多模态大语言模型技术发展报告

主编单位：中科算网算泥社区

2026年2月

主编单位：中科算网科技有限公司　算泥 AI 开发者社区（https://c.sumw.com.cn）

目 录

序言

1. 研究背景与动机

人工智能的发展正进入一个以多模态融合为核心标志的新纪元。继大型语言模型（Large Language Models，LLMs）在自然语言处理领域取得革命性突破之后 AI 研究的焦点正迅速转向能够同时理解和生成文本、图像、音频、视频乃至更复杂模态信息的统一模型。2025 年我们见证了多模态大语言模型的爆发式增长其技术迭代速度和能力边界的拓展远超预期，深刻地重塑着人机交互的范式、内容创作的流程以及科学研究的边界。

从早期的双流架构探索如 ViLBERT 和 LXMERT 到 CLIP 凭借对比学习实现视觉与语言的深度对齐多模态技术的发展历经了漫长的积累。然而直到 2023 年随着 LLaVA 等工作的出现将视觉编码器与大型语言模型相结合的“指令微调”（Instruction Tuning）范式才真正点燃了社区的热情使得模型能够以前所未有的方式遵循人类指令来执行多模态任务。这一阶段开源社区的繁荣特别是 LLaMA 系列模型的开放极大地加速了技术的普及与创新。

进入 2024 年研究的重点转向了“统一建模”。以 Meta 的 Chameleon 和谷歌的 VITRON 为代表的模型开始尝试在单一架构内统一理解与生成任务打破了两者之间的壁垒。Show-o 等工作更是探索了自回归（Autoregressive）与扩散（Diffusion）两种生成范式的混合旨在兼顾生成质量与效率。这些探索为 2025 年的技术爆发奠定了坚实的基础。

2025 年我们目睹了多模态技术从“统一”走向“全能”的飞跃。以 Janus 为代表的“解耦设计”理念通过为理解和生成任务提供独立的视觉编码路径显著提升了模型的综合性能解决了早期融合架构的内在冲突。紧接着以 JanusFlow 和 NExT-OMNI 为代表的模型创新性地引入了整流流（Rectified Flow）和离散流匹配（Discrete Flow Matching）等更先进的生成范式进一步提升了生成质量和效率。在应用层面 VITA-1.5 在实时视觉-语音交互方面取得了接近 GPT-4o 的性能而阿里巴巴的 Qwen3-Omni 则首次在单一原生全模态模型中实现了跨越文本、图像、音频、视频所有主流模态的最先进性能。与此同时 Mogao 在交错多模态内容生成方面的突破预示着 AI 在内容创作领域将扮演更为核心的角色。

在这一波澜壮阔的技术浪潮中新的架构、训练方法、数据集和评估基准层出不穷知识的更新速度呈指数级增长。然而信息的碎片化和技术细节的复杂性也为

研究人员、开发者和决策者带来了巨大的挑战。系统性地梳理多模态大语言模型的技术脉络评估其能力边界洞察其未来走向变得至关重要且异常紧迫。

在此背景下，作为国内领先的 AI 大模型开发服务平台，算泥社区秉持“技术专业、生态开放、开发者友好”的理念，联合社区众多资深分析师与技术专家、学者，共同撰写并发布《2025 多模态大语言模型技术发展报告》。我们的目标是提供一份全面、权威且具有前瞻性的技术报告，系统性地回顾多模态大语言模型的发展历程，深度剖析截至目前涌现的核心技术创新，详细梳理关键的数据来源与评估基准，全面展示其在各个领域的应用实践并客观分析当前面临的挑战与未来的发展机遇。我们希望通过这份报告为学术界的研究人员提供清晰的技术路线图，为工业界的开发者提供可靠的实践指南，为相关领域的决策者提供科学的战略参考，共同推动多模态人工智能技术健康、快速地发展。

2. 多模态大语言模型的定义与范畴

为了系统性地展开本报告的论述首先必须对“多模态大语言模型”（Multimodal Large Language Models，MLLMs）的核心概念及其范畴进行清晰的界定。广义上多模态大语言模型是指一类能够处理、理解、关联和生成两种或两种以上不同模态信息的人工智能大语言模型。这些模型通常以一个强大的大型语言模型（LLM）为核心通过特定的架构设计将 LLM 的语言能力扩展到非文本模态从而实现跨模态的智能处理。

模态（Modality）在本报告中指代信息的特定表现形式。当前多模态大语言模型研究涵盖的主要模态包括：

文本（Text）：作为所有 MLLMs 的基础提供核心的语义理解、逻辑推理和指令遵循能力。

视觉（Vision）：包括静态图像（Image）和动态视频（Video）是当前研究最活跃、应用最广泛的非文本模态。

音频（Audio）：涵盖语音（Speech）、音乐（Music）和通用声音事件（Sound Events）是实现自然人机交互的关键。

动作（Action）：主要应用于具身智能（Embodied AI）和机器人领域指代模型输出的物理或虚拟环境中的动作序列。

其他模态：还包括三维（3D）表示、热成像、表格、图表、分子结构等更专业的模态这些模态的整合正在成为新的研究前沿。

基于模型对不同模态的处理能力和架构设计我们可以从以下几个维度对多

模态大语言模型进行划分：

表 1：多模态大语言模型的分类维度

分类维度	类型一	类型二	关键区别
模态覆盖范围	专用多模态模型	全模态模型 (Omni-Modal Models)	专用模型通常处理两种或三种特定模态（如文本-图像）而全模态模型旨在统一处理所有主流模态（如文本、图像、音频、视频等）。2025 年的 Qwen3-Omni 和 NExT-OMNI 是全模态模型的杰出代表。
任务统一性	分离式模型	统一模型 (Unified Models)	分离式模型为理解和生成任务使用不同的模型或模块而统一模型则在单一架构内同时支持理解和生成。Chameleon 和 Janus 是统一模型的重要里程碑。
架构设计	外部专家集成	端到端统一模型	外部专家集成模型（如 Visual ChatGPT）通过 LLM 调用外部工具来处理多模态任务而端到端模型则在单一网络中完成所有处理是当前的主流发展方向。

一个核心的演进趋势是从理解到生成的统一。早期的多模态模型主要聚焦于“理解”任务如视觉问答（VQA）或图像描述。然而随着生成模型特别是扩散模型和流模型的成熟新一代的多模态大语言模型已经具备了强大的“生成”能力能够根据文本或多模态输入创造出全新的图像、视频或音频内容。这种理解与生成的统一是衡量现代多模态大语言模型能力的关键标准。

本报告将重点关注那些致力于实现任务统一和端到端设计的多模态大语言模型特别是那些在 2024 年至 2026 年间发布、推动技术边界向前发展的模型。我们将深入探讨它们如何通过创新的架构设计和训练方法逐步实现对更多模态的覆盖并最终迈向能够处理任意模态输入和输出的“全模态智能”这一宏伟目标。

3. 报告研究方法与数据来源

本报告通过多源信息交叉验证力求客观、准确地反映 2025 年多模态大语言模型的技术全景。

报告的核心信息来源于对全球顶级学术会议和预印本平台的系统性文献检索。我们的主要信息来源包括：

顶级人工智能会议：重点关注计算机视觉（CVPR，ICCV，ECCV）、机器学习（NeurIPS，ICLR，ICML）和自然语言处理（ACL，EMNLP）领域的顶级会议论文特别是 2024 年和 2025 年的最新发表成果。这些经过同行评议的论文构成了本报告最核心的技术依据。

arXiv 平台：持续追踪 arXiv（特别是 cs.CV，cs.CL，cs.AI，cs.LG 等子领

域）的最新动态。鉴于多模态领域技术迭代速度极快许多重要的研究成果（如 Chameleon，Janus）首先在 arXiv 上发布是获取最前沿信息不可或缺的渠道。

官方技术报告与博客：直接参考由顶尖研究机构（如 Meta AI，Google AI，OpenAI，阿里巴巴）发布的官方技术报告、白皮书和博客文章。这些资料为我们理解闭源模型（如 GPT-4o）和工业界最新产品（如 Qwen3-Omni）提供了权威的一手信息。

在文献筛选上我们优先选择那些被广泛引用、在重要会议上获得奖项（如 NeurIPS Highlight，ICLR Oral）、或由知名研究团队发布的论文。

采用了多维度的分析框架将宏观趋势与微观技术细节相结合：

时间序列分析：以技术发展的里程碑事件为节点将多模态大语言模型的演进划分为四个主要阶段（早期探索、快速发展、统一建模、全模态爆发）清晰地展示了技术演进的脉络。

技术主题分析：围绕“核心技术架构”这一主线对建模范式、编码器设计、对齐机制、生成范式和训练方法等关键技术主题进行深度剖析揭示其内在的演进逻辑。

应用驱动分析：从实际应用场景出发分析多模态技术如何在视觉理解、内容创作、实时交互和具身智能等领域创造价值并结合典型案例进行说明。

通过这一分析框架本报告旨在避免单纯的技术罗列而是力图构建一个结构化、有深度的知识体系帮助读者更好地理解“为什么”和“怎么样”而不仅仅是“是什么”。报告中的每一项关键论断、数据和结论均在附录部分提供了详尽的参考文献和来源链接以供读者查证和深入研究。

尽管我们力求全面和客观，但必须承认本报告仍存在许多局限性：

信息时效性。多模态技术发展极快，最新进展可能在数月后就被新的突破所超越。我们建议读者将本报告作为理解技术脉络和把握发展趋势的参考，而非最终定论。

模型数目庞大，尤其是开源模型，报告仅选取一些极具代表性的模型，从技术发展的视角来解读，难以覆盖多模态模型发展的复杂技术进步和创新细节。

闭源模型信息不完整。对于 GPT-4o、Gemini、文心 5.0 等闭源商业模型，由于技术细节未完全公开，我们的分析主要基于官方发布的技术报告和公开演示，可能无法完全反映其内部实现。

具身智能及世界模型是新兴的前沿领域，本报告只是少量提及，展示其与多

模态大语言模型的紧密联系。

4. 报告结构与阅读指南

本报告主体部分共分五章辅以详尽的附录旨在为读者提供一条从宏观到微观、从理论到实践的清晰认知路径。我们建议读者根据自身需求选择性阅读亦可通读全文以建立系统性认知。

第一章：多模态大语言模型发展历程。本章将以时间为轴系统回顾多模态技术从早期探索到 2025 年全模态爆发的四个关键阶段重点梳理各阶段的里程碑事件和代表性工作为理解当前的技术格局提供历史视角。

第二章：核心技术架构与训练方法的进化。我们将深度剖析支撑现代多模态大语言模型的关键技术包括建模范式、视觉编码器、对齐机制、生成范式以及训练方法的演进。本章将重点解读 2025 年涌现的解耦设计、流模型等前沿技术。

第三章：数据来源与评估基准。本章将系统梳理多模态模型的“养料”与“标尺”详细介绍主流的预训练数据集、指令微调数据集并对各类评估基准（如 MME，Video-MME）进行分析为读者提供评估和选择模型的技术参考。

第四章：应用场景与实践。本章将展示多模态技术如何从实验室走向现实世界覆盖视觉理解、内容创作、实时交互、具身智能等多个热门应用领域并结合 VITA-1.5、ChartMoE 等模型的实践案例进行说明。

第五章：当前挑战与未来展望。本章将客观分析多模态技术在计算资源、数据、安全性等方面面临的挑战并在此基础上对世界模型、通用人工智能等未来发展方向进行展望。

附录部分包含了详尽的专业术语表、参考文献列表和推荐资源是本报告的重要组成部分。所有在正文中出现的专业术语和引用文献均可在附录中找到详细的解释和来源链接。本报告的目标读者群体广泛包括但不限于人工智能领域的研究人员、算法工程师、产品经理、高校师生、科技行业的投资者以及对前沿科技感兴趣的公众。对于有开发需求的团队或个人，算泥社区平台通过整合国产异构算资源，为开发者提供了经济高效的算选择。

5. 核心发现与关键洞察

经过系统性的研究与分析本报告提炼出以下关于 2025 年多模态大语言模型发展的核心发现与关键洞察：

2025 年是“全模态元年”技术范式发生根本性转变。技术演进的核心驱动力

从“统一理解与生成”转向“追求全能与实时”。以解耦设计（Decoupling）、流模型（Flow Models）和 原生全模态（Native Omni-Modal）为代表的三大技术突破共同定义了 2025 年的技术新高度使得模型在能力边界和交互体验上取得了质的飞跃。

混合生成范式成为主流流模型潜力巨大。纯粹的自回归或扩散模型正被更高效、更高质量的混合范式所取代。特别是以 Rectified Flow 和 Discrete Flow 为代表的流模型因其理论上的优雅性和实践中的高效性在 JanusFlow 和 NExT-OMNI 等前沿工作中展现出巨大潜力有望成为下一代生成模型的核心技术。

实时交互与交错生成是应用落地的关键。以 VITA-1.5 为代表的实时视觉-语音交互能力以及以 Mogao 为代表的交错多模态内容生成能力极大地提升了用户体验和 AI 的实用价值。这标志着多模态技术正从“可用”迈向“好用”为在消费电子、内容创作、在线教育等领域的规模化应用铺平了道路。

开源生态持续繁荣但与顶级闭源模型的差距依然存在。以 Qwen3-Omni、VITA 系列等为代表的开源模型在 2025 年取得了长足进步部分能力已能对标 GPT-4o 等顶级闭源模型。然而在模型的稳定性、长上下文处理能力和复杂推理的可靠性方面差距依然存在。开源社区的快速迭代和工业界的持续投入将是弥合差距的关键。

数据和评估的挑战日益凸显。随着模型能力的增强对高质量、多样化的多模态数据（特别是视频和交错数据）的需求变得空前迫切。同时现有的评估基准在衡量模型的真实世界能力特别是交互能力和安全性方面仍显不足。构建更全面的数据生态和更科学的评估体系是推动领域健康发展的当务之急。

综上所述 2025 年的多模态大语言模型领域呈现出技术加速迭代、应用场景快速拓展、开源与闭源激烈竞争的繁荣景象。我们正处在一个由多模态技术定义的“AI2.0”时代的开端其深远影响将在未来几年内持续显现。

第一章：多模态大语言模型发展历程

1.1 早期探索阶段（2017-2020）：奠基与探索

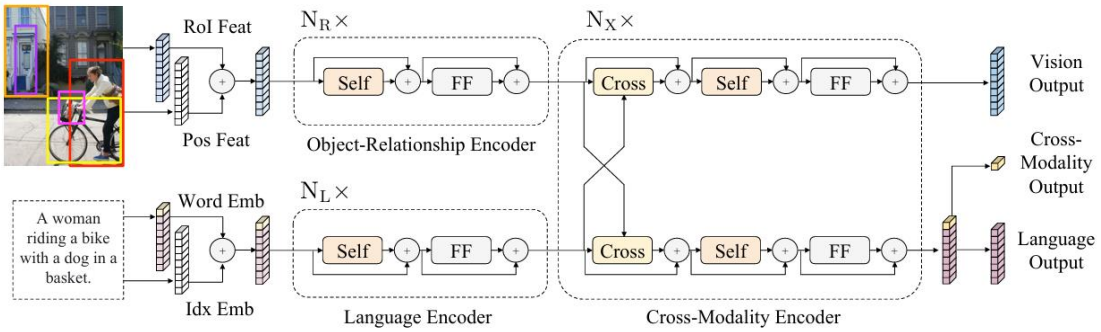
多模态大语言模型的历史根植于深度学习在自然语言处理（NLP）和计算机视觉（CV）两大领域的独立突破。2017 年 Vaswani 等人提出的 Transformer 架构以其强大的并行计算能力和对长距离依赖的建模优势彻底改变了 NLP 领域。这一成功激发了研究者们将其扩展到多模态领域的雄心。这一时期的核心议题是：

如何有效地融合基于 Transformer 的语言理解能力与视觉表示以解决跨模态的理解任务。因此这一阶段可以被视为多模态大语言模型的奠基与探索期其主要特征是双流架构的流行和对比学习的萌芽。

1.1.1 视觉-语言模型的起源：双流架构的探索

在 Transformer 的启发下第一批真正意义上的视觉-语言预训练模型（Vision-Language Pre-training，VLP）在 2019 年集中涌现。其中的代表性工作是 ViLBERT 和 LXMERT。这些模型开创性地采用了双流（Two-Stream）架构。其核心思想是为视觉和文本模态分别设置独立的 Transformer 编码器以充分学习各自模态内的特征然后再通过一个跨模态（Cross-Modal）Transformer 编码器进行深度融合。

以 LXMERT 为例其架构包含一个对象关系编码器（基于 Faster R-CNN 提取的区域特征）、一个语言编码器和一个跨模态编码器。这种设计允许模型在融合前对每个模态进行独立的上下文建模。



为了训练这些复杂的模型研究者们设计了一系列新颖的预训练任务。这些任务借鉴了 NLP 领域的成功经验（如 BERT 的掩码语言模型）并将其扩展到多模态场景。常见的任务包括：

掩码多模态建模（Masked Multi-Modal Modeling）：随机掩盖输入文本中的部分单词或图像中的部分区域特征然后让模型根据剩余的上下文进行预测。这迫使模型学习模态内部和模态之间的细粒度关联。

跨模态对齐预测（Cross-Modal Alignment Prediction）：向模型输入一对图像和文本让其判断两者是否匹配。这个任务旨在让模型学习更高层次的图文语义对应关系。

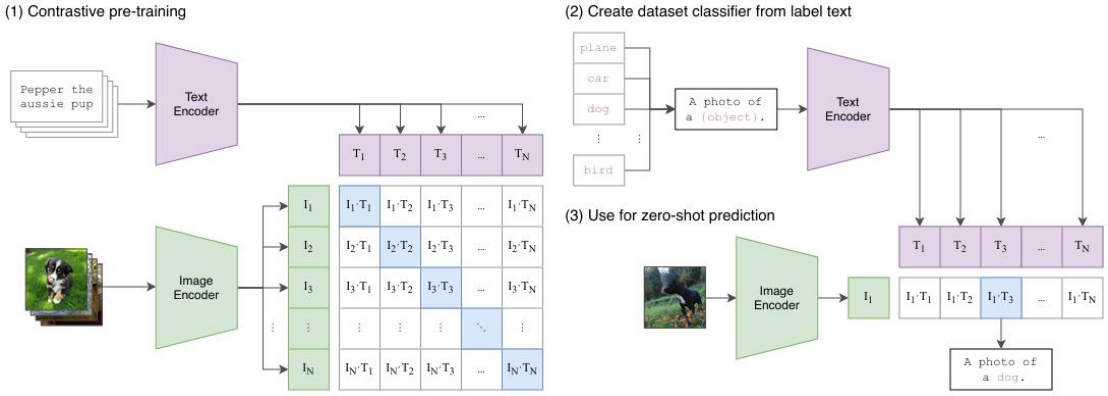
这些早期的双流模型在多个下游视觉-语言任务（如视觉问答 VQA、视觉常识推理 VCR、图文检索）上取得了显著的性能提升证明了大规模跨模态预训练

