多模态思维链[26]等认知能力。

2.1　基础框架

如图 3 所示，跨模态认知大模型范式的主要思想是以语言为中心，将文本语言大模型用作处理各种模态编码的通用协议接口。这种策略可以理解为将不同模态的信息作为“外语”输入到语言大模型中，以实现联合建模。其优势在于能够将各种跨模态下游任务的预测统一转化为开放式文本生成，从而充分发挥语言大模型在小样本上下文学习和思维链推理方面的能力，实现处理复杂非特定任务的“通用智能”。这种方法与认知科学的双系统理论相符，其中，各模态的编码器可以视为系统一，负责快速的模态感知，而语言大模型可以视为系统二，它对感知到的多模态信息进行深度融合，并通过内部的认知系统推理预测输出结果。MetaLM[27]和 Kosmos[28, 29]就采用了这种策略，以端到端的方式从头训练出了以语言模型为核心的跨模态认知大模型。

随着语言模型规模的持续扩大，从头开始进行大规模跨模态预训练将带来巨大的计算成本。近期的一些研究工作选择采用预训练的大规模语言模型和单模态编码器进行初始化，并冻结了大部分基础模型的参数，只对模态映射部分的少量参数进行训练。例如，PaLM-E[30]将各种模态的编码器（如 ViT、传感器状态编码器等）与语言大模型 PaLM[31]结合，仅对编码器部分进行训练。同样，Frozen[32]也采取了这种策略。BLIP2[33]则选择冻结视觉编码

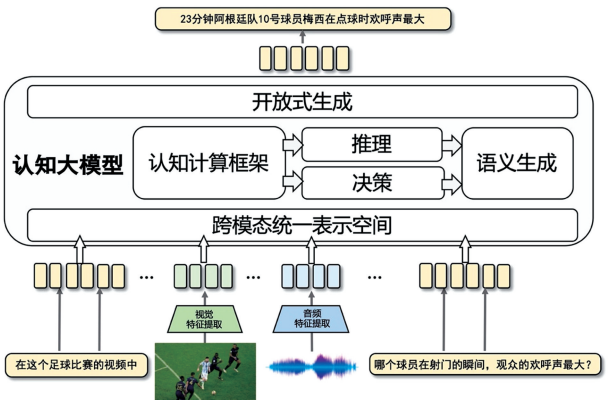

图 3　跨模态认知大模型示意图

器和语言大模型的参数，接着使用 Q-Former，以一组可学习的查询向量，通过交叉注意力机制建立模态之间的桥梁，使得语言大模型能理解并进一步处理视觉表征。

在保持原有认知推理能力的同时，为了增强对多模态数据的理解，一些研究工作试图在语言大模型的基础上插入额外的可学习参数，也即适配器（Adapter），以实现高效的微调。例如，Flamingo[34]在语言大模型中添加了额外的交叉注意力模块，以便引入多模态信息。通过在大规模多模态数据上训练，Flamingo 展现出了优异的多模态上下文学习能力。令人注目的是，即使不对模型进行微调，Flamingo 依然在 6 个任务上超越了当前最先进微调模型的表现。

另一方面，LLaMA-Adapter[35]则采用了一组可学习的多模态提示向量来引入多模态信息，并采用零初始化的注意力机制以确保训练效果及稳定性。在此基础上，LLaMA-AdapterV2[36]释放了更多的可学习参数，并整合了专家系统（如搜索引擎或 OCR 模型），以进一步提升对多模态数据的理解能力。

2.2　跨模态指令微调

在自然语言处理领域，指令微调（Instruction Tuning）是一种对语言大模型进行微调的技术，其目标是使模型能够根据各种自然语言指令执行多样化的任务。这种方法首先通过自然语言指令详细描述任务，然后在这些与人类意图高度一致的指令数据上对语言大模型进行微调。这样做可以显著增强模型的泛化能力，并激发出如复杂决策和思维链等显式认知能力。

LLaVA[37]率先在跨模态领域引入了指令微调的策略。他们利用现有的图像描述数据，结合 ChatGPT，构建了专门的跨模态指令微调数据集。这样的训练使得模型展现出与 GPT-4 论文描述相符的高级跨模态认知能力。例如，模型能够按照用户指令对图片进行详尽的描述和解析、处理图像中的问题，并能够详细展示解答的步骤。在经过微调后，LLaVA 在多模态科学多选题数据集 ScienceQA[38]上刷新了当前的最佳成绩。