的有效性。然而它们也存在明显的局限：复杂的网络结构导致计算成本高昂且模态间的交互仅发生在顶层的融合模块限制了更深层次的特征融合。

1.1.2 跨模态对齐的突破：CLIP 与对比学习

双流模型证明了跨模态预训练的可行性 OpenAI 在 2021 年初发布的 CLIP（Contrastive Language-Image Pre-training）则彻底改变了该领域的游戏规则。CLIP 的贡献是革命性的它摒弃了复杂的融合模块和像素级的预测任务转而采用一种更为简洁、高效且可扩展的对比学习（Contrastive Learning）范式。

CLIP 的核心思想是：直接从互联网上收集的海量（4 亿）图文对数据中学习一种统一的跨模态嵌入空间。它包含一个图像编码器和一个文本编码器。



在训练过程中对于一个批次内的 N 个图文对模型的目标是正确地将 N 个图像与其对应的 N 个文本描述匹配起来同时将不匹配的 N^2-N 个组合推开。通过这种方式 CLIP 学习到的视觉特征与自然语言在语义上深度对齐。

表 2：早期双流模型与 CLIP 的对比

特征	早期双流模型 (如 LXMERT)	CLIP
核心思想	融合与预测	对齐与匹配
架构	双流独立编码 + 跨模态融合	双流独立编码无显式融合模块
训练目标	掩码预测、对齐分类	对比损失（Contrastive Loss）
数据规模	百万级	亿级（4 亿）
下游任务范式	预训练-微调（Pre-train &Fine-tune）	零样本/少样本预测（Zero/Few-shot）

CLIP 最惊人的能力在于其强大的零样本泛化能力。由于其视觉概念是与自然语言直接关联的因此无需任何微调就可以通过构建文本提示（如“一张...的照片”）来完成任意视觉分类任务其性能甚至超过了在特定数据集上监督训练的 ResNet 模型。这一突破打破了长期以来“预训练-微 调”的范式为后续的多模态大语言模型发展指明了新的方向：即利用海量自然监督信号通过对比学习构建一

个统一的、可泛化的多模态语义空间。

1.1.3 技术局限与挑战

尽管取得了显著进展但这一早期探索阶段的多模态模型仍面临诸多挑战这些挑战也预示了未来的研究方向：

生成能力的缺失：无论是双流模型还是 CLIP 其设计都主要面向理解任务。它们能够判断图文是否匹配或对图像进行分类但无法根据文本描述生成一张全新的图像。这种生成能力的缺失是该阶段模型最大的局限。

模态融合的深度不足：双流模型虽然有跨模态融合模块但融合发生在较高层次限制了模态间更底层的交互。而 CLIP 则完全没有显式的融合机制其对齐是“全局”而非“局部”的难以处理需要细粒度对应关系的任务（如视觉定位）。

对高质量标注数据的依赖：早期的 VLP 模型依赖于经过清洗和标注的数据集（如 COCO，Visual Genome）规模受限。虽然 CLIP 使用了更大规模的带噪声网络数据但如何有效利用这些数据以及数据偏见带来的问题仍是悬而未决的挑战。

计算资源的巨大消耗：双流架构的复杂性和大规模预训练的需求使得这些模型的训练成本极高只有少数大型研究机构能够承担阻碍了更广泛的研究和应用。

总而言之 2017 年至 2020 年的早期探索阶段成功地将 Transformer 架构引入多模态领域并通过双流架构和对比学习两种不同的路径验证了大规模预训练在视觉-语言任务上的巨大潜力。特别是 CLIP 的出现为后续研究奠定了“对齐”这一核心思想。然而生成能力的缺失和模态融合的浅层性等问题也为下一阶段的技术突破埋下了伏笔。

1.2 快速发展阶段（2021-2023）：LLM 驱动的范式革命

进入 2021 年尤其是在 2022 年末 ChatGPT 发布之后大型语言模型（LLMs）展现出的强大零样本学习、指令遵循和上下文学习能力为整个人工智能领域带来了深刻的范式革命。多模态领域迅速捕捉到这一变革信号研究重心从“从零开始设计复杂的跨模态融合架构”转向“如何将强大的预训练 LLM 适配到多模态任务中”。这一阶段的核心特征是 LLM 作为多模态智能的核心以及视觉指令微调（Visual Instruction Tuning）成为主流技术路线。

1.2.1 大语言模型的崛起及其对多模态的启发

GPT-3 的发布及其后续模型的演进揭示了一个关键事实：当模型规模足够

大并在海量文本数据上进行预训练后会涌现出惊人的泛化能力。模型不再仅仅是学习语言的统计规律而是开始具备一定程度的常识推理和世界知识。这为多模态研究者提供了新的思路：与其构建复杂的专用模型不如利用 LLM 已经具备的强大推理和语言能力仅需教会它“看懂”图像即可。

这一思路的转变带来了几个关键优势：

继承 LLM 的强大能力：可以直接利用 LLM 的语言生成、代码理解、逻辑推理等高级能力并将其自然地迁移到多模态对话和任务中。

简化架构设计：无需再设计复杂的跨模态融合模块只需一个轻量级的“适配器”将视觉特征连接到 LLM 即可。

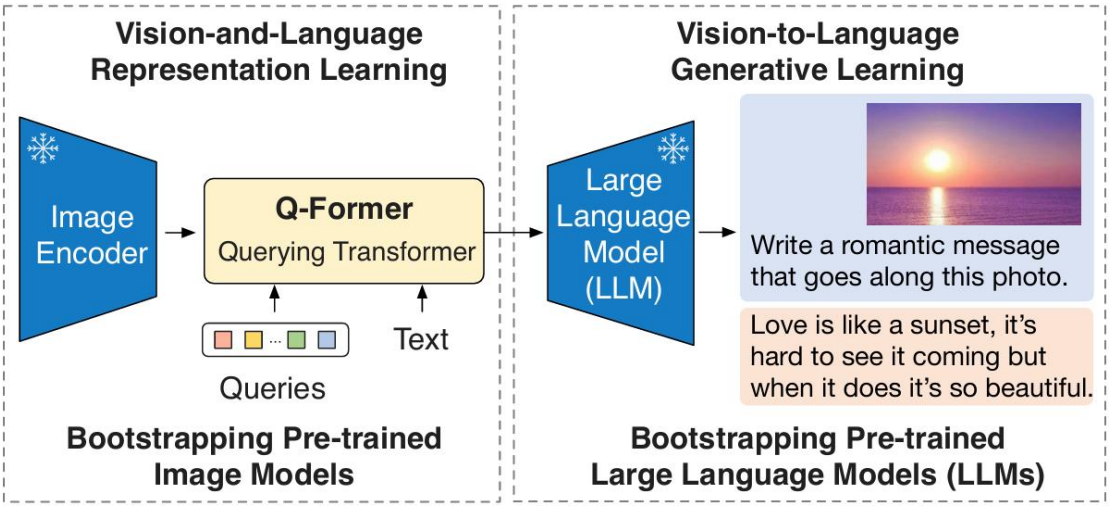
提升数据效率：由于 LLM 已经经过大规模预训练多模态的训练过程可以更聚焦于学习“对齐”从而降低对海量图文对数据的依赖。

1.2.2 视觉-语言预训练的突破：BLIP 系列的演进

在将 LLM 与视觉模态结合的道路上 Salesforce 研究院的 BLIP 系列工作扮演了至关重要的角色。它们通过一系列创新的架构和预训练任务高效地实现了视觉模态与语言模型的对齐。

BLIP(2022)：针对网络图文对数据中普遍存在的噪声问题 BLIP 提出了一种多模态混合编码器（Multimodal Mixture of Encoder-Decoder，MED）并设计了“字幕与过滤”（Captioning and Filtering，CapFilt）机制能够自动生成高质量的字幕并过滤掉噪声数据显著提升了预训练的效率和效果。

BLIP-2(2023)：BLIP-2 是这一阶段的标志性工作。



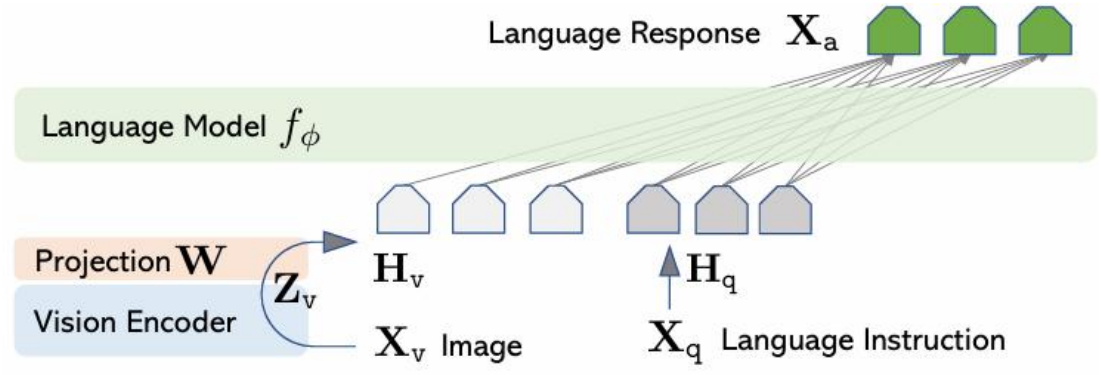
它提出了一个名为 Q-Former（Querying Transformer）的轻量级对齐模块。Q-Former 充当了冻结的视觉编码器（如 CLIP ViT）和冻结的 LLM 之间的“桥

梁”。它通过一小组可学习的查询向量从视觉编码器中提取与文本最相关的视觉特征然后将这些精炼后的特征输入给 LLM。这种“冻结主干只训练适配器”的设计极大地降低了训练成本使得在消费级硬件上训练强大的多模态模型成为可能。BLIP-2 的成功为后续几乎所有的视觉指令微调工作奠定了架构基础。

1.2.3 多模态指令微调的兴起：LLaVA 的开创性工作

BLIP-2 提供了高效的架构 LLaVA（Large Language and Vision Assistant）则开创了高效的训练方法。

2023 年 4 月发布的 LLaVA 首次将 LLM 领域的“指令微调”（Instruction Tuning）概念成功地引入多模态领域。



LLaVA 的洞察非常简洁：人类是通过语言指令与世界交互的那么也应该通过指令来教模型理解图像。其核心贡献在于构建了一个名为 LLaVA-Instruct-158K 的数据集。该数据集利用 GPT-4 强大的 API 将 COCO 数据集中已有的图像标注（如边界框、描述）转化为更丰富的多轮对话或问答形式。例如对于一张包含“一只猫在沙发上”的图像 GPT-4 可以生成诸如“这张图里有什么？”、“猫是什么颜色的？”、“它在做什么？”等一系列指令和回答。

LLaVA 的训练过程分为两个简单的阶段：

特征对齐阶段：使用简单的图文对数据训练一个线性投影层将 CLIP 视觉编码器的输出映射到 LLM 的词嵌入空间实现初步的模态对齐。

指令微调阶段：使用 LLaVA-Instruct-158K 数据集对整个模型（包括 LLM 部分）进行端到端的微调教会模型遵循指令进行多模态对话。

LLaVA 以其简洁的架构、高效的训练方法和令人印象深刻的对话能力迅速引爆了开源社区。它证明了即使使用相对较小规模的指令数据也能“解锁”LLM 在多模态场景下的强大能力。一时间基于 LLaVA 进行改进和扩展的工作层出不穷如 InstructBLIP、MiniGPT-4 等共同推动了多模态指令微调技术的成熟。

1.2.4 开源生态的繁荣

这一阶段的快速发展离不开开源社区的巨大推动力。特别是 Meta 在 2023 年发布的 LLaMA 系列模型其卓越的性能和开放的许可证为研究者们提供了一个强大的、可自由修改的 LLM 基座。这一时期几乎所有主流的开源多模态模型（包括 LLaVA、MiniGPT-4 等）都是基于 LLaMA 构建的。这形成了一个良性循环：强大的开源 LLM 基座降低了多模态研究的门槛而涌现出的优秀多模态模型又进一步丰富了 LLM 的生态吸引了更多开发者投身其中。

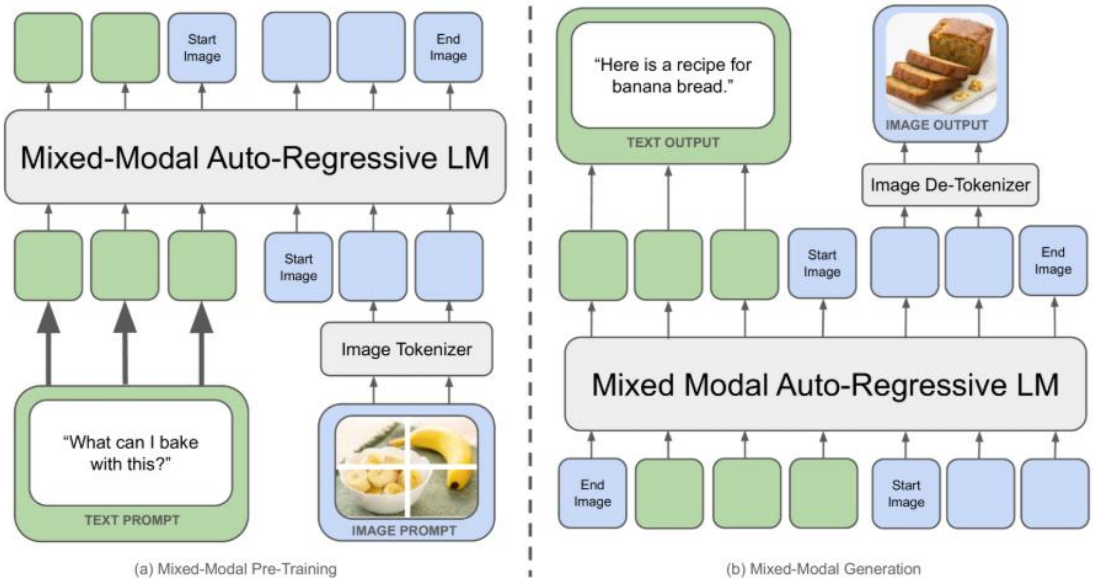
总结而言 2021 年至 2023 年是多模态技术由 LLM 驱动、发生范式革命的快速发展阶段。以 BLIP-2 的 Q-Former 架构和 LLaVA 的指令微调方法为两大支柱研究者们找到了一条将 LLM 的强大能力高效迁移到多模态场景的康庄大道。开源生态的繁荣则为这场技术革命提供了源源不断的动力。然而这一阶段的模型大多仍停留在“看懂”和“描述”的层面如何实现更高级的“生成”和“全模态”处理成为了下一阶段亟待解决的核心问题。

1.3 统一建模阶段（2024）：走向理解与生成的融合

随着多模态指令微调技术的成熟研究界的目光在 2024 年转向了一个更具挑战性的目标：在单一模型内统一多模态的理解与生成能力。此前理解任务（如 VQA）和生成任务（如文生图）通常由不同的模型负责。这一阶段的核心议题是打破两者之间的壁垒构建能够“既看懂又会画”的统一模型。这一时期的探索呈现出多样化的技术路径主要特征是早期融合架构的尝试、混合生成范式的出现以及全模态模型的初步探索。

1.3.1 理解与生成的统一尝试：Chameleon 与 VITRON

2024 年 5 月 Meta AI 发布的 Chameleon 是这一阶段的开创性工作。它大胆地提出了一种早期融合（Early-fusion）的思路。与此前主流的“晚期融合”（即先分别编码再连接）不同 Chameleon 在模型的最底层就将图像和文本转换为统一的离散 Token 序列然后将这些混合模态的 Token 序列直接输入到一个标准的、仅包含解码器（Decoder-only）的 LLM 中进行处理。这种设计的优势在于它最大程度地保留了 LLM 的原始架构理论上可以实现更深层次的模态交互。



Chameleon 通过一个新颖的图像分词器（Image Tokenizer）将图像转换为离散的 Token 其方式类似于文本的分词。这使得图像和文本可以在同一个序列中被 LLM 无差别地处理从而自然地统一了理解（模型根据图文上下文预测文本 Token）和生成（模型根据图文上下文预测图像 Token）任务。

几乎在同一时期来自昆仑万维、新加坡国立大学、新加坡南洋理工大学团队的研究者们发布了 VITRON 提出了另一种统一建模的思路。VITRON 的核心是统一的像素级视觉表示。