尽管现有的跨模态认知模型已经展示出了一定的认知和推理能力，但目前还缺乏系统的跨模态指令微调数据集。此外，使用冻结编码器的方法可能会限制模态的感知能力，存在大语言模型固有的“幻觉（hallucination）”问题。特别是在专业领域，例如科学领域，这些问题可能会更加显著。

3 分布式智能体系统

端到端的跨模态认知大模型将不同模态的感知模块和语言大模型进行强耦合，虽然实现了较强的多模态感知和跨模态推理能力，但是仍然存在两方面的问题：一是扩展性较差，模型训练完成后不能动态加入处理其他模态数据的能力，二是不具有根据历史经验进行长期进化的能力。为了解决上述问题，最近越来越多的研究开始将大模型的感知能力和认知能力解耦，以认知大模型作为中心控制器，通过灵活模块化的方式调用其他多模态感知模型或工具，构建起了如图 4 所示的分布式“感知—认知”智能体系统。更进一步，这些系统通过与外部环境的交互实现持续学习，从而构建出一个能持续进化的分布式智能体系统，这种系统不需要进行大规模参数调整，就能实现多模态感知、认知决策和长期进化。

3.1 外部工具使用

如图 4 所示，跨模态分布式智能体系统主要由几个核心部分组成：首先是作为控制核心的认知模型控制器，通常采用文本语言大模型；其次是各种模态的模块化工具，这些外部“工具”包括各个模态的预训练模型以及各类感知或执行工具；最后，系统还会与能够提供反馈的外部环境进行互动，这些“环境”可能包括待处理的图像、视频、音频，或者包含各种信息的可交互环境等。

Socratic[39]模型是对此类系统的一种早期尝试。该模型首先利用各类预训练小模型（例如视觉—语言模型和音频—文本模型）对多模态信号进行感知，然后将小模型的的文本输出输入到文本语言大模型进行推理和决策，进而使文本语言大模型拥有了处理多模态信息的能力。SayCan[40]通过集成文本语言大模型的输出和与机器人行动策略相关的可供性函数，实现了高级的机器人决策规划。同时，Inner Monologue[41]采用文本形式将环境状态反馈给文本语言大模型，使其能感知到环境的变化，并据此动态调整其动作规划，以做出更精准的决策。

Visual ChatGPT[42]通过将视觉基础模型的描述和使用方法以提示的形式传递给 ChatGPT，使 ChatGPT 能够自主选择并调用特定的模型和方法，进而实现了基于文本语言大模型的多模态对话。HuggingGPT[43]赋予了 GPT 调用 HuggingFace 上预训练模型的能力，使其能够根据需要处理的任务，自主决定调用任意特定领域的专家模型。而 ViperGPT[44]和 CaP[45]则以代码形式将多模态模型和机器人控制功能提供给文本语言大模型，使其能够生成用于完成任务的可执行代码。其中 ViperGPT 在 OK-VQA[46]任务上超越了端到端模型 Flamingo 的零样本迁移性能。

3.2 记忆增强的持续学习

除了能够学习使用外部工具，人类作为具备通用智能的智能体，拥有持续与现实多模态世界互动来适应新环境和掌握新技能的能力，这使得人类具备了终身学习的可能性。然而，现行的文本语言大模型只能依赖输入的上下文进行规划和推理，它们没有记忆和更新机制，不能从以往的经验中学习和累积知识，这限制了它们的持续学习和能力提升。因此，为了建立一个具有持续学习能力的智能体系统，需要在大型模型系统中加入记忆、反思和长期规划的功能。

GITM[47]和 VOYAGER[48]模型在文本语言大模型的基础上，结合了外部多模态工具的使用，进一步引入了记忆模块和互联网知识库，成功构建了能

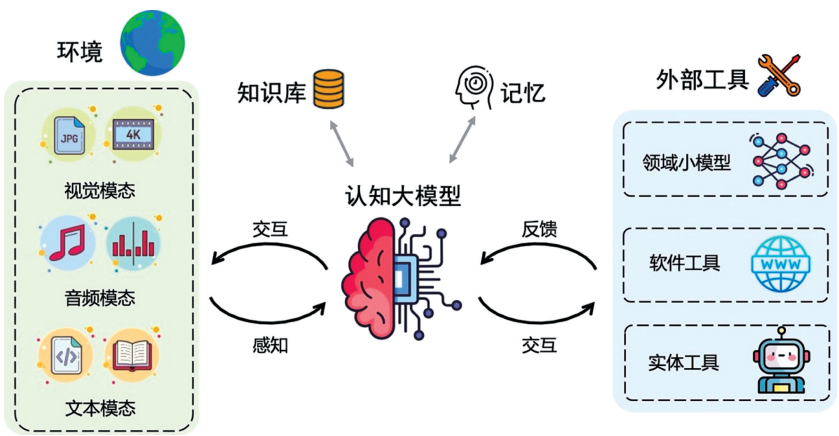

图 4 分布式智能体示意图

进行持续学习的多模态智能体。这种智能体不仅可以设定并追求长期目标，还能通过与环境的持续交互，积累经验并进行自我反思，以实现技能库的不断扩展，这种模式代表了一种全新的终身学习方式。斯坦福大学的研究团队在一个模拟小镇环境中构建了多个类似的智能体[49]，赋予它们记忆、长期规划和反思能力。这些智能体随即展现出了复杂的群体行为，包括信息的传递、新关系的建立等。

4　评测基准

客观、全面、多角度的评测基准的建立是多模态模型研究发展的重要组成部分。早期的评测基准往往局限于某一特定任务，如图像音频描述[50, 51]或多模态问答（例如 VQA[8]、ScienceQA[38]）。这类基准往往覆盖任务有限，尤其是高级认知推理部分往往覆盖不足。

最近随着语言大模型的兴起，出现了如（Massive Multitask Language Understanding，MMLU）[52]、C-Eval[53]等评测基准，旨在全面评估语言大模型的语言理解、生成和推理能力。它们的出现大大推动了文本语言大模型的快速迭代和能力提升。受此启示，研究者们也针对跨模态语言大模型提出了相应的评测基准，代表性的评测基准如表 1 所示。其中，OwlEval[54]包含了 82 个人工构建的问题，覆盖了从自然图像理解到 OCR 等多种测试任务，采用人工打分进行评测。MME[55]基准（Multimodal Large Language Model Evaluation）从感知和认知这两大维度设计了涵盖 14 个子任务的问题，所有问题为人工构造的“是/否”判断题，去除了大模型输出多样性的影响。其中 BLIP2 和 MiniGPT-4 在感知与认知两个维度上各自展现出了卓越的性能，同时实验结果也揭示了现有模型在指令跟随能力和推理能力等方面仍存在不足。在 MME 的基础上，MMBench[56]进一步精细化了跨模态能力的层次结构，分为 L-1 至 L-3 三个层级，并增加了更多的测试问题，所有的问题都设计为选择题形式。Kosmos-2 在该任务上取得了最佳成绩，表明在训练过程中引入目标定位数据可以提升模型的能力。LAMM-Benchmark（Language-assisted Multi-modal Benchmark）[57]则对多种传统数据集进行了整合，并扩展至三维点云模态。除了利用传统评估指标，他们还引入了 GPT 来对答案的相关性与准确性进行评估。LVLM-eHub（Large Vision-language Models Evaluation Hub）[58]则整合了六大类的跨模态能力，其中还涵盖了需要模型决策、规划的具身智能领域。LVLM-eHub 推出了“多模态大模型竞技场”，允许用户自由提出问题并对答案进行投票。在此榜单中，mPLUG-Owl[54]和 MiniGPT-4 均位列前三，这进一步凸显了跨模态指令微调训练的关键作用。

总之，跨模态语言大模型的评测基准正在从针对特定任务的评估进化为更加全面、立体和层次分明的评估方法，这为跨模态研究的深入发展提供了坚实的支撑。然而，现有的评测基准仍存在以下几个明显的不足之处：（1）覆盖场景局限：许多基准只针对少量特定的任务或应用场景，限制了其在通用场景中的适用性；（2）缺乏动态性：现有评测基准大都还是采用固定（静态）的样例式问答对，在内容上不能动态变化，在大模型时代容易造成数据泄露和信息过时；（3）评测不一致性：当前很多评测基准采用人工打分，这可能会受到评估者的主观偏见影响；尽管一些基准采用了 ChatGPT 进行自动评估，但由于 ChatGPT 对提示词的敏感性，可能会产生不一致的评测结果。因此，面对上述挑战，未来研究者需要进一步优化和完善跨模态语言大模型的评测基准，使其更具广泛性、动态性和一致性。

5　总结与展望

本文从多模态感知大模型、跨模态认知大模型以及分布式智能体系统三大范式系统性地介绍了跨模态语言大模型的技术进展和发展趋势。多模态感知大模型主要针对不同模态的信号进行感知、对齐