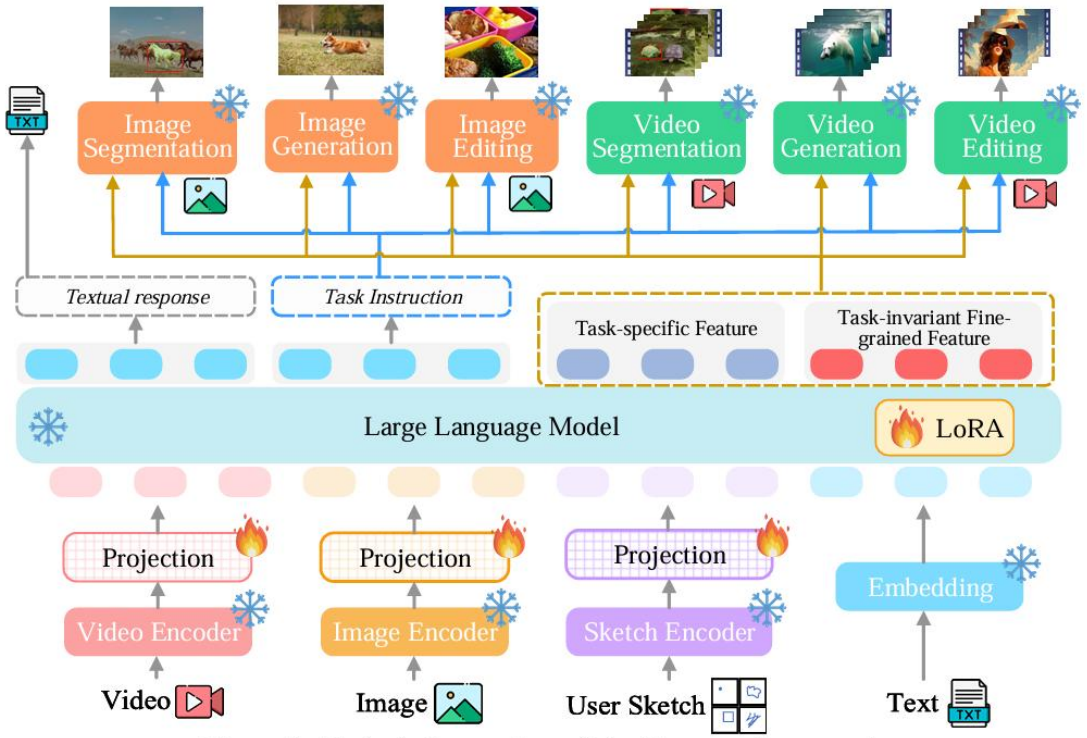


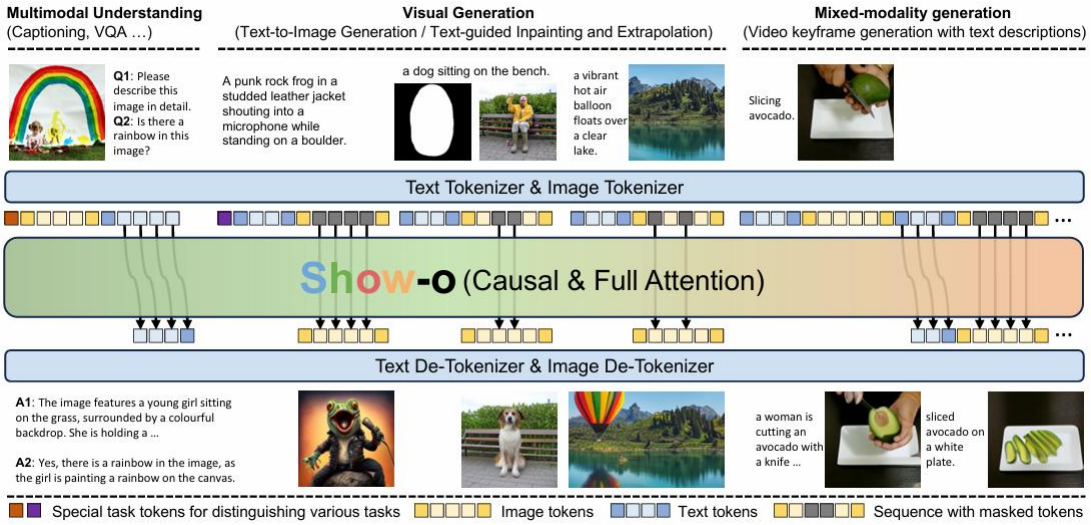
Figure 2: Technical overview of the VITRON framework.

它将各种视觉任务无论是高级语义理解（如 VQA）还是低级像素处理（如图像分割、编辑）都统一为“像素到像素”的生成任务。通过这种方式 VITRON 在单一模型内实现了对图像的理解、生成、分割和编辑四大核心能力展现了强大的通用视觉能力。

1.3.2 混合生成范式的出现：Show-o 的探索

在探索统一建模的过程中如何平衡生成质量、速度和多样性成为一个关键挑战。传统的自回归（Autoregressive，AR）模型虽然在文本生成上表现出色但在图像生成上存在速度慢、容易出现重复性伪影等问题。而扩散模型（Diffusion Models）虽然生成质量高但推理速度又是一大瓶颈。

为了解决这一问题 Show-o 提出了一种创新的混合生成范式。



它巧妙地将自回归模型和离散扩散模型结合在同一个统一的 Transformer 架构中。在生成图像时模型首先以自回归的方式快速生成一个全局的、低分辨率的草图（或称之为“计划”）然后再利用离散扩散模型对这个草图进行逐步的细化和高清化。这种“先规划后细化”的策略既利用了自回归模型在结构化预测上的优势又发挥了扩散模型在细节纹理生成上的长处实现了生成质量和效率的有效平衡。Show-o 的成功为后续的生成模型发展开辟了新的思路即不同生成范式并非相互排斥而是可以协同工作的。

1.3.3 全模态模型的萌芽

在视觉-语言统一建模取得进展的同时研究者们也开始将目光投向更广阔的“全模态”领域即在模型中进一步整合音频（Audio）和视频（Video）模态。这一时期的探索尚处于萌芽阶段主要通过在现有视觉-语言模型的基础上进行扩展。

例如一些工作开始尝试将音频频谱图（Spectrogram）作为一种特殊的“图像”输入给模型从而利用已有的视觉编码器来处理音频信号。对于视频则通常采用采样关键帧并将其作为多张图像输入的方式进行处理。这些早期的尝试虽然在架构上略显“朴素”但它们验证了将更多模态纳入统一 LLM 框架的可行性为 2025 年全模态模型的爆发积累了宝贵的经验。

1.3.4 工业界的激烈竞争：GPT-4V 与 Gemini

2024 年也是工业界巨头在多模态领域激烈竞争的一年。OpenAI 正式向公众发布了其强大的多模态模型 GPT-4V(ision)其在复杂的视觉推理、OCR 和少样本学习任务上展现出的惊人能力为整个领域树立了新的标杆。紧随其后 Google 也发布了其原生多模态模型 Gemini 系列特别是其旗舰版本 Gemini Ultra 在多个多模态基准测试中都表现出与 GPT-4V 相媲美甚至超越的性能。这两大闭源模型的发布一方面展示了多模态技术巨大的商业潜力另一方面也激发了开源社区更大的追赶热情形成了“闭源引领开源追赶”的竞争格局。

总结来说 2024 年是多模态技术从“分离”走向“统一”的关键一年。研究者们通过早期融合、混合生成范式等多种路径成功地在单一模型内实现了理解与生成的统一。同时对音频、视频等更多模态的整合也开始萌芽。工业界巨头的入场则进一步加速了技术的成熟和应用落地。然而这一阶段的统一模型在架构上仍有待完善生成质量和效率仍有提升空间这些都为 2025 年更深层次的技术变革创造了契机。

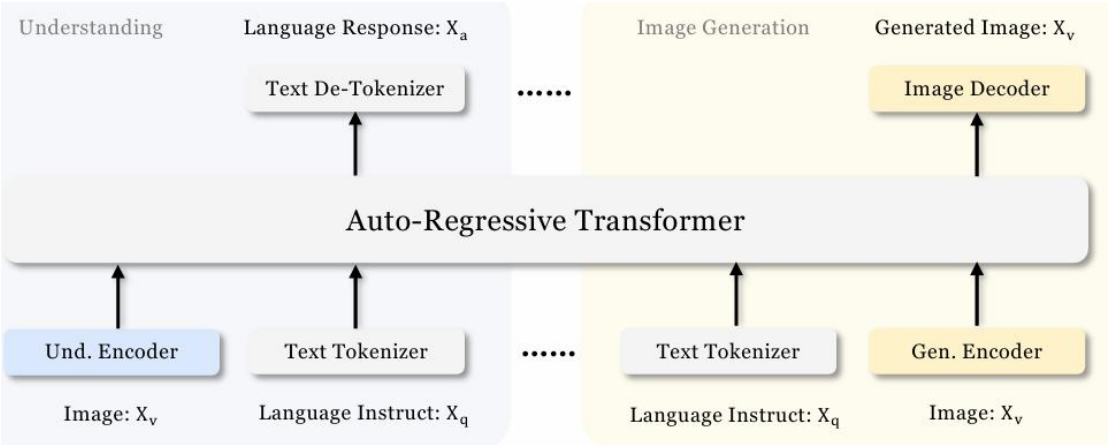
1.4 全模态爆发阶段（2025）：迈向“全能”与“实时”

2024 年是统一建模的探索年 2025 年则是多模态技术全面爆发、走向“全能”与“实时”的一年。在这一年里技术演进的核心驱动力从“如何在单一模型中统一理解与生成”转向“如何更高效、更高质量地统一所有主流模态并实现流畅的实时交互”。一系列具有里程碑意义的工作集中涌现它们在模型架构、生成范式和应用体验上都取得了质的飞跃。这一阶段的主要技术特征是解耦设计的成熟、流模型的崛起、原生全模态架构的实现以及交错生成能力的突破。

1.4.1 解耦设计的突破：Janus 的启示

2024 年末由 DeepSeek、香港大学、北大联合团队提出的 Janus 模型为解决早期融合架构（如 Chameleon）中存在的理解与生成能力难以兼顾的问题提供了全新的“解耦设计”（Decoupled Design）思路。Janus 的核心洞察是：视觉理解

任务需要的是全局、抽象的语义信息而视觉生成任务则更需要局部、精细的像素级细节。将两者耦合在同一个视觉编码路径中必然会导致性能上的妥协。



为此 Janus 创新性地设计了双路径视觉编码器：

理解路径（Understanding Path）：使用一个类似于 CLIP 的视觉编码器将图像编码为一组紧凑的、蕴含高级语义的特征向量专门用于 VQA、图像描述等理解任务。

生成路径（Generation Path）：使用一个 VQ-VAE 等图像分词器将图像转换为离散的、保留了丰富空间细节的视觉 Token 专门用于图像生成和编辑任务。

这两条路径的输出被同时输入到 LLM 中。LLM 可以根据当前任务的需要自主选择关注来自哪条路径的视觉信息。这种解耦设计使得模型的理解和生成能力可以得到独立的、更充分的优化从而在两大类任务上都取得了当时的最先进性能。Janus 的设计理念迅速被后续的许多工作所借鉴成为 2025 年高性能多模态模型的主流架构思想。

1.4.2 流模型的崛起：JanusFlow 与 NExT-OMNI

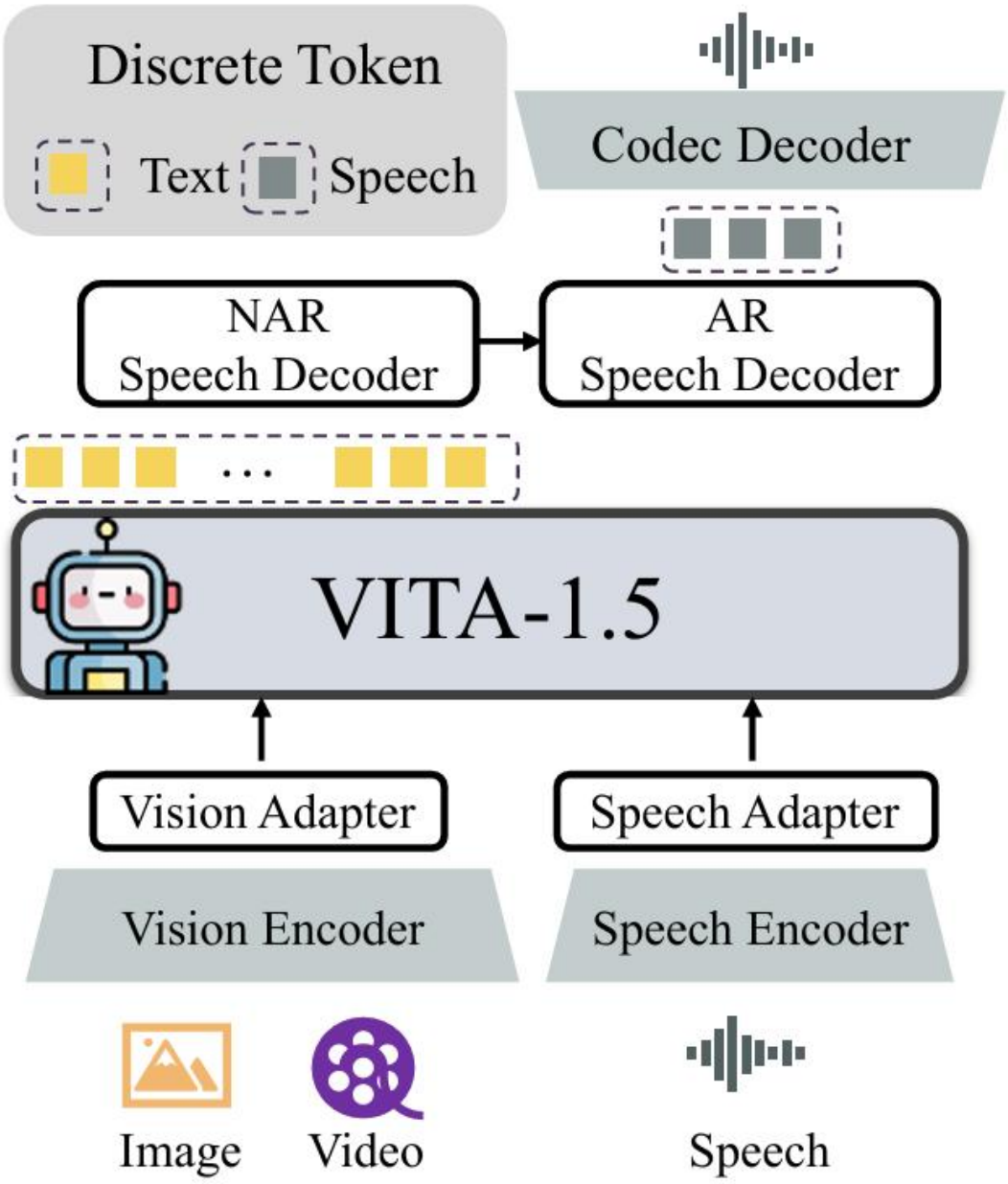
在生成范式上 2025 年见证了流模型（Flow Models）的全面崛起。相比于需要多次迭代去噪的扩散模型流模型旨在学习一个能够将简单的高斯噪声分布一步或数步映射到复杂数据分布的常微分方程（ODE）。

JanusFlow：作为 Janus 的后续工作 JanusFlow 将整流流（Rectified Flow）这一新兴的流模型技术引入多模态生成。它通过一种巧妙的方式协调了自回归（AR）模型和整流流。在生成时模型首先以 AR 方式生成一个“草稿”然后利用整流流进行一次或几次精炼即可得到高质量的图像。这种“AR + Flow”的混合范式在保持高质量的同时显著提升了推理速度通常只需 1-8 个采样步骤即可完成生成远快于扩散模型动辄数十上百步的采样。

NExT-OMNI：该工作则探索了另一种更前沿的流模型技术——离散流匹配（Discrete Flow Matching）。它将所有模态（文本、图像、音频、视频）都统一为离散的 Token 序列然后通过学习这些 Token 序列之间的流场变换实现了“任意模态到任意模态”（Any-to-Any）的生成。NExT-OMNI 是首个能够处理四种主流模态并实现任意转换的统一模型代表了全模态生成技术的前沿方向。

1.4.3 实时交互的实现：VITA-1.5 的突破

在提升用户体验方面实现流畅的实时交互是 2025 年的一个核心目标。VITA-1.5 在这方面取得了重大突破。

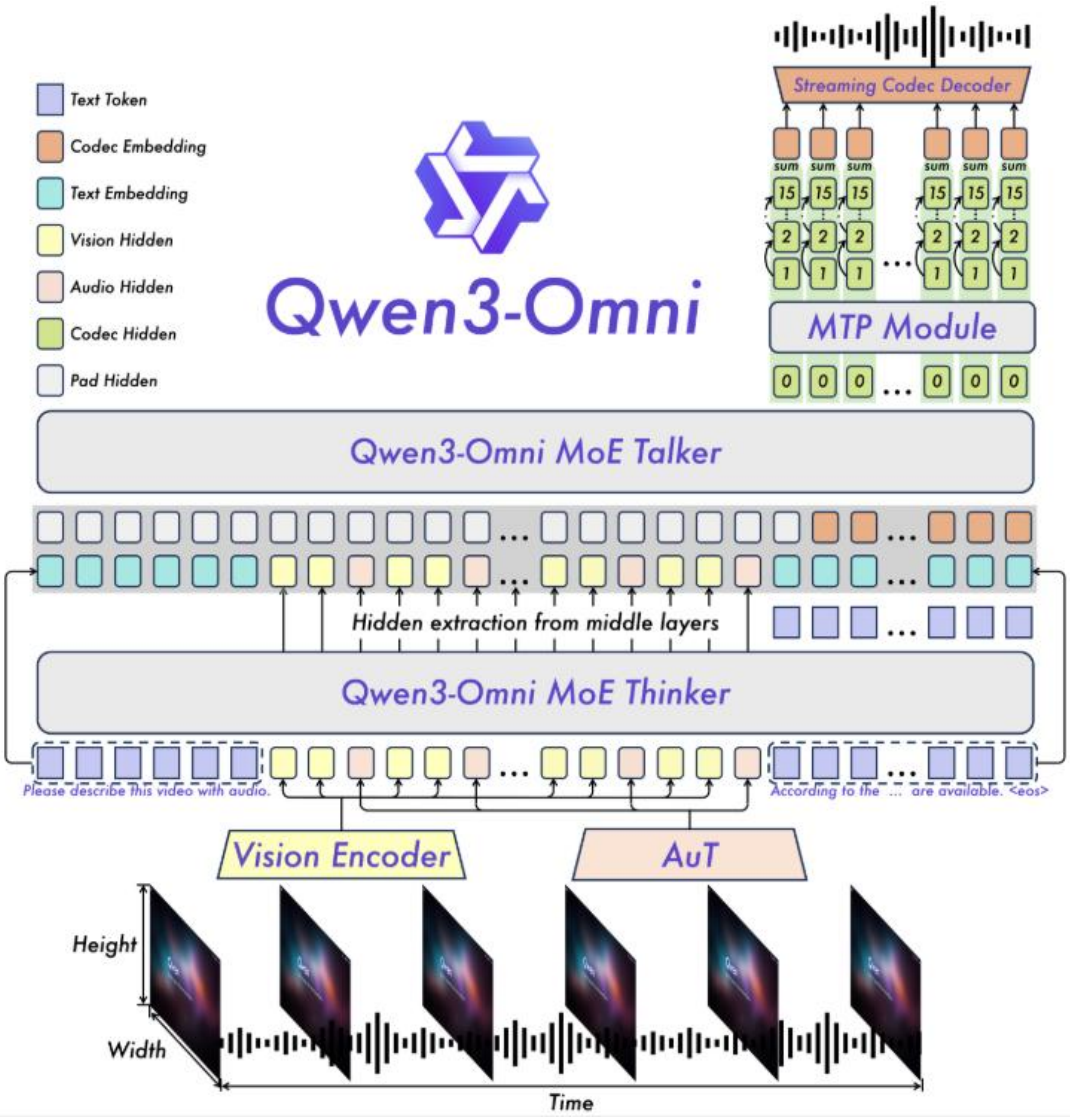


该模型通过精心设计的多阶段渐进式训练方法成功地将视觉和语音信息高

效地整合到一个 LLM 中。其最引人注目的成就是实现了接近 GPT-4o 水平的实时视觉-语音交互能力。用户可以向模型流式地输入语音指令同时展示摄像头捕捉到的实时画面模型能够即时地理解并作出语音回应延迟极低。这一突破极大地提升了多模态模型的实用性使其有望成为真正的个人智能助手。