表 1　跨模态语言大模型评测基准

评测基准	覆盖能力	数据来源	评测方法
OwlEval[54]	指令理解、视觉理解、OCR、知识迁移、推理、多轮对话	人工构建	人工主观评测
MME[55]	感知能力、认知能力	人工构建	客观评测
MMBench[56]	感知能力、推理能力	人工构建	客观评测
LAMM-Benchmark[57]	图像理解、点云理解	已有数据集	客观评测
LVLM-eHub[58]	视觉感知、视觉知识获取、视觉推理、视觉常识、具身智能、幻觉	已有数据集＋在线评测平台	客观评测＋竞技场人工主观评测

和融合，形成了双编码器、融合编码器、统一骨干网络等各种典型模型架构，模型的多模态信号处理能力得到了显著提升。跨模态认知大模型使用以文本语言大模型为基座，在统一的语义空间内融合了各个模态的信息，实现了跨模态的语义理解和推理，在此基础上多模态指令微调则进一步提高了模型的认知和推理能力。分布式智能体系统将大模型的感知和认知能力解耦，以认知大模型为核心，通过外部工具调用，加入记忆、反思机制，初步实现类人的外部工具使用能力和根据外部反馈持续进化的能力。虽然跨模态大语言模型得到了快速发展，但是整个研究领域仍然处于初级阶段，面临着诸多挑战和机遇：

（1）在感知层面，目前研究主要集中在文本、语音、自然图像等多模态信号，而针对其他的多模态信号，例如传感器信号、文档中的图表、科学数据中的分子式等，研究较少。未来在跨模态语言大模型框架下亟需探索更多形式的多模态信号与大模型的结合，其中包括需要进一步研究多类型模态数据的表示方式和跨模态大模型的架构设计，重点包括设计统一的多模态特征编码空间，同时需要平衡语义信息和信号重建的能力。

（2）在认知层面，目前的跨模态语言大模型，虽然利用语言大模型本身强大的认知能力，实现了一定的跨模态小样本学习、思维链推理、序列决策等认知能力，但是整体认知能力仍然较弱，如何实现跨模态之间强认知能力的对齐是一个重要的研究问题。此外，另一个重要方向是提升跨模态语言大模型在推理和决策中的可解释性，为推理和决策提供清晰有效的论据支撑。

（3）在系统层面，虽然分布式智能体系统已经表现出了一定的持续学习能力，但是它们的长期规划和多步决策能力仍然显著不足。一方面目前的智能体系统通常被限定在特定的虚拟环境中，需要为每个环境设计特定的智能体框架，导致泛化和迁移能力较弱，无法实现真正的持续学习。另一方面，由于这些系统依赖于上下文学习，当引入更多的外部工具和专业小模型时，复杂情境下交互稳定性可能变弱，交互轮数会越来越多，可能触及模型输入长度上限。因此我们需要一方面研究大模型中超长多轮上下文的高效建模方法，另一方面探索外部工具库和历史记忆的高效存储、精准操控和快速检索方法。

（4）最后，构建全面的跨模态语言大模型评测基准也是一个关键的研究方向。亟需要建立覆盖场景广泛、动态性强、具有一致性的评估标准，以全方位地评估跨模态语言大模型的性能，这将对跨模态语言大模型研发具有重要的指导意义。

参 考 文 献

[1] Baevski A, Zhou H, Mohamed A, et al. Wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 2020: 12449—12460.

[2] 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.

[3] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. (2021-02-26)/[2023-06-29]. https://arxiv.org/pdf/2103.00020.pdf.

[4] Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423—443.

[5] Guo WZ, Wang JW, Wang SP. Deep multimodal representation learning: a survey. IEEE Access, 2019, 7: 63373—63394.

[6] Zhang C, Yang ZC, He XD, et al. Multimodal intelligence: representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 478—493.

[7] Gan Z, Li LJ, Li CY, et al. Vision-language pre-training: basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 2022, 14(3/4): 163—352.

[8] 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.

[9] Yin SK, Fu CY, Zhao SR, et al. A survey on multimodal large language models. (2023-06-23)/[2023-06-29]. https://arxiv.org/pdf/2306.13549.pdf.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000—6010.

[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. (2020-10-22)/[2023-06-29]. https://arxiv.org/pdf/2010.11929.pdf.

[12] Guzhov A, Raue F, Hees J, et al. Audioclip: extending clip to image, text and audio// Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2022: 976—980.