1.4.4 原生全模态的成熟：Qwen3-Omni 的工业级实现

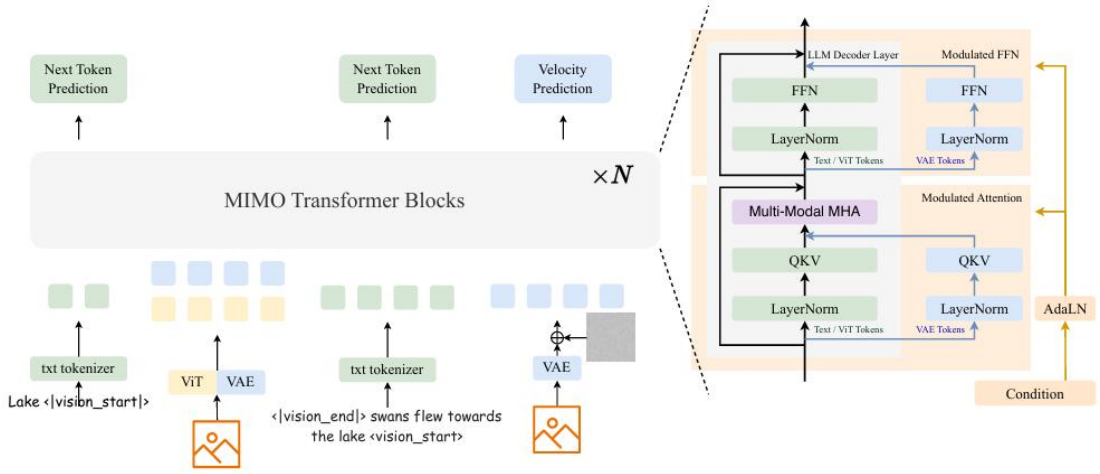
阿里巴巴在 2025 年 9 月发布的 Qwen3-Omni 代表了原生全模态（Natively Omni-Modal）技术在工业界的成熟落地。与那些依赖外部工具或模块拼接的模型不同 Qwen3-Omni 在一个统一的、端到端的架构内原生支持文本、图像、音频、视频四种模态的输入和输出。



据其技术报告显示 Qwen3-Omni 是首个在所有四种模态的主流基准测试上全面达到最先进性能的单一模型。这标志着全模态技术已经从理论走向实践具备了构建强大、可靠的商业应用的基础。

1.4.5 交错生成的创新：Mogao 的涌现能力

除了处理单一模态的输入输出 2025 年的另一个重要进展是交错多模态生成（Interleaved Multi-Modal Generation）。Mogao 在这方面做出了开创性贡献。



它能够生成包含文本、图像等多种模态交错出现的内容序列例如生成一篇图文并茂的博客文章。Mogao 通过一种基于因果建模的方法实现了这一能力并且展现出了令人惊奇的“涌现能力”如零样本的图像编辑和组合生成。这种能力使得 AI 不再仅仅是任务执行的工具更有可能成为内容创作的合作伙伴。

表 3：2025 年代表性多模态大语言模型技术特征

模型	核心技术创新	主要贡献
Janus	解耦双路径视觉编码	解决了统一模型中理解与生成的冲突
JanusFlow	AR + 整流流	实现了高质量、高效率的图像生成
NExT-OMNI	离散流匹配	实现了任意模态到任意模态的生成
VITA-1.5	多阶段渐进式训练	实现了接近 GPT-4o 的实时语音-视觉交互
Qwen3-Omni	原生全模态架构	首个在所有主流模态上达到 SOTA 的工业级模型
Mogao	因果交错生成	实现了图文并茂的复杂内容生成

1.4.6 多模态走进物理世界

具身智能是 2025 年最激动人心的应用方向。VLA“视觉-语言-动作”（Vision-Language-Action, VLA）模型通过统一视觉、语言和动作数据，使机器人能够跨任务、跨具身形态、跨环境泛化。OpenVLA 作为首个完全开源、可商用的 VLA 模型，在 2024 年 6 月发布后迅速成为机器人研究的基础模型。

1.4.7 国内代表性模型的崛起与特色

在 2025 年全球多模态技术浪潮中 ，中国科技力量同样取得了举世瞩目的成就，涌现出一批具有鲜明技术特色和强大实力的代表性模型。这些模型不仅在性能上追赶甚至超越了国际顶尖水平，更在架构设计和应用场景上展现了独特的创新思路。

深度求索 DeepSeek-OCR：DeepSeek-AI 另辟蹊径，从“光学压缩”这一独特视角切入，推出了 DeepSeek-OCR。该模型的核心创新在于将高分辨率的文档页面高效压缩为极少量的视觉 token，再由一个轻量级的 MoE 语言模型进行解码。这种“视觉作为压缩介质”的范式，在保证高精度 OCR 的同时，将处理长文档的 token 开销降低了 7-20 倍，为解决 LLM 的长上下文难题提供了一条极具潜力的技术路径。

通义千问 Qwen3-VL：阿里巴巴发布的 Qwen3-VL 系列 是其中的佼佼者。它不仅在传统的图文理解任务上表现出色，更通过增强的交错 MRoPE 和 DeepStack 等架构创新，实现了对长视频和复杂文档的深度理解。其原生支持 256K 的交错上下文处理能力，使其在长视频问答、文档分析等场景中具备显著优势，标志着国内模型在长上下文多模态处理能力上达到了新的高度。

文心 5.0 原生全模态：百度发布的文心 5.0 是中国首个真正意义上的“原生全模态”大模型。其核心理念是从训练伊始就将文本、图像、音频、视频等所有主流模态置于统一架构下进行联合建模，而非后期“拼接”。这种原生设计使得模型能够在底层形成跨模态的内在关联，从而在全模态的理解与生成任务上展现出更强的协同效应和一致性。其高达 2.4 万亿的参数规模和超稀疏激活的 MoE 架构，也代表了国内在大模型规模化探索上的最新成果。

智源 Emu3.5：北京智源人工智能研究院（BAAI）开源的 Emu3.5 则将多模态模型的能力边界从“理解世界”推向了“模拟世界”。作为一个大规模多模态世界模型，Emu3.5 不仅能处理交错的视文输入输出，更重要的是能够原生预测世界的下一个状态，展现出时空一致的世界探索和开放世界具身操作的能力。其提出的 DiDA（离散扩散适配）技术，在不牺牲性能的前提下将推理速度提升 20 倍，为世界模型的实际应用扫清了障碍。

这些模型的涌现，不仅丰富了全球多模态技术生态，也展示了中国在 AI 核心技术领域的深厚积累和创新活力。它们在不同技术路线上进行的探索，共同推动了多模态技术向着更高效、更通用、更智能的方向发展。

综上所述，2025 年是多模态大语言模型技术发展史上承前启后、全面爆发

的一年。在这一年里模型架构、生成范式和交互体验都迈上了新的台阶。解耦设计解决了核心的性能瓶颈流模型提供了更优的生成方案而实时交互和原生全模态的实现则宣告了“全能 AI 助手”时代的曙光。

第二章：核心技术架构与训练方法的进化

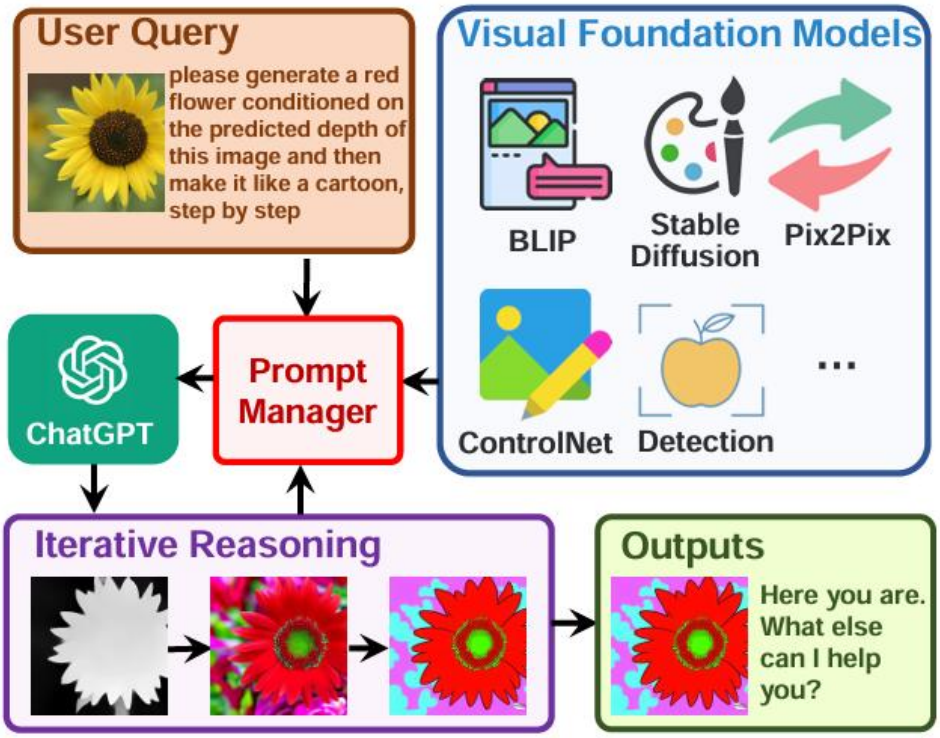
多模态大语言模型在 2025 年的爆发式增长其根源在于核心技术架构与训练方法的系统性进化。研究者们不再满足于简单地将不同模态的模块进行拼接而是从更根本的层面思考如何构建一个高效、统一且可扩展的智能系统。本章将深入剖析支撑这一轮技术浪潮的关键技术系统梳理其从早期探索到当前最前沿的演进脉络。我们将重点探讨五大核心主题：建模范式的演进、视觉编码器的设计、模态对齐机制、生成范式的革命以及训练方法的创新。通过对这些技术细节的深度解读本章旨在为读者揭示现代多模态大语言模型强大能力背后的“第一性原理”。

2.1 建模范式的演进：从外部集成到原生统一

多模态大语言模型的架构演进本质上是关于如何组织和协调不同模态信息处理流程的探索。回顾其发展历程我们可以清晰地看到一条从“外部专家集成”到“模块化联合建模”再到最终“端到端统一建模”的演进路径。这条路径反映了研究界对于模态融合深度和模型通用性不断提升的追求。

2.1.1 外部专家集成建模（Pre-2023）：LLM 作为“大脑”协调器

在多模态指令微调技术成熟之前一种简单而直接的思路是利用大型语言模型（LLM）作为中央“大脑”通过调用各种现成的、成熟的单模态“专家模型”（如视觉问答模型、图像生成模型）来协同完成复杂的多模态任务。这一范式的代表性工作是 Visual ChatGPT 和 HuggingGPT。



Visual ChatGPT 的核心机制是围绕一个基于聊天的界面将用户的多模态指令（如“帮我把这张图里的猫 P 掉然后生成一张狗的图片”）分解为多个子任务。然后 LLM 会根据其对任务的理解生成调用不同视觉基础模型（Visual Foundation Models，VFMs）的代码或指令并整合它们的输出来完成用户的最终请求。

表 4：外部专家集成建模范式分析

特征	描述
核心思想	LLM 作为任务规划器和协调器调度外部专家模型。
优势	1.快速实现：可充分利用现有的大量高质量专家模型无需从零训练。 2.能力全面：理论上可以集成任意数量和种类的专家快速获得新能力。 3.可解释性强：LLM 的决策过程（调用了哪个模型）相对透明。
局限性	1.深度融合缺失：模态间的交互是“任务级”而非“特征级”的信息在传递过程中损失严重。 2.效率低下：多次调用不同模型带来巨大的通信开销和推理延迟。 3.错误累积：整个任务链条的成功率受限于最弱的那个专家模型错误会逐级放大。

尽管存在明显局限外部专家集成范式在当时起到了重要的承上启下的作用。它首次展示了将 LLM 的通用推理能力应用于复杂多模态任务的巨大潜力为后续更深度的融合建模提供了宝贵的思路启示。然而其固有的浅层交互模式决定了它只能是一种过渡方案。

2.1.2 模块化联合建模（2023-2024）：寻找最佳“连接”方式

随着 LLaVA 等工作的成功研究界迅速转向探索如何在单一模型内部更好地连接视觉编码器和 LLM。这一“模块化联合建模”阶段的核心议题是设计一个高效的“适配器”或“连接器”在冻结大部分主干网络参数的同时实现高质量的模态对齐。根据连接方式的不同这一范式又可以细分为“提示中介建模”和“混合接口建模”。

A. 提示中介建模（Prompt-mediated Modeling）

这类方法的核心思想是将视觉信息转化为一种特殊的“软提示”（Soft Prompt）并将其插入到 LLM 的输入层。LLM 在处理文本时会同时“看到”这些代表了图像内容的软提示从而实现多模态理解。BLIP-2 的 Q-Former 就是这种范式的典型代表。Q-Former 通过一小组可学习的查询向量将视觉编码器输出的大量特征“压缩”成一小段固定长度的软提示既高效又有效。

B. 混合接口建模（Hybrid-interface Modeling）

随着研究的深入研究者们发现仅仅在输入层进行连接可能还不够。混合接口建模则尝试在 LLM 的更多层次上建立视觉与语言的连接。例如一些工作不仅在输入层注入视觉提示还在 LLM 的中间层或输出层引入额外的跨模态交互模块。2024 年的 VITRON 和 M2-Omni 就是这一方向的代表。它们通过在模型的不同深度设置多个“接口”让视觉信息能够更灵活、更深入地参与到 LLM 的思考过程中。

模块化联合建模是 2023 年至 2024 年的主流范式。它在成本和性能之间取得了很好的平衡催生了大量优秀的开源模型。然而这种“外挂”式的连接方式终究不是最理想的解决方案。适配器的设计本身就需要大量的经验和技巧而且冻结的主干网络也限制了模型进行更深层次的跨模态联合优化的可能性。

2.1.3 端到端统一建模（2024-2025）：迈向原生多模态

追求更彻底的融合与更优雅的架构是技术演进的必然方向。2024 年下半年至 2025 年端到端统一建模（End-to-End Unified Modeling）成为最前沿的探索方向。这一范式的核心目标是构建一个“原生”的多模态模型它不再区分视觉模块或语言模块而是从一开始就在一个统一的架构内处理所有模态的信息。

早期融合的尝试：如前文所述 Meta 的 Chameleon 是这一方向的先行者。它通过将所有模态都“Token 化”实现了在模型最底层的早期融合。这种设计的优点是架构统一、简洁但缺点也同样明显：不同模态的统计特性和信息密度差异巨大（例如图像通常比文本包含更多冗余信息）强行统一处理容易导致次优解。

解耦设计的成熟：为了解决早期融合的弊端 2025 年的 Janus 提出了精巧的“解耦设计”。它虽然仍在一个统一的 LLM 框架内但为视觉的理解和生成任务提供了不同的编码路径。这种“分而治之”再“统一处理”的思路被证明是当前实现高性能统一建模的有效路径。

原生全模态的实现：最终极的目标是构建一个无需任何特殊设计、能够自然处理所有模态的单一模型。2025 年 9 月发布的 Qwen3-Omni 在这条道路上迈出了关键一步。它通过在大规模、多样化的多模态数据上进行端到端的联合训练让一个标准的 Transformer 模型“自然地”学会了处理和关联不同模态信息。这种“大力出奇迹”的思路虽然对数据和算力提出了极高的要求但它所代表的“原生全模态”方向无疑是通往通用人工智能的最有希望的路径之一。

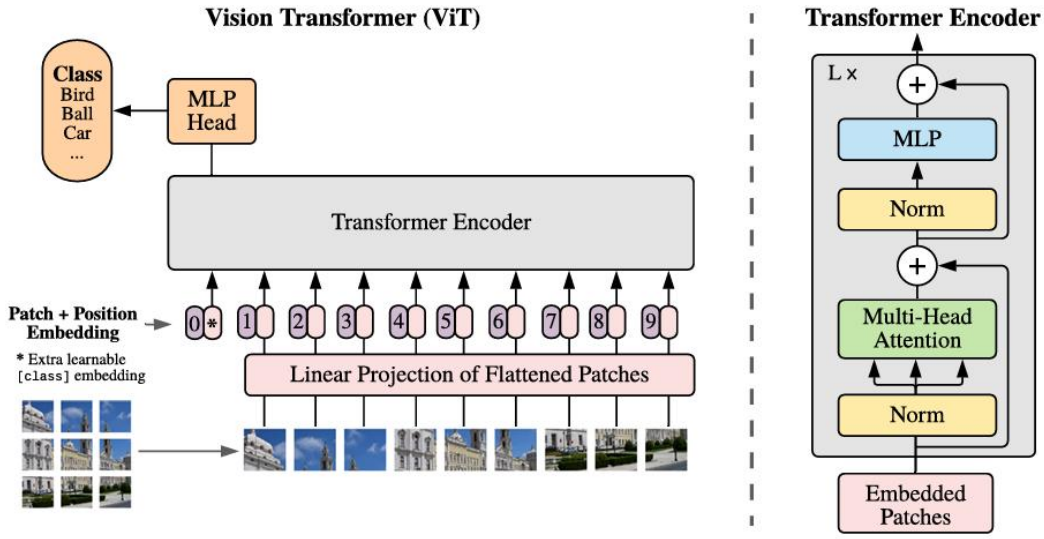
从外部集成到模块化连接再到最终的原生统一建模范式的演进清晰地展示了多模态技术不断追求更深层次融合、更广阔通用性和更优雅架构设计的坚定步伐。正是这一系列架构上的革新为 2025 年多模态大语言模型的全面爆发奠定了坚实的基础。

2.2 视觉编码器的设计：从单一特征到解耦表示

LLM 是多模态模型的“大脑”视觉编码器（Visual Encoder）是其“眼睛”。视觉编码器的核心任务是将输入的图像或视频帧转换为一系列 LLM 能够理解的特征向量。其设计的优劣直接决定了模型能够从视觉世界中汲取信息的深度和广度。其演进过程反映了研究界对于如何提取更丰富、更灵活、更适应下游任务的视觉表示的持续探索大致经历了从“单一通用特征”到“多分辨率协同”再到“任务导向解耦”的演进路径。

2.2.1 传统视觉编码器：ViT 与 CLIP 的奠基

现代多模态大语言模型普遍采用基于 Vision Transformer(ViT)的架构作为视觉编码器的骨干。ViT 的革命性在于它将 Transformer 架构成功地从 NLP 领域迁移到了 CV 领域。



它将图像分割成一系列固定大小的图块（Patches）并将这些图块线性嵌入后像处理单词一样输入到 Transformer 编码器中。这种设计使得模型能够捕捉图像中的全局依赖关系相比于传统的卷积神经网络（CNNs）具有更好的可扩展性。

在 ViT 的基础上 CLIP 的视觉编码器（通常也是 ViT 架构）通过在海量的图文对数据上进行对比学习获得了强大的语义表征能力。因此冻结的 CLIP ViT 成为了 2023 年以来绝大多数模块化多模态模型（如 LLaVA、BLIP-2）的首选视觉编码器。使用预训练的 CLIP ViT 具有两大优势：

强大的语义特征：其输出的特征与自然语言在语义上深度对齐极大地降低了后续模态对齐的难度。

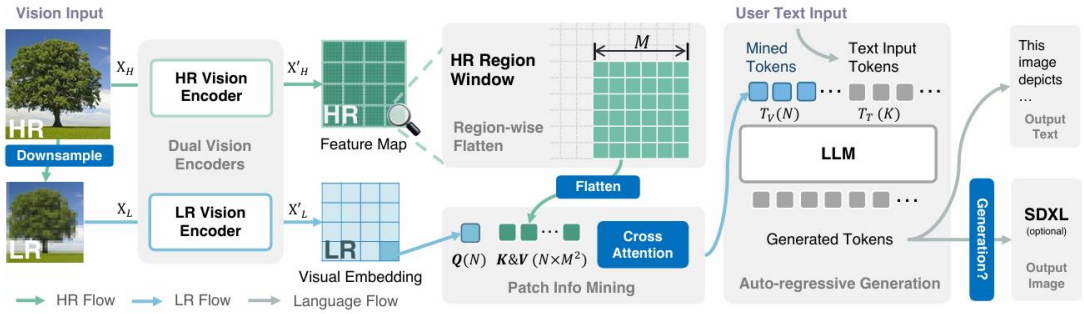
训练效率高：由于其参数被冻结在多模态训练阶段无需更新显著节省了计算资源。

然而标准的 CLIP ViT 也存在其固有的局限性：它通常在较低的分辨率（如 224x224 或 336x336）下进行预训练这导致其在处理需要高分辨率细节的任务（如 OCR、细粒度识别）时表现不佳。

2.2.2 高分辨率处理：应对细节挑战

为了解决低分辨率带来的信息损失问题研究者们探索了多种策略来让模型“看清”图像的细节。一个直接的方法是直接提高 ViT 的输入分辨率但这会导致图块数量的平方级增长带来巨大的计算和内存开销。因此更精巧的设计应运而生。

以 Mini-Gemini 为代表的工作提出了一种双编码器（Dual Encoder）或多分辨率（Multi-resolution）的策略。



它在保留原有低分辨率 CLIP ViT 的同时额外增加了一个处理高分辨率图像块的编码器。在处理图像时模型首先用低分辨率编码器获取全局上下文信息然后可以根据需要将图像中的特定区域（或整个图像）以更高的分辨率输入给第二个编码器以获取局部细节。这种“全局概览+局部详查”的机制在不显著增加整体计算成本的前提下有效地提升了模型处理高分辨率细节的能力。

2.2.3 解耦视觉编码：Janus 的革命性设计

2024 年底最重要的架构创新之一便是 Janus 提出的解耦视觉编码（Decoupled Visual Encoding）。这一设计的核心洞察在于不同的下游任务对视觉特征的需求是不同的。

理解任务（如 VQA）需要的是抽象的、高级的语义信息。

生成任务（如文生图）需要的是具体的、低级的像素细节。

传统的单一编码器试图用一套特征来满足两种截然不同的需求这本质上是一种“妥协”。而 Janus 则通过其双路径设计彻底解决了这一矛盾：

理解路径：沿用 CLIP ViT 输出紧凑的、富含语义的特征向量专门服务于理解任务。

生成路径：采用一个类似于 VQ-GAN 的视觉分词器（Visual Tokenizer）将图像无损地重建为离散的视觉 Token 序列。这个序列保留了图像的所有像素级信息专门服务于生成任务。

这种解耦设计带来了巨大的优势：两条路径可以被独立优化使得模型的理解和生成能力不再相互掣肘从而在各自的领域都能达到更高的性能。这一思想迅速成为 2025 年高性能统一模型的设计标准标志着视觉编码器设计从“一刀切”走向了“因材施教”的精细化阶段。

2.2.4 像素级编码：VITRON 的统一表示

与 Janus 的“解耦”思想相对应另一条技术路线则追求极致的“统一”。VITRON 提出的像素级编码（Pixel-level Encoding）就是其中的代表。它尝试将所

有与视觉相关的任务无论是高级理解还是低级处理都统一到像素级别的表示上。

VITRON 的设计使得模型不仅能输出描述图像的文本（理解）还能直接输出修改后的图像像素（编辑）或分割掩码（分割）。这种端到端的像素级生成能力使得模型在图像编辑、修复和分割等任务上展现出传统 MLLM 难以企及的精确控制力。虽然这种设计的计算成本相对较高但它为构建全能的“Photoshop AI”提供了可能代表了视觉编码器在通用性和任务广度上的一个重要探索方向。

表 5：不同视觉编码器设计范式对比

范式	代表模型	核心思想	优势	局限性
单一通用编码	LLaVA，BLIP-2	使用冻结的 CLIP ViT 提取通用语义特征。	简单高效语义对齐良好。	分辨率低细节信息丢失。
多分辨率编码	Mini-Gemini	结合低分辨率全局编码器和高分辨率局部编码器。	兼顾全局上下文和局部细节。	架构更复杂计算成本增加。
解释编码	Janus	为理解和生成任务提供独立的编码路径。	理解和生成能力可独立优化性能更高。	模型参数量更大。
像素级编码	VITRON	将所有视觉任务统一为像素到像素的生成。	精确控制像素级任务（编辑、分割）。	计算成本高对高级语义任务不一定最优。

从单一的 CLIP ViT 到应对高分辨率挑战的多编码器再到为不同任务量身定制的解耦路径以及追求极致统一的像素级表示视觉编码器的演进之路清晰地反映了多模态模型对视觉信息日益增长的精细化、多样化和专业化需求。一个设计精良的视觉编码器是模型通往更强大、更通用多模态智能的坚实基石。

2.3 语言模型骨干网络：多模态智能的“思考中枢”

在多模态大语言模型（MLLM）的架构中大型语言模型（LLM）扮演着无可替代的“思考中枢”角色。它负责接收来自不同模态编码器的信息并进行高级的语义理解、逻辑推理、指令遵循和内容生成。LLM 骨干网络的性能直接决定了整个多模态系统的智能上限。因此选择一个强大且合适的 LLM 骨干并对其进行有效的多模态适配是构建高性能 MLLM 的关键步骤。

2.3.1 主流 LLM 骨干的选择：开源社区的赋能

2023 年以来开源 LLM 的蓬勃发展极大地推动了多模态研究的进程。研究者们得以站在巨人的肩膀上将精力更聚焦于多模态的特定挑战而非从零开始训练一个庞大的语言模型。当前主流的开源 MLLM 主要围绕以下几个系列的 LLM 进行构建：

LLaMA 系列（Meta AI）：从 LLaMA 到 LLaMA 2 再到 LLaMA 3，Meta 发布的系列模型以其卓越的性能、庞大的社区支持和相对开放的许可证成为了构建 MLLM 的“黄金标准”。绝大多数有影响力的开源 MLLM 包括 LLaVA、Janus、VITA-1.5 等都采用了 LLaMA 系列作为其语言骨干。这形成了一个强大的生态系统相关的研究和改进可以方便地相互借鉴。

Phi 系列（Microsoft）：Phi 系列特别是 Phi-3 以其“小模型、大能力”（Small Language Models，SLMs）的特点受到了广泛关注。它们通过在高质量、经过精心筛选的“教科书级别”数据上进行训练在相对较小的参数规模（如 3.8B）下达到了与更大模型相媲美的性能。对于追求在端侧设备或资源受限环境中部署的多模态应用而言 Phi 系列是一个极具吸引力的选择。

DeepSeek 系列（DeepSeek AI）：DeepSeek-LLM 和 DeepSeek-Coder 等模型以其强大的代码生成和数学推理能力而闻名。对于那些需要处理包含大量代码、公式或需要严谨逻辑推理的专业领域多模态任务（如科学文献理解、UI 设计自动化）而言采用 DeepSeek 系列作为骨干网络可能带来独特的优势。

Qwen 系列（Alibaba）：从 Qwen 到 Qwen2 再到 2025 年的 Qwen3 阿里巴巴的 Qwen 系列模型以其强大的多语言能力和持续的全模态扩展而著称。特别是 Qwen3-Omni 其语言骨干从设计之初就考虑了与多种模态的深度协同是原生全模态模型的重要代表。

2.3.2 参数规模的影响：越大越好但需权衡

与纯文本 LLM 类似 MLLM 的性能也与其语言骨干的参数规模显著相关。通常来说更大的模型（如 70B 级别）在处理复杂指令、进行深度推理和生成高质量内容方面要优于较小的模型（如 7B 或 13B 级别）。许多前沿的 MLLM 研究为了追求更高的性能上限都会选择最大规模的开源 LLM 作为实验基础。

然而“越大越好”并非没有代价。巨大的模型尺寸带来了高昂的训练和推理成本限制了其在现实世界中的广泛部署。因此如何在模型性能和部署效率之间做出权衡是所有 MLLM 研究者和开发者必须面对的问题。这也催生了模型量化（Quantization）、知识蒸馏（Knowledge Distillation）等一系列模型压缩和加速技术在多模态领域的应用。

2.3.3 架构的微调与适配

尽管现代 MLLM 倾向于“冻结”LLM 骨干的大部分参数以节省训练成本但

为了更好地整合多模态信息一些微小的架构调整仍然是必要的。

词嵌入空间的扩展：LLM 原始的词嵌入空间只包含文本 Token。为了让 LLM 能够“看到”视觉 Token 需要将视觉编码器输出的特征向量投影到与文本 Token 相同的维度并将其视为一种特殊的“视觉词汇”添加到 LLM 的输入序列中。

注意力机制的调整：在处理包含视觉 Token 的混合序列时 LLM 的自注意力机制（Self-Attention）能够自然地学习文本与视觉、视觉与视觉之间的复杂关联。在某些设计中研究者还会引入额外的跨注意力模块以更显式地加强模态间的交互。

位置编码的扩展：对于需要处理多张图像或视频帧的输入如何设计有效的位置编码以告知 LLM 不同图像或帧的空间/时间关系是一个重要且开放的研究问题。例如需要让模型理解一张图片在另一张的“左边”或者一个视频帧在另一个的“之前”。

总而言之 LLM 作为多模态系统的“思考中枢”其选择和适配是整个系统设计的重中之重。开源 LLM 的繁荣为 MLLM 的快速发展提供了坚实的基础而如何在不同规模、不同特性的 LLM 骨干之间做出选择并对其进行精巧的多模态适配将持续是推动该领域向前发展的关键技术环节。

2.4 模态对齐机制：搭建跨模态沟通的桥梁

视觉编码器和 LLM 骨干是两个独立的“王国”模态对齐机制（Modality Alignment Mechanism）是连接这两个王国的“桥梁”。它的核心任务是将来自不同模态（如视觉）的特征信息高效、准确地转换为 LLM 能够理解和处理的“语言”。对齐机制设计的优劣直接关系到信息在跨模态传递过程中的保真度和有效性。其演进过程是从简单的线性投影到精巧的查询压缩再到更具适应性的专家混合网络体现了对更高对齐质量和效率的不懈追求。

2.4.1 线性投影层：最简单的连接

在早期的探索中最简单直接的对齐方法是使用一个或多个线性投影层（Linear Projection Layer）。其作用是将视觉编码器输出的特征向量（例如 CLIP ViT 输出的特征维度通常是 1024 或 768）通过一个可训练的权重矩阵直接映射到与 LLM 词嵌入向量相同的维度（例如 LLaMA 的词嵌入维度是 4096）。

这种方法的优点是极其简单、计算成本低。在 LLaVA 的第一阶段预训练中就是通过训练一个简单的线性投影层实现了视觉特征与语言模型初步的语义对

齐。然而其缺点也同样明显：

信息瓶颈：一个简单的线性变换可能难以捕捉视觉特征与语言语义之间复杂的非线性关系。

长度不匹配：视觉编码器通常会为一张图片生成数百个特征向量（每个对应一个图像块）而 LLM 在处理长序列时会面临巨大的计算压力。如何将这些大量的视觉特征有效地呈现给 LLM 是一个挑战。

2.4.2 Q-Former 架构：高效的查询压缩

为了解决上述挑战 BLIP-2 提出了革命性的 Q-Former（Querying Transformer）架构。Q-Former 可以被看作是一个精巧的“信息压缩器”和“转换器”它在冻结的视觉编码器和冻结的 LLM 之间扮演了关键的桥梁角色。

Q-Former 的核心机制是引入了一小组（例如 32 个）可学习的查询向量（Learnable Queries）。这些查询向量通过一个专属的 Transformer 网络与来自视觉编码器的海量图像块特征进行交互（通过交叉注意力机制）。在这个过程中查询向量被训练来“主动地”从图像中提取与特定文本描述最相关的视觉信息。最终这些“吸收”了关键视觉信息的查询向量其输出的特征就被作为软提示输入给 LLM。

Q-Former 的优势是多方面的：

高效压缩：无论原始图像编码器输出多少特征 Q-Former 总能将其压缩为一小段固定长度（如 32 个）的序列极大地减轻了后续 LLM 的处理负担。

强大的对齐能力：通过专门的预训练任务（如图像-文本对比学习、图像-文本匹配、图像引导的文本生成）Q-Former 能够学习到高质量的、蕴含丰富语义的视觉表示。

灵活性与可扩展性：Q-Former 的设计是模块化的可以方便地接入任何视觉编码器和 LLM 具有很强的通用性。

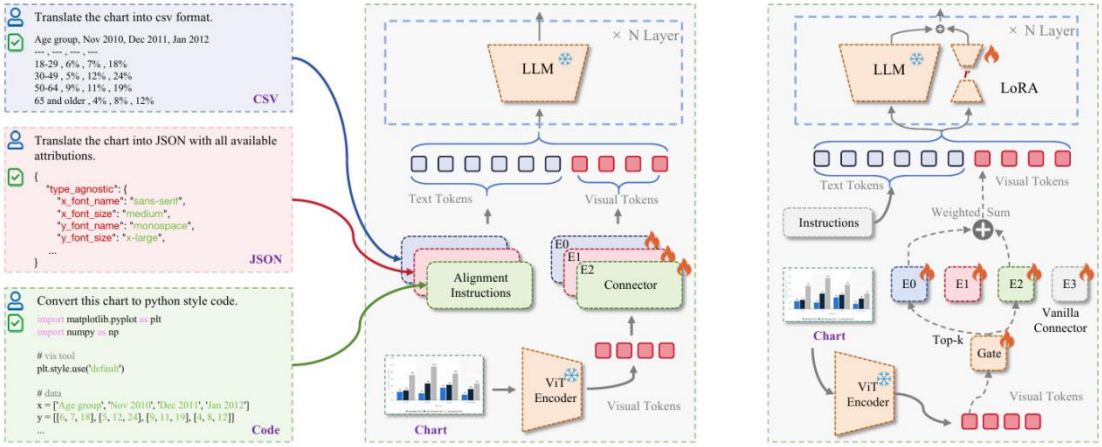
Q-Former 的成功使得“冻结主干、只训练适配器”的训练范式成为可能并被后续大量的 MLLM 工作（如 InstructBLIP）所采纳成为模块化联合建模时代最核心的对齐技术。

2.4.3 MoE 连接器：专家网络实现自适应对齐

进入 2025 年随着模型需要处理的模态和任务越来越多样化研究者们发现单一的、通用的对齐模块可能已不足以应对所有情况。例如理解一张照片所需的视

觉特征和理解一张科学图表所需的视觉特征其侧重点可能完全不同。为了实现更具适应性的对齐专家混合网络（Mixture-of-Experts，MoE）的思想被引入到对齐模块的设计中。