[13] Jia C, Yang YF, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. (2021-02-11)/[2023-06-29]. https://arxiv.org/pdf/2102.05918.pdf.

[14] Elizalde B, Deshmukh S, Al Ismail M, et al. CLAP learning audio concepts from natural language supervision// Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2023: 1—5.

[15] Lu JS, Batra D, Parikh D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 2019: 13—23.

[16] Chen YC, Li LJ, Yu LC, et al. Uniter: universal image-text representation learning// Proceedings of the 2020 European Conference on Computer Vision. Cham: Springer, 2020: 104—120.

[17] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137—1149.

[18] Shi BW, Hsu WN, Lakhotia K, et al. Learning audio-visual speech representation by masked multimodal cluster prediction. (2022-01-05)/[2023-06-29]. https://arxiv.org/pdf/2201.02184.pdf.

[19] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770—778.

[20] Kim W, Son B, Kim I. ViLT: vision-and-language transformer without convolution or region supervision. (2021-02-05)/[2023-06-29]. https://arxiv.org/pdf/2102.03334.pdf.

[21] Li JN, Selvaraju RR, Gotmare AD, et al. Align before fuse: vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 2021: 9694—9705.

[22] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171—4186.

[23] Paraskevopoulos G, Parthasarathy S, Khare A, et al. Multiresolution and multimodal speech recognition with transformers// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 2381—2387.

[24] Wang ZR, Yu JH, Yu AW, et al. SimVLM: simple visual language model pretraining with weak supervision. (2021-08-24)/[2023-06-29]. https://arxiv.org/pdf/2108.10904.pdf.

[25] Ao JY, Wang R, Zhou L, et al. Speecht5: unified-modal encoder-decoder pre-training for spoken language processing// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 5723—5738.

[26] Wei J, Wang XZ, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 2022: 24824—24837.

[27] Hao YR, Song HY, Dong L, et al. Language models are general-purpose interfaces. (2022-06-13)/[2023-06-29]. https://arxiv.org/pdf/2206.06336.pdf.

[28] Huang SH, Dong L, Wang WH, et al. Language is not all You need: aligning perception with language models. (2023-02-27)/[2023-06-29]. https://arxiv.org/pdf/2302.14045.pdf.

[29] Peng ZL, Wang WH, Dong L, et al. Kosmos-2: grounding multimodal large language models to the world. (2023-06-26)/[2023-06-29]. https://arxiv.org/pdf/2306.14824.pdf.

[30] Driess D, Xia F, Sajjadi MSM, et al. PaLM-E: an embodied multimodal language model. (2023-03-06)/[2023-06-29]. https://arxiv.org/pdf/2303.03378.pdf.

[31] Chowdhery A, Narang SR, Devlin J, et al. PaLM: scaling language modeling with pathways. (2022-04-05)/[2023-06-29]. https://arxiv.org/pdf/2204.02311.pdf.

[32] Tsimpoukelli M, Menick J, Cabi S, et al. Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 2021: 200—212.

[33] 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)/[2023-06-29]. https://arxiv.org/pdf/2301.12597.pdf.

[34] 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.

[35] Zhang RR, Han JM, Liu C, et al. LLaMA-adapter: efficient fine-tuning of language models with zero-init attention. (2023-03-28)/[2023-06-29]. https://arxiv.org/pdf/2303.16199.pdf.

[36] Gao P, Han JM, Zhang RR, et al. LLaMA-adapter V2: parameter-efficient visual instruction model. (2023-04-28)/[2023-06-29]. https://arxiv.org/pdf/2304.15010.pdf.

[37] Liu HT, Li CY, Wu QY, et al. Visual instruction tuning. (2023-04-17)/[2023-06-29]. https://arxiv.org/pdf/2304.08485.pdf.

[38] Lu P, Mishra S, Xia T, et al. Learn to explain: multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 2022, 35: 2507—2521.

[39] Zeng A, Attarian M, Ichter B, et al. Socratic models: composing zero-shot multimodal reasoning with language. (2022-04-01)/[2023-06-29]. https://arxiv.org/pdf/2204.00598.pdf.

[40] Ahn M, Brohan A, Brown N, et al. Do As I can, not As I say: grounding language in robotic affordances. (2022-04-04)/[2023-06-29]. https://arxiv.org/pdf/2204.01691.pdf.

[41] Huang WL, Xia F, Xiao T, et al. Inner monologue: embodied reasoning through planning with language models. (2022-07-12)/[2023-06-29]. https://arxiv.org/pdf/2207.05608.pdf.