ICLR 2025 的 Oral 论文 ChartMoE 是这一方向的杰出代表。



(a) Alignment Instructions (b) Alignment Pretraining (c) Supervised Finetuning

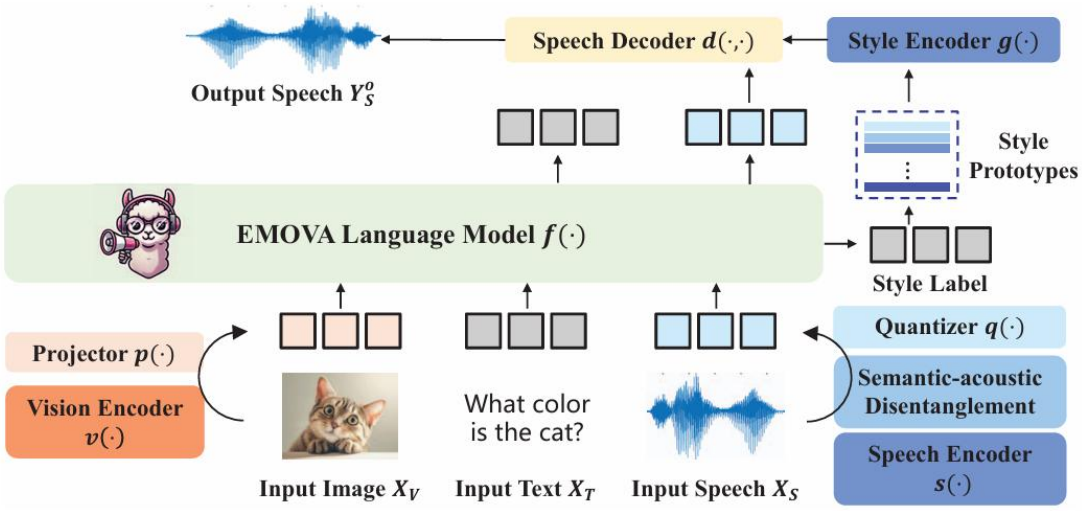
在处理复杂的图表理解任务时 ChartMoE 设计了一个 MoE 连接器。该连接器包含多个并行的“专家网络”（每个专家通常是一个小型的 MLP 或 Transformer）以及一个“门控网络”（Gating Network）。当视觉特征输入时门控网络会根据特征的类型和内容动态地决定将这些特征分配给哪个或哪些专家进行处理并赋予不同的权重。例如一些专家可能擅长处理图表中的文本和数字另一些专家则可能擅长理解图表的结构和布局。

通过这种方式 MoE 连接器实现了自适应的、任务导向的对齐。它能够根据输入数据的特点动态地组合不同专家的能力从而比单一的连接器获得更精细、更准确的视觉表示。ChartMoE 凭借其创新的 MoE 连接器在多个图表理解基准上取得了超过 16%的巨幅性能提升充分证明了这种自适应对齐机制的优越性。

2.4.4 全模态对齐的挑战与发现

当模型需要处理的模态从视觉-语言扩展到音频、视频等更多模态时对齐机制面临着新的挑战。一个核心问题是：在统一的语义空间中对齐多种模态是否会因为“模态冲突”而损害各自的性能？

传统的观点认为不同模态的最优表示空间可能存在差异强行将它们对齐可能会导致性能下降。然而 2025 年的 EMOVA 模型通过其在视觉-语言-语音三模态上的实验得出了一个令人振奋的发现：精心设计的全模态对齐不仅不会损害性能反而能够产生“增强效应”（Enhancement Effect）。



例如在对齐了语音模态后模型在纯视觉-语言任务上的性能也得到了提升。这可能是因为不同模态的信息可以相互印证、相互补充从而帮助模型学习到更鲁棒、更抽象的通用语义表示。

这一发现为构建更强大的全模态模型注入了强心剂。它表明追求更广泛的模态覆盖不仅是为了扩展模型的功能其本身就是一条通往更深层次智能的有效路径。如何设计能够最大化这种“增强效应”的全模态对齐机制将是未来研究的一个重要方向。

从简单的线性投影到高效的 Q-Former 再到自适应的 MoE 连接器模态对齐机制的演进之路是多模态模型不断追求更高效、更精准、更智能的跨模态“沟通”方式的缩影。正是这些日益精巧的“桥梁”让不同模态的信息得以在 LLM 的“思考中枢”里顺畅地流动、碰撞与融合最终涌现出强大的多模态智能。

2.5 生成范式的革命：追求质量、速度与统一

理解是智能的输入生成就是智能的输出。多模态大语言模型不仅要“看懂”世界更要“创造”世界。生成范式的演进是 2024 年至 2025 年多模态技术发展最为活跃、最具突破性的领域之一。这场革命的核心是在追求更高生成质量的同时不断提升生成速度并最终将不同的生成模型统一到一个优雅的框架之下。其演进路径主要围绕着自回归（AR）、扩散（Diffusion）和流（Flow）这三大主流范式展开并最终走向了高效的混合生成。

2.5.1 传统生成范式：自回归与扩散的权衡

在多模态生成领域长期以来主要由两种范式主导：

自回归模型（Autoregressive Models，AR）：这类模型将生成过程视为一

个序列决策过程。在图像生成中它们通常先将图像“展平”为一个一维的像素或Token 序列然后像生成文本一样逐个像素或逐个 Token 地进行预测和生成。其优点是架构与语言模型天然统一可以直接利用 LLM 进行生成。但缺点也十分突出：

速度慢：串行的生成方式导致推理速度与图像大小成正比难以用于实时应用。

误差累积：生成过程中的一个错误可能会被后续步骤不断放大导致生成的图像在全局结构上出现问题。

单向性：只能从左到右、从上到下地生成缺乏灵活性。

扩散模型（Diffusion Models）：自 2020 年 DDPM 提出以来扩散模型以其卓越的生成质量和多样性迅速成为高质量图像生成的主流。它通过一个“加噪-去噪”的过程来学习数据的分布。在生成时模型从一个纯噪声图像开始通过数十上百次的迭代去噪逐步恢复出清晰的图像。其优点是生成质量极高能够产生逼真的细节和纹理。但其核心痛点在于：

推理速度极慢：多次迭代去噪的过程非常耗时严重限制了其应用场景。

与 LLM 架构不兼容：扩散模型（通常基于 U-Net 架构）与 LLM（基于 Transformer 架构）在结构上存在差异难以实现完美的统一。

在 2024 年之前研究者们通常需要在这两种范式之间做出艰难的权衡：要么选择与 LLM 架构统一但速度慢、质量稍逊的自回归模型要么选择质量高但速度极慢且难以统一的扩散模型。

2.5.2 混合生成范式的探索：Show-o 的启示

为了打破上述僵局 2024 年的 Show-o 提出了一种创新的混合生成范式。它巧妙地在同一个 Transformer 架构内将自回归与离散扩散（Discrete Diffusion）结合起来。其生成过程分为两个阶段：

全局规划（AR）：模型首先以自回归的方式快速生成一个低分辨率的、包含图像全局结构和布局的“草图”或“计划”。

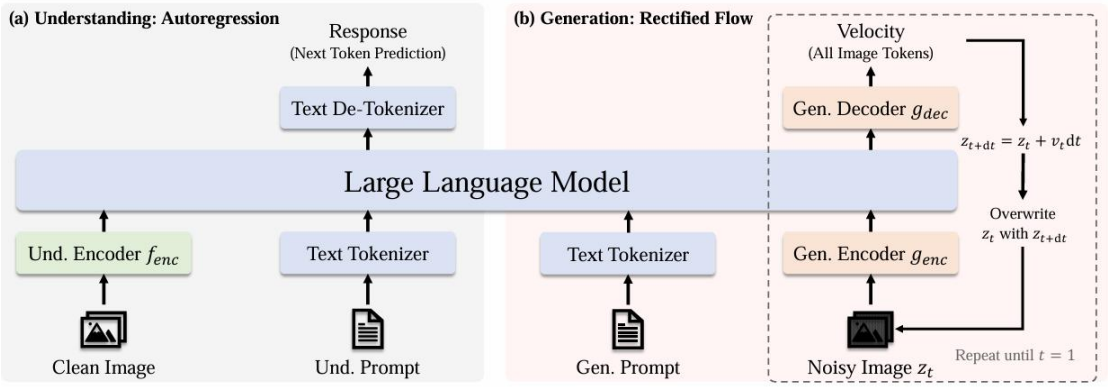
局部细化（Diffusion）：然后模型将这个草图作为条件利用并行的离散扩散过程对草图进行细节的填充和高清化。

这种“先规划后细化”的策略既发挥了自回归模型在把握全局结构上的优势又利用了扩散模型在生成高质量细节上的长处实现了质量与速度的有效平衡。Show-o 的探索证明了不同生成范式并非不可调和而是可以协同工作为后续的生成模型发展开辟了新的道路。

2.5.3 流模型的崛起：JanusFlow 与 NExT-OMNI 的突破

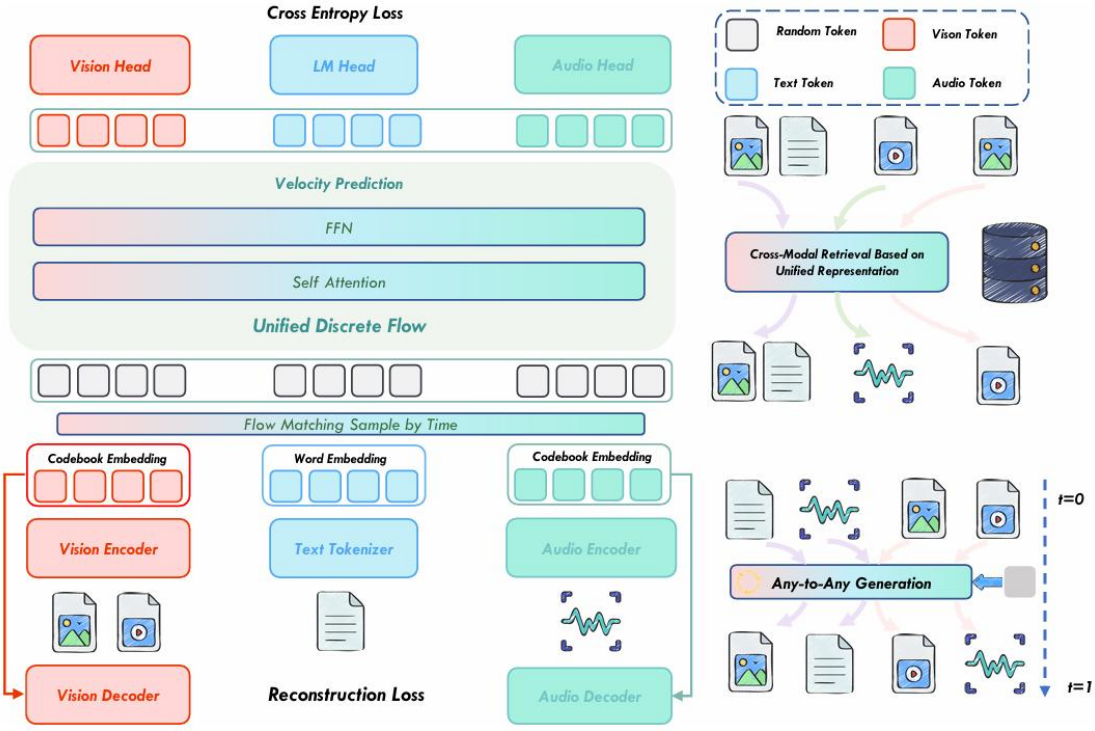
2025 年生成范式革命的真正主角是流模型（Flow Models）。流模型旨在学习一个从简单先验分布（如高斯噪声）到复杂数据分布的直接映射函数。相比于扩散模型的多步迭代理想的流模型仅需一步或极少步就能完成高质量的生成从而在理论上拥有巨大的速度优势。

整流流（Rectified Flow）：JanusFlow 是将整流流成功应用于大规模多模态生成的开创性工作。整流流通过一种特殊的“重参数化”技巧将复杂的非线性变换路径“拉直”使得模型更容易学习。



JanusFlow 提出的“AR + Flow”混合范式在生成时先用 AR 模型快速生成一个高质量的“起点”然后通过 1-8 步整流流采样就能达到与数百步扩散模型相媲美的生成质量。这在保持高质量的同时将生成速度提升了数十甚至上百倍。

离散流匹配（Discrete Flow Matching）：NExT-OMNI 则探索了更前沿的离散流匹配技术。



它将所有模态（文本、图像、音频、视频）都统一为离散的 Token 序列然后直接学习这些离散 Token 分布之间的变换流场。这使得模型能够实现“任意模态到任意模态”的生成例如文生图、图生文、文生视频、视频生音频等都在一个统一的流模型框架下得以实现。NExT-OMNI 的成功标志着流模型在处理离散数据和实现全模态统一生成方面的巨大潜力。

表 6：主流生成范式对比（2025 视角）

范式	代表模型	生成质量	生成速度	统一性	核心优势/劣势
自回归 (AR)	Chameleon	中-高	慢	高	优势：与 LLM 架构天然统一。劣势：速度慢误差累积。
扩散 (Diffusion)	Stable Diffusion	极高	极慢	低	优势：生成质量和多样性极佳。劣势：速度是主要瓶颈与 LLM 不兼容。
混合 (AR+Diffusion)	Show-o	高	中	中-高	优势：结合 AR 的全局规划和 Diffusion 的局部细节。劣势：架构相对复杂。
流(Flow)	JanusFlow, NExT-OMNI	极高	极快	高	优势：兼具极高质量和极快速度理论完备。劣势：训练相对不稳定技术较新。

从自回归与扩散的艰难权衡到混合范式的巧妙融合再到流模型的全面崛起生成范式的革命性演进是 2025 年多模态技术最激动人心的篇章。流模型以其兼具高质量、高速度和高统一性的巨大潜力正迅速成为下一代生成模型的标准范式为实现更强大、更实时的多模态内容创作奠定了坚实的技术基础。

2.6 训练方法的创新：追求数据效率与能力对齐

拥有了先进的架构如何高效地“教导”这些庞大的模型是决定其最终能力的关键。训练方法（Training Methods）的创新与架构设计本身同等重要。在多模态大语言模型的演进过程中训练方法的核心目标始终围绕着两个方面：提升数据效率（如何用更少、更易获取的数据达到更好的效果）和实现能力对齐（如何让模型真正理解并遵循人类的意图）。其发展脉络是从大规模的无监督预训练到有监督的指令微调再到更精细化的多阶段渐进式训练。

2.6.1 两阶段训练范式：预训练 + 指令微调

自 LLaVA 以来一种经典的两阶段训练范式成为了开源社区的主流：

第一阶段：视觉-语言预训练（Vision-Language Pre-training）

目标：实现视觉特征与语言模型在语义层面的初步对齐。

数据：通常使用大规模、相对原始的图文对数据如 CC3M、LAION 等。这些数据包含数十亿的图文对但质量参差不齐。

方法：训练一个连接模块（如线性投影层或 Q-Former）使其能够将视觉编码器输出的特征映射到 LLM 的输入空间。训练任务通常是简单的图像-文本匹配或图像引导的文本生成。

关键：这一阶段只训练连接模块 LLM 和视觉编码器的骨干参数通常被冻结因此训练成本相对较低。

第二阶段：多模态指令微调（Multimodal Instruction Tuning）

目标：教会模型遵循人类的指令进行复杂的、对话式的多模态任务。

数据：使用高质量、经过精心构建的指令遵循数据集。这些数据集通常规模不大（数十万到数百万级别）但格式多样覆盖了从简单的图像描述到复杂的多轮视觉推理等各种任务。LLaVA-Instruct-158K 是这类数据集的开山之作。

方法：在指令数据集上对整个模型（或 LLM 骨干的大部分参数）进行端到端的有监督微调（Supervised Fine-tuning，SFT）。

关键：数据的质量和多样性远比数量更重要。许多工作都致力于如何利用 GPT-4 等更强大的模型来自动生成更高质量的指令数据。

这一“先对齐后微调”的两阶段范式在成本和效果之间取得了很好的平衡被证明是训练强大的多模态理解模型的有效路径。

2.6.2 多阶段渐进式训练：VITA-1.5 的精细化策略

随着模型需要处理的模态越来越多（如加入语音）以及对实时交互等更高能力的要求简单的两阶段训练可能已不足以实现最优的对齐。为了更精细地协调不同模态的学习过程 2025 年的 VITA-1.5 提出了一种多阶段渐进式训练（Multi-stage Progressive Training）策略。

VITA-1.5 的训练过程被分解为四个精心设计的阶段层层递进逐步解锁模型的能力：

阶段一：语言-视觉对齐。与传统方法类似使用大规模图文对数据对齐视觉编码器和 LLM。

阶段二：语言-音频对齐。在第一阶段的基础上加入音频模态。使用大规模的“音频-文本”对数据（如语音识别数据）训练一个音频编码器与 LLM 的对齐。

阶段三：多模态指令微调。使用包含图像、音频和文本的混合指令数据集对整个模型进行微调教会模型处理图文、声文混合的指令。

阶段四：对话能力微调。最后使用真实世界的多模态对话数据进一步提升模型的交互流畅度和上下文理解能力。

这种渐进式的训练策略如同精心设计的课程让模型在每个阶段都聚焦于一个特定的学习目标。它避免了在训练初期就用过于复杂的混合模态数据“淹没”模型从而实现了更稳定、更高效的训练过程。VITA-1.5 凭借这一策略成功地在单一模型中高效地整合了视觉和语音两大核心交互模态并实现了出色的实时性能。

2.6.3 数据策略的创新：从海量噪声到高质量合成

训练方法的创新离不开数据本身的创新。在数据层面一个清晰的趋势是从追求“量”转向追求“质”。

早期（2022-2023）：研究者们主要依赖于从网络上爬取的、未经清洗的数十亿级别的图文对数据（如 LAION-5B）。这种“大力出奇迹”的方式虽然有效但也带来了数据偏见、内容不可控等一系列问题。

中期（2023-2024）：随着 GPT-4 等强大模型的出现数据合成（Data Synthesis）成为主流。研究者们发现利用 GPT-4 的 API 可以从少量的人类标注数据出发“生成”出海量、高质量、多样化的指令微调数据。LLaVA-Instruct-158K 的成功充分证明了合成数据在激发 LLM 多模态能力方面的巨大潜力。

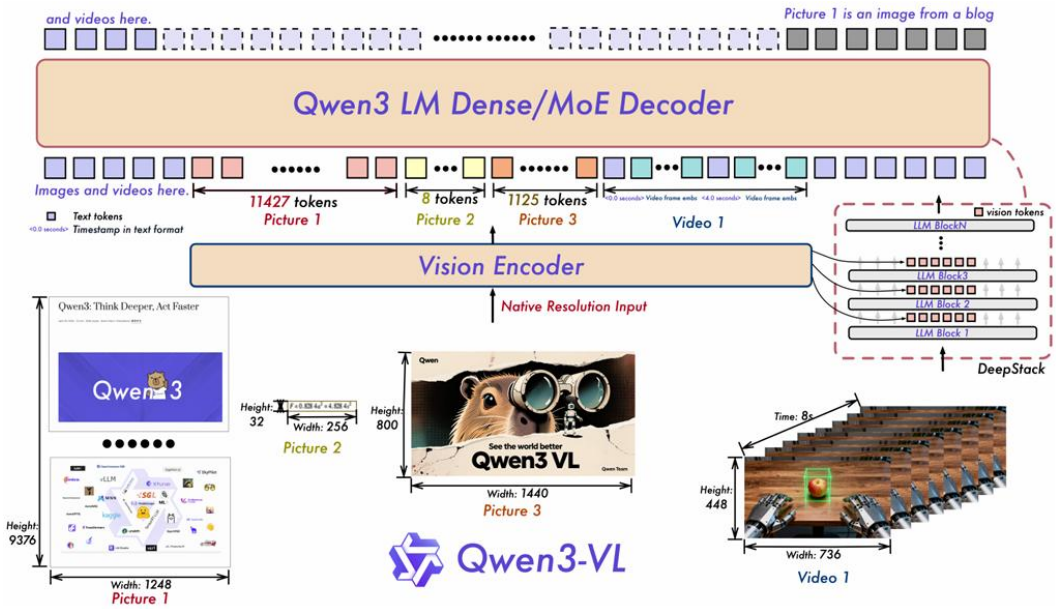
近期（2025）：数据策略变得更加精细化。例如 ChartMoE 为了训练其图表理解能力专门构建了一个包含 100 万个“图表-表格-JSON-代码”四元组的庞大数据集通过多任务学习让模型深刻理解图表的结构化信息。EMOVA 为了让模型理解情感专门设计了包含丰富情感标注的语音-文本数据集。这种面向特定能力的、高度结构化的数据构建正在成为提升模型专业能力的关键。

总结而言训练方法的创新是一场在数据效率、能力对齐和训练成本之间不断寻求最优解的探索。从经典的两阶段范式到更精细化的多阶段渐进式训练再到数据策略从“量”到“质”的转变这些创新共同确保了多模态大语言模型能够在有限的资源下被高效地“塑造”成我们所期望的、功能强大的智能体。

2.7 国内代表性模型的架构创新

2025 年国内涌现出一批在架构设计上极具创新性的多模态大语言模型，它们针对特定问题提出了独特的解决方案。

Qwen3-VL 的“深”与“长”：Qwen3-VL 的架构设计核心在于解决多模态长上下文的挑战。



其增强的交错 MRoPE（Interleaved Multi-head Rotational Positional Embedding）是对传统旋转位置编码的改进，使其能更好地处理视频帧之间、以及图文交错内容中的时空关系。而 DeepStack 技术则借鉴了特征金字塔网络的思想，将视觉编码器（ViT）不同层级的特征进行有效融合，使得语言模型不仅能看到高层的语义信息，也能获取底层的细节纹理，从而实现更精细的视觉-语言对齐。这种“深”度特征融合与“长”上下文处理能力的结合，使其在处理长视频和复杂文档时表现突出。

DeepSeek-OCR 的“轻”与“巧”：DeepSeek-OCR 的思路则完全不同，它巧妙地将“语言问题”转化为“视觉问题”来降维。

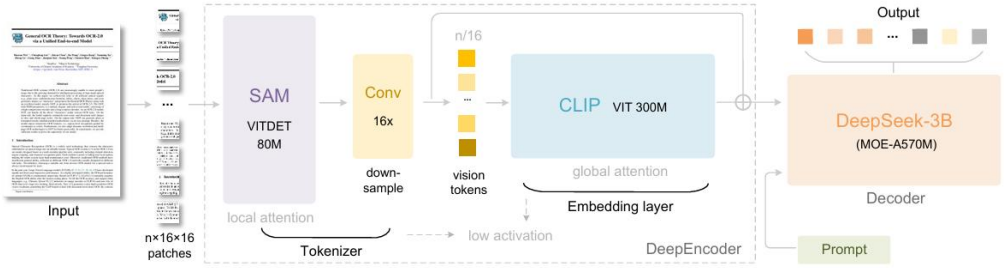


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three components: a SAM [17] for perception dominated by window attention, a CLIP [29] for knowledge with dense global attention, and a 16× token compressor that bridges between them.

其核心组件 DeepEncoder 通过串联窗口注意力、16 倍卷积压缩器和全局注意力，实现了在高分辨率输入下依然能产出极少量（通常少于 100 个）视觉 tok

en 的壮举。这使得后续的 3B MoE 语言模型可以轻松地进行“解压缩”（即 OCR 识别）。这种“先压缩再解压”的模式，本质上是一种“以空间换时间”的策略，极大地降低了长文档处理的计算复杂度，体现了架构设计的“轻”与“巧”。

文心 5.0 的“原生”与“统一”：文心 5.0 最大的特点在于其“原生全模态”的设计哲学。与大多数先分别预训练单模态编码器，再通过连接模块进行对齐的“胶水”模型不同，文心 5.0 从一开始就将所有模态的数据（文本、图像、音频、视频）放入一个统一的 Transformer 架构中进行端到端的联合训练。这种“大一统”的方法理论上能让模型在最底层就学习到不同模态之间最本质的关联，从而在需要深度跨模态推理的任务上展现出更强的性能和更好的一致性。这是对多模态建模范式的一次大胆探索。

Emu3.5 的“预测”与“加速”：Emu3.5 的架构服务于其“世界模型”的定位。

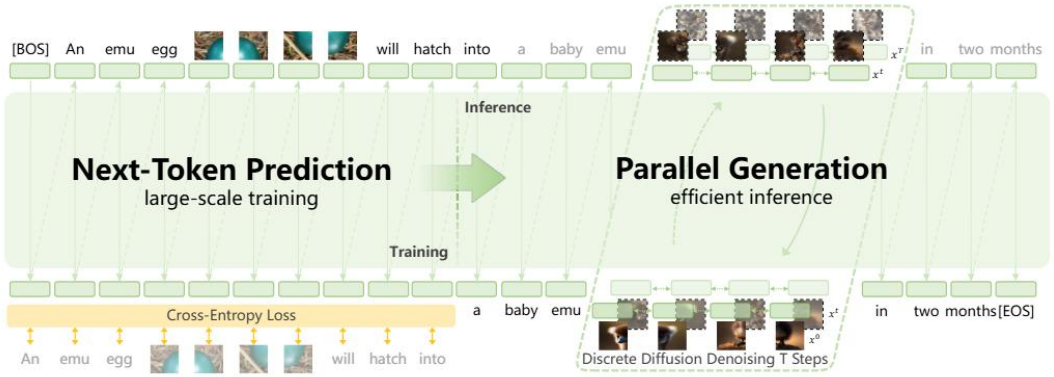


Figure 3: Overview of the Emu3.5 architecture. The model is trained end-to-end at scale with a unified next-token prediction objective. During inference, single-token prediction is accelerated via discrete diffusion adaptation, enabling bidirectional parallel generation per image.

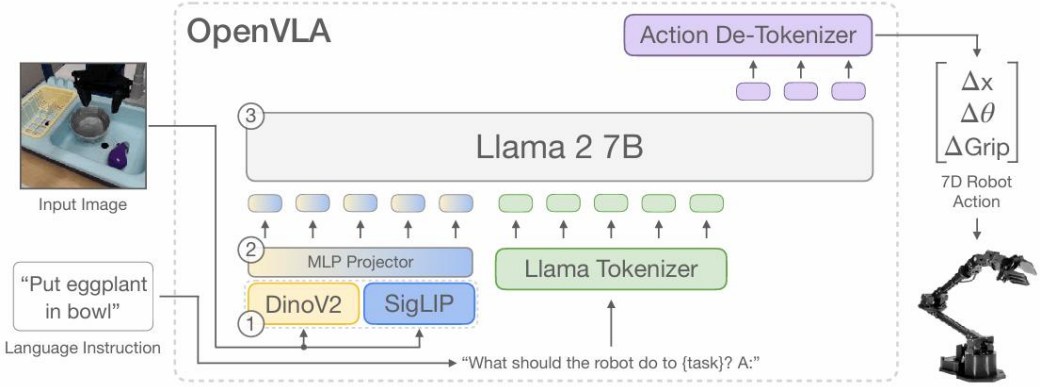
它将所有任务都统一为“预测下一个 token”，无论是文本、图像还是动作。这种极简的统一目标使其能够学习到世界状态的演化规律。然而，传统的自回归生成方式效率低下，无法满足世界模型实时交互的需求。为此，Emu3.5 引入了 DiDA（离散扩散适配），这是一种将自回归的逐 token 生成，巧妙地转换为可以并行计算的双向预测的技术。通过 DiDA，Emu3.5 在生成图像等任务上实现了约 20 倍的推理加速，使其世界模型的能力从理论走向了实用。

这些国内模型的架构创新，从不同角度为多模态技术的发展贡献了宝贵的思路，共同构成了 2025 年多模态技术百花齐放的繁荣景象。

2.8 OpenVLA：开启开源机器人操控新时代

OpenVLA 是首个完全开源的“视觉-语言-动作”（Vision-Language-Action,

VLA）模型。它在高达 97 万条真实世界机器人演示数据上进行训练，为通用机器人操控策略树立了新的性能标杆，并极大地推动了具身智能领域的研究。



OpenVLA 的架构设计精巧而高效。它创新性地采用了双视觉编码器设计，融合了 DINOv2 和 SigLIP 两个强大的预训练视觉模型的特征。DINOv2 擅长提供低层次的空间几何信息，而 SigLIP 则能提供高层次的语义信息。这种“双剑合璧”的设计，显著增强了模型对复杂场景的视觉泛化能力，这对于机器人需要在多变环境中进行精确操作至关重要。视觉特征通过一个小型 MLP 投影器映射到语言嵌入空间，并与基于 Llama 2 7B 的大语言模型骨干相结合。整个模型通过端到端的方式进行训练，直接将视觉-语言模型微调用于生成机器人的控制动作。

在动作表示上，OpenVLA 采用了一种简单而有效的方法。它将连续的 7 维机器人动作（例如，手臂末端执行器的三维位置、三维姿态和夹爪状态）离散化为一系列的整数“tokens”。每个动作维度被独立地划分为 256 个区间（bins），然后将这些离散的动作 tokens 直接覆写到语言模型词汇表中，从而将动作生成问题巧妙地转化为了一个标准的“下一个 token 预测”问题。这种方法不仅简化了模型架构，也使得利用现有的大语言模型训练框架成为可能。

得益于其强大的架构和在 Open X-Embodiment 这一大规模、多样化数据集上的训练，OpenVLA 在 29 个评估任务和多种机器人硬件上，其绝对任务成功率比之前最先进的、参数量大 7 倍的闭源模型 RT-2-X 高出 16.5%。更重要的是，OpenVLA 支持通过 LoRA 等参数高效微调方法，在消费级 GPU 上快速适应新的机器人和任务，并且其模型权重、训练数据和代码库完全开源。这极大地降低了机器人研究的门槛，为整个社区探索更先进的 VLA 模型铺平了道路。

第三章：数据来源与评估基准

先进的架构和训练方法是多模态大语言模型的“骨架”与“经络”海量、高质量的数据则是其赖以生存和成长的“血液”与“养料”。与此同时科学、全面的评估基准则是衡量模型能力、指引技术发展的“标尺”与“灯塔”。本章将系统性地梳理支撑现代多模态大语言模型发展的两大基石：数据来源与评估基准。我们将详细介绍主流的预训练数据集和指令微调数据集揭示数据策略从“规模”到“质量”再到“结构化”的演进趋势。同时本章还将深度剖析当前流行的多模态评估基准分析其各自的侧重点、优势与局限性为读者在纷繁复杂的模型选择中提供一套清晰、可靠的度量衡。

3.1 数据来源：多模态智能的基石

多模态大语言模型的训练过程本质上是一个从海量数据中学习知识、对齐语义的过程。根据在训练流程中所扮演的角色这些数据可以被分为两大类：用于奠定模型基础能力的预训练数据集以及用于教会模型遵循人类意图的指令微调数据集。

3.1.1 预训练数据集：奠定通用视觉-语言基础

预训练阶段的目标是让模型学习到通用的、可迁移的视觉-语言关联知识。这一阶段的数据特点是规模巨大、来源广泛、但质量参差不齐。它们主要来源于公开的学术数据集和更大规模的网络爬取数据。

A.学术图文对数据集

在早期研究者们主要使用一些经典的、经过人工标注和清洗的学术数据集。这些数据集质量高但规模相对较小。

COCO(Common Objects in Context)：包含约 33 万张图像每张图像有 5 个独立的描述字幕。它是 VQA、图像描述等任务最经典的基准之一。

Visual Genome(VG)：规模更大包含约 10 万张图像但标注更密集涵盖了区域描述、对象关系、属性等丰富信息非常适合学习细粒度的视觉概念。

SBU Captions：包含约 100 万张从 Flickr 上收集的、带有用户提供描述的图像是早期向更大规模网络数据探索的尝试。

B.大规模网络图文对数据集

为了构建更强大的通用模型研究者们逐渐转向从互联网上爬取更大规模的图文对数据。这些数据集的规模可达数十亿甚至数万亿级别为模型提供了前所未

有的广阔视野。

LAION 系列：由德国非营利组织 LAION 维护是目前应用最广泛的网络图文数据集。其代表性版本包括：

LAION-400M：包含 4 亿个从网络上筛选出的英文图文对。

LAION-5B：规模扩大到 58.5 亿个多语言图文对并通过 CLIP 相似度评分、水印检测等多种方式进行了过滤是训练 Stable Diffusion 等知名生成模型的核心数据源。

DataComp：这是一个由 Meta AI 领导的、旨在构建更高质量网络数据集的社区项目。DataComp 不仅提供了一个包含 128 亿图文对的庞大数据集更重要的是它提出了一套系统性的数据过滤和去重策略证明了数据质量比单纯的数据规模更重要。使用 DataComp 的 12.8B 子集训练的 CLIP 模型性能优于使用 LAION -5B 的 5 倍数据量训练的模型。

表 7：代表性预训练数据集对比

数据集	规模	来源	主要特点
COCO	33 万	学术标注	质量极高多角度描述但场景有限。
Visual Genome	10 万	学术标注	标注密集包含场景图适合细粒度学习。
LAION-5B	58.5 亿	网络爬取	规模巨大多语言但噪声较大。
DataComp-12.8B	128 亿	网络爬取	规模极大强调高质量过滤性能优越。

3.1.2 指令微调数据集：对齐人类意图的关键

预训练赋予了模型通用的视觉-语言知识但要让模型变得“听话”、“好用”还需要通过指令微调来对齐人类的意图。指令微调数据集的特点是规模相对较小、但结构化、多样化并紧密围绕“指令-响应”的格式构建。

A. 通用视觉-语言指令数据集

这类数据集旨在提升模型在各种通用场景下的对话和问答能力。其构建的核心思想是利用更强大的“教师模型”（如 GPT-4）来生成高质量的指令数据这一方法被称为“LLM-as-a-Judge”或数据合成。

LLaVA-Instruct-158K：开创性的工作。它将 COCO 数据集中的图像与标注信息（描述、边界框）输入给 GPT-4 并设计了巧妙的提示让 GPT-4 生成关于这些图像的对话、问答和推理等多种类型的指令数据。

ShareGPT4V：一个更大规模的社区驱动项目收集了大量用户与 GPT-4V 的真实对话数据并将其整理成一个包含 120 万条指令的数据集。这些数据更加自然、多样反映了真实世界用户的需求。

LRV-Instruction：另一个高质量的指令数据集它不仅包含正向的问答还包含了对模型错误回答的批评和修正这有助于训练出更具批判性思维、更不容易产生幻觉的模型。

B.面向特定能力的指令数据集

随着研究的深入研究者们开始构建面向特定能力的数据集以“靶向”提升模型在某一专业领域的性能。

图表理解：ChartQA 和 ChartMoE 使用的数据集包含了大量的图表图像及其对应的表格数据、摘要和问题专门用于训练模型理解和分析图表的能力。

文档与 OCR：M-VQA 等数据集专注于包含大量文本的文档图像训练模型进行文档级的 OCR 和信息提取。

视频理解：Video-MME 和 ActivityNet-QA 等包含了视频片段和关于视频内容的时间性问题用于训练模型的动态事件理解和时序推理能力。

情感与语音：EMOVA 为了训练模型的情感语音生成能力专门构建了一个包含文本、语音和对应情感标签的数据集。

从大规模、高噪声的网络数据到小规模、高质量的合成指令数据再到面向特定能力的结构化专业数据数据来源的演进趋势清晰地表明：未来的多模态大语言模型竞争将不仅仅是模型架构的竞争更是数据策略和数据生态的竞争。拥有独特、高质量的专有数据将成为构建差异化、高性能模型的关键壁垒。

3.2 评估基准：度量多模态智能的标尺

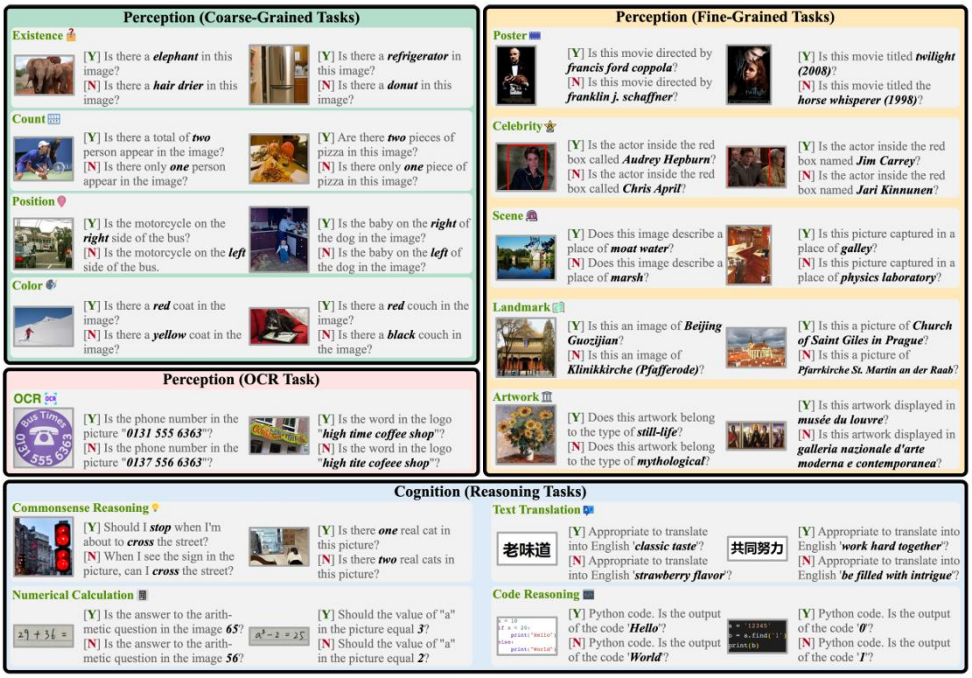
随着多模态大语言模型能力的日益增强和多样化如何科学、全面、公正地评估这些模型成为了一个极具挑战性的课题。一个好的评估基准不仅是衡量模型性能的“标尺”更是指引未来研究方向的“灯塔”。本节将梳理当前主流的多模态评估基准并根据其评估的侧重点将其分为通用能力评估、特定任务评估和交互式评估三大类。