[42] Wu CF, Yin SM, Qi WZ, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models. (2023-03-08)/[2023-06-29]. https://arxiv.org/pdf/2303.04671.pdf.

[43] Shen YL, Song KT, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face. (2023-03-30)/[2023-06-29]. https://arxiv.org/pdf/2303.17580.pdf.

[44] Surís D, Menon S, Vondrick C. ViperGPT: visual inference via python execution for reasoning. (2023-03-14)/[2023-06-29]. https://arxiv.org/pdf/2303.08128.pdf.

[45] Liang J, Huang WL, Xia F, et al. Code as policies: language model programs for embodied control// Proceedings of the 2023 IEEE International Conference on Robotics and Automation. New York: IEEE, 2023: 9493—9500.

[46] Marino K, Rastegari M, Farhadi A, et al. OK-VQA: a visual question answering benchmark requiring external knowledge// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 3190—3199.

[47] Zhu XZ, Chen YT, Tian H, et al. Ghost in the minecraft: generally capable agents for open-world environments via large language models with text-based knowledge and memory. (2023-05-25)/[2023-06-29]. https://arxiv.org/pdf/2305.17144.pdf.

[48] Wang GZ, Xie YQ, Jiang YF, et al. Voyager: an open-ended embodied agent with large language models. (2023-05-25)/[2023-06-29]. https://arxiv.org/pdf/2305.16291.pdf.

[49] Park JS, O'Brien JC, Cai CJ, et al. Generative agents: interactive simulacra of human behavior. (2023-04-07)/[2023-06-29]. https://arxiv.org/pdf/2304.03442.pdf.

[50] 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.

[51] Kim CD, Kim B, Lee H, et al. Audiocaps: generating captions for audios in the wild// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019: 119—132.

[52] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. (2020-09-07)/[2023-06-29]. https://arxiv.org/pdf/2009.03300.pdf.

[53] Huang YZ, Bai YZ, Zhu ZH, et al. C-eval: a multi-level multi-discipline Chinese evaluation suite for foundation models. (2023-05-15)/[2023-06-29]. https://arxiv.org/pdf/2305.08322.pdf.

[54] Ye QH, Xu HY, Xu GH, et al. mPLUG-owl: modularization empowers large language models with multimodality. (2023-04-27)/[2023-06-29]. https://arxiv.org/pdf/2304.14178.pdf.

[55] Fu CY, Chen PX, Shen YH, et al. MME: a comprehensive evaluation benchmark for multimodal large language models. (2023-06-23)/[2023-09-05]. https://arxiv.org/pdf/2306.13394.pdf.

[56] Liu Y, Duan HD, Zhang YH, et al. MMBench: is your multi-modal model an all-around player? (2023-07-12)/[2023-09-05]. https://arxiv.org/pdf/2307.06281.pdf.

[57] Yin ZF, Wang J, Cao JJ, et al. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. (2023-06-11)/[2023-09-05]. https://arxiv.org/pdf/2306.06687.pdf.

[58] Xu P, Shao WQ, Zhang KP, et al. LVLM-eHub: a comprehensive evaluation benchmark for large vision-language models. (2023-06-15)/[2023-09-05]. https://arxiv.org/pdf/2306.09265.pdf.

Cross-modal Large Language Models: Progress and Prospects

Lu Chen[1,2]　Situo Zhang[1,2]　Kai Yu[1,2*]

1. *X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240
2. *MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai* 200240

Abstract　Conversational large language models (LLMs), such as ChatGPT, have achieved remarkable advancements in in-context learning and reasoning abilities by utilizing massive training data and large-scale model parameters. Building upon the breakthroughs in text-based language models, there has recently been a significant technological trend towards understanding and generating other modalities, such as speech, images, and graphics. This trend has led to the transition into cross-modal LLMs. With the rapid development of large models, cross-modal LLMs have gradually acquired strong multimodal perception and initial cross-modal cognitive abilities. This article first provides a comprehensive overview of the evolution of cross-modal LLM technology from three perspectives: multimodal large perception models, cross-modal large cognitive models, and distributed agent systems, then summarizes the relevant evaluation benchmarks. Additionally, the article discusses the technical challenges and potential research directions that cross-modal LLMs are currently facing.

Keywords　large language model; multimodal perception; cross-modal cognition; distributed agent

(责任编辑　崔国增　张　强)

* Corresponding Author, Email: kai.yu@sjtu.edu.cn