3.2.1 通用能力评估基准：全面考察综合素质

这类基准旨在全面地、多维度地评估一个多模态大语言模型的通用视觉-语言理解能力。它们通常包含大量精心设计的、覆盖不同能力象限的题目并采用统一的评分标准。

MME(A Comprehensive Evaluation Benchmark for Multimodal Large Language Models)：MME 是目前最流行、最被广泛认可的通用能力评估基准之一。

它的核心特点是全面性和对幻觉的鲁棒性。



MME 总共包含 2000 个题目覆盖了 14 个二级能力方向包括存在性判断、海报理解、名人识别、场景理解、OCR、常识推理等。其所有题目都被设计为“是/否”问答题这极大地简化了评估过程并有效避免了因 LLM 输出的文本风格不同而带来的评分偏差。MME 的得分同时考虑了准确率和幻觉率是一个综合性的评价指标。

MM-Vet(A Comprehensive Benchmark for Evaluating Vision-Language Models)：MM-Vet 是另一个重要的通用能力基准。它创新性地引入了基于 GPT-4 的评分机制。

MM-Vet 包含 200 个问题这些问题被设计为开放式回答更能激发模型的自由表达能力。在评分时 MM-Vet 将模型的回答与人类专家的标准答案一同提交给 GPT-4 由 GPT-4 来判断模型的回答是否正确、全面。这种“LLM-as-a-Judge”的评分方式虽然引入了一定的不确定性但能够更好地评估模型回答的质量和信息量。

SEED-Bench：SEED-Bench 则更侧重于评估模型在高级认知和推理能力上的表现。它包含约 1.9 万个多项选择题覆盖了从场景理解到科学知识问答等 12 个维度。其题目设计更具挑战性需要模型进行更深层次的推理。

表 8：主流通用能力评估基准对比

基准	题目数量	题型	评分方式	主要特点
MME	2，000	是/否问答	精确匹配	全面鲁棒抗幻觉易于评估。

MM-Vet	200	开放式问答	GPT-4 评分	评估回答质量更接近人类判断。
SEED-Bench	19，000	多项选择	精确匹配	侧重高级认知与推理能力。

3.2.2 特定任务评估基准：衡量专业领域能力

除了通用能力模型在特定专业领域的能力也至关重要。为此研究者们开发了一系列针对特定任务的评估基准。

图表理解：ChartQA 是该领域的标准基准。它包含数千个关于图表的问题这些问题既需要视觉上的信息提取（如“哪个柱子最高？”）也需要逻辑上的推理计算（如“A 比 B 高多少？”）。

视频理解：Video-MME 是 MME 在视频领域的扩展专门用于评估模型对动态视频内容的理解能力。它包含对视频中动作、关系、时序的提问。ActivityNet-QA 则提供了更具挑战性的、需要对视频进行时间定位和推理的问答。

数学推理：MathVista 是一个极具挑战性的数学问题基准其题目通常以图像形式呈现（如几何图形、函数图像）需要模型结合视觉理解和数学推理才能解答。

OCR 与文档理解：DocVQA 和 TextVQA 等基准专注于评估模型从包含大量文本的图像（如扫描文档、街景照片）中读取和理解信息的能力。

3.2.3 交互式与动态评估：走向真实世界

传统的静态评估基准（模型对一张图片或一个问题给出一次性回答）难以完全反映模型在真实世界中的交互能力。因此新的评估范式正在兴起它们更侧重于评估模型在动态、多轮交互过程中的表现。

VITA-Eval(VITA-1.5 提出): 这是一个专门为评估实时视觉-语音交互能力而设计的基准。评估人员会与模型进行实时的、多轮的语音对话同时向模型展示物体或场景。评估指标包括模型的响应延迟、语音识别的准确性、视觉理解的精确度以及对话的流畅性。这种真人评测的方式虽然成本高但最能反映模型的真实用户体验。

Arena-Style Battle Platforms：受到 LMSYS Chatbot Arena 的启发一些平台（如 LLaVA-Arena）开始采用“竞技场”模式。它们会随机展示两个不同多模态模型的回答让用户来投票选择哪个更好。通过大量的用户投票可以得到一个相对客观的模型排名。这种众包式的评估方式能够有效地捕捉用户的主观偏好。

评估基准的演进从静态的、单轮的问答到动态的、多轮的、交互式的评测反映了研究界对模型“真实世界能力”的日益重视。未来的评估基准将更加关注模

型的鲁棒性、安全性、公平性以及与人类价值观的对齐。构建一个既能全面衡量模型能力又能高效、低成本部署的评估体系将是多模态领域持续面临的重要挑战。

3.3 数据质量与模型性能的关系

数据是 AI 的燃料，但并非所有的燃料都具有相同的品质。近年来的研究越来越清晰地表明，数据的质量远比数量更重要。本节将深入探讨数据质量的多个维度，以及它们如何影响多模态大语言模型的最终性能。

3.3.1 图文对齐质量的重要性

在视觉-语言预训练中，图文对的对齐质量是决定模型能否学到有意义的跨模态关联的关键。所谓对齐质量，指的是图像内容与其对应文本描述之间的语义一致性和相关性程度。

噪声数据的负面影响。从网络上爬取的图文对数据，不可避免地包含大量噪声。这些噪声可能来自多个方面：文本与图像完全无关（如网页上的广告文字被错误地与新闻图片配对）；文本只描述了图像的一小部分内容；文本包含错误或误导性信息；图像质量低下（模糊、截断、水印覆盖）等。研究表明，在包含高比例噪声数据的数据集上训练的模型，不仅性能提升缓慢，还可能学到错误的跨模态关联，导致在下游任务中出现系统性偏差。

CLIP 相似度过滤。为了提升数据质量，LAION 系列数据集引入了基于 CLIP 相似度的过滤机制。具体而言，对于每一个图文对，使用预训练的 CLIP 模型计算图像 embedding 和文本 embedding 之间的余弦相似度。只有相似度超过某个阈值（如 0.3）的图文对才会被保留。这种过滤方法的有效性已经得到了广泛验证：在经过 CLIP 过滤的数据上训练的模型，性能显著优于在原始未过滤数据上训练的模型。然而，这种方法也存在一个潜在的问题：它可能会引入“选择偏差”——那些与 CLIP 模型的训练数据分布相似的样本更容易被保留，而那些新颖的、罕见的样本则可能被过滤掉，从而限制了模型学习到的知识的多样性。

人工标注与合成数据的价值。相比于网络爬取的数据，人工标注的数据具有更高的质量保证。COCO 和 Visual Genome 等经典数据集，虽然规模相对较小（数十万级别），但由于其精确的标注和丰富的语义信息，在模型训练中仍然发挥着不可替代的作用。特别是在指令微调阶段，高质量的人工标注数据是必不可少的。然而，人工标注的成本极高，难以扩展到数十亿级别的规模。为了在质量和规模之间取得平衡，利用强大的 LLM（如 GPT-4）来自动生成高质量的合成

数据，已经成为一种主流策略。LLaVA-Instruct-158K 的成功充分证明了这一策略的有效性。

3.3.2 数据多样性与模型泛化能力

除了质量，数据的多样性也是影响模型泛化能力的关键因素。多样性可以体现在多个维度：视觉内容的多样性（不同的物体、场景、风格）、语言表达的多样性（不同的句式、词汇、语言）、任务类型的多样性（描述、问答、推理、生成）等。

长尾分布的挑战。真实世界的数据分布通常呈现长尾特征：少数常见的类别（如“猫”、“狗”、“汽车”）占据了数据的大部分，而大量罕见的类别（如特定品种的动物、专业领域的物体）只有极少的样本。在这种不平衡的数据上训练的模型，往往在常见类别上表现良好，但在罕见类别上性能急剧下降。这种“长尾问题”在多模态领域尤为突出，因为视觉世界的多样性远超文本世界。为了缓解这一问题，一些工作探索了数据重采样、类别平衡损失函数、以及利用外部知识库来增强罕见类别的表示等方法。

跨领域泛化。多模态模型的一个重要应用场景是跨领域迁移，即在一个领域（如自然图像）上训练的模型，能够泛化到另一个领域（如医疗影像、卫星图像）。然而，不同领域的数据分布可能存在巨大差异，直接迁移往往效果不佳。为了提升跨领域泛化能力，训练数据需要覆盖尽可能多样的领域。DataComp 项目在构建数据集时，特别强调了领域多样性，通过从不同来源、不同主题的网站上爬取数据，确保数据集能够覆盖广泛的视觉和语言分布。实验表明，在多样性更高的数据上训练的模型，在跨领域任务上的泛化能力显著提升。

多语言与文化多样性。当前的多模态研究主要集中在英语数据上，这导致模型在处理非英语语言和非西方文化内容时性能下降。为了构建真正的全球化 AI，数据的多语言和文化多样性至关重要。LAION-5B 包含了多种语言的图文对，是朝这一方向迈出的重要一步。然而，即使在多语言数据集中，不同语言的数据量也极不平衡，英语仍然占据主导地位。如何收集和利用更多的非英语、非西方文化的数据，是未来需要重点关注的方向。

3.4 评估基准的演进与局限性

评估基准是衡量技术进步的标尺，但任何标尺都有其局限性。本节将深入分析当前多模态评估基准的演进趋势和存在的问题。

3.4.1 从单一任务到综合能力评估

早期的多模态评估主要聚焦于单一任务，如图像描述（Image Captioning）、视觉问答（VQA）等。这些任务虽然经典，但往往只能反映模型在某一特定能力维度上的表现，难以全面评估模型的综合能力。随着多模态大语言模型能力的日益强大和多样化，评估基准也在向更综合、更全面的方向演进。

MME 的综合性设计。MME 是这一演进趋势的典型代表。它包含 14 个二级能力方向，覆盖了从低级的感知能力（如 OCR、物体识别）到高级的认知能力（如常识推理、数学计算）。更重要的是，MME 在设计时特别考虑了对“幻觉”的鲁棒性评估。它的所有题目都采用“是/否”问答格式，并且包含大量的负样本（即答案为“否”的问题）。这种设计可以有效地检测模型是否倾向于过度自信地给出肯定答案，从而量化模型的幻觉程度。MME 的得分同时考虑了准确率和幻觉率，是一个更为全面的综合指标。

基准饱和与持续创新。然而，随着模型性能的快速提升，一些经典的评估基准开始出现“饱和”现象，即顶尖模型的得分已经接近或达到人类水平，基准失去了区分不同模型能力的作用。例如，在 COCO 图像描述任务上，许多模型的自动评估指标（如 CIDEr、BLEU）已经超过了人类标注者。这种饱和现象促使研究者们不断开发更具挑战性的新基准。MathVista 就是这样一个例子，它专注于需要视觉理解和数学推理相结合的复杂问题，即使是最先进的模型在这个基准上的表现也远未达到饱和。

3.4.2 自动评估与人工评估的权衡

在多模态任务中，特别是生成类任务（如图像描述、图像生成），如何进行评估是一个长期存在的难题。自动评估指标（如 BLEU、CIDEr、FID）虽然高效、可复现，但往往无法完全捕捉生成内容的质量和多样性。人工评估虽然更准确，但成本高昂、主观性强、难以大规模部署。

LLM-as-a-Judge 的兴起。近年来，一种新的评估范式——“LLM-as-a-Judge”——开始流行。这种方法使用强大的 LLM（如 GPT-4）来评估模型的输出质量。具体而言，将模型生成的回答与参考答案一同提交给 GPT-4，并设计详细的评分标准（rubric），让 GPT-4 给出评分和评价。MM-Vet 就采用了这种评估方式。研究表明，GPT-4 的评分与人类专家的评分具有较高的一致性，同时又具有自动评估的高效性。然而，这种方法也存在一些问题：它依赖于特定的 LLM（如

GPT-4），而这些 LLM 本身可能存在偏见；评估结果可能受到提示词（prompt）设计的影响，缺乏稳定性；更重要的是，当被评估的模型本身就是基于类似 LLM 构建的，可能会出现“自我强化”的偏差——模型学会了生成更符合 GPT-4 偏好的回答，而不一定是更符合人类真实需求的回答。

多维度评估的必要性。鉴于单一评估指标的局限性，越来越多的研究开始采用多维度评估框架。例如，在评估图像生成模型时，不仅要评估生成图像的视觉质量（如 FID、IS），还要评估其与文本提示的一致性（如 CLIP Score）、生成的多样性（如 Diversity Score）、以及是否包含不当内容（如 NSFW 检测）。这种多维度评估可以更全面地反映模型的能力和局限性，但也增加了评估的复杂度和成本。

第四章：应用场景与实践

技术的发展终将服务于现实世界的需求。多模态大语言模型以其跨越多种信息媒介的独特能力正以前所未有的深度和广度渗透到社会生产和个人生活的方方面面。本章将聚焦于多模态技术的“落地实践”系统性地梳理其在四大核心应用领域的现状、关键技术挑战及未来发展趋势。我们将通过具体的模型案例和实践分析展示多模态技术如何从抽象的算法演变为解决实际问题、创造巨大价值的强大工具。这四大领域分别是：高级视觉理解、多模态内容创作、实时交互式助手以及具身智能与机器人。

4.1 高级视觉理解：超越“看图说话”

高级视觉理解是多模态大语言模型最基础、也是最核心的应用领域。它要求模型不仅能“看到”图像或视频中的物体更能“理解”它们之间的关系、背后的逻辑以及所处的复杂情境。随着模型能力的增强其应用早已超越了简单的“看图说话”正朝着更专业、更需要深度推理的方向发展。

4.1.1 复杂场景与常识推理

现代多模态大语言模型特别是像 GPT-4V 和 Gemini 这样的大型闭源模型已经展现出惊人的常识推理能力。它们能够理解图片中的幽默、讽刺和隐含的因果关系。例如向模型展示一张“宇航员在月球上骑马”的图片并提问“这张图有什么不寻常之处？”模型能够准确地指出“马不可能出现在月球上因为月球没有支持马生存所需的大气和生态系统”。这种能力使其在教育、信息检索和内容审核

等领域具有巨大潜力。

4.1.2 专业领域的视觉分析

将多模态技术应用于专业领域是提升生产效率的关键。当前在多个垂直领域多模态视觉分析已经取得了显著进展。

医疗影像分析：一些研究工作尝试利用多模态模型来分析 X 光片、CT 扫描等医疗影像。通过将影像与对应的诊断报告进行联合训练模型可以学习到识别病灶、辅助医生进行诊断的能力。这有望缓解医疗资源紧张、提升诊断效率。

金融图表分析：金融领域充斥着大量的 K 线图、财报图表。传统的分析方法依赖于人工解读效率低下且容易出错。以 ChartMoE 为代表的模型通过在海量的图表数据上进行训练能够精准地回答关于图表的复杂问题如“哪个季度的增长率最高？”或“预测下一阶段的趋势”。这为量化交易、投资分析等场景提供了强大的自动化工具。

自动驾驶感知：在自动驾驶领域多模态模型能够融合来自摄像头、激光雷达（LiDAR）、毫米波雷达等多种传感器的数据形成对周围环境更全面、更鲁棒的理解。这有助于模型在恶劣天气（如雨、雪、雾）或光照不足等单一传感器容易失效的场景下做出更安全的决策。

4.1.3 视频内容理解与摘要

随着短视频和流媒体的爆发如何高效地从海量视频中提取信息成为了一个巨大的挑战。多模态大语言模型在视频理解方面展现出巨大潜力。

事件检测与定位：模型能够识别视频中的关键事件（如“进球瞬间”、“颁奖典礼”）并精确定位其在视频中的起止时间。Video-MME 等基准专门用于评估这种能力。

视频摘要与问答：对于长视频（如电影、会议记录）模型可以自动生成简洁的内容摘要或根据用户提问快速找到并回答视频中的相关内容（如“CEO 在会议的哪个部分讨论了下一季度的财报？”）。这极大地提升了视频信息的检索和利用效率。

表 9：高级视觉理解应用领域与关键技术

应用领域	关键挑战	代表性技术/模型	商业价值
常识推理	抽象概念理解反事实推理	GPT-4V,Gemini	内容审核智能搜索引擎教育辅导
金融图表分析	结构化数据提取逻辑推理	ChartMoE	自动化交易策略智能投研报告

医疗影像分析	细粒度特征识别专业知识对齐	Med-PaLM	辅助诊断降低漏诊率提升医生效率
视频内容摘要	时序关系理解长上下文处理	Video-MME,Qwen3-Omni	智能媒资管理高效视频检索会议纪要生成

高级视觉理解是多模态技术创造价值的起点。随着模型在专业知识对齐和深度推理能力的进一步提升我们有理由相信它将在更多垂直领域扮演“超级专家”的角色成为各行各业不可或缺的生产力工具。

4.2 多模态内容创作：人机协同的新范式

视觉理解是让 AI“读懂”世界多模态内容创作（Multimodal Content Creation）则是让 AI“描绘”世界。2025 年随着生成范式（特别是流模型）的革命性突破多模态大语言模型在内容创作领域的角色正从一个简单的“工具”演变为一个深度参与创作流程的“合作伙伴”。这不仅极大地提升了内容生产的效率更催生了全新的创意表达方式。

4.2.1 高质量、高效率的图像与视频生成

生成质量和生成速度是衡量内容创作工具好坏的核心标准。2025 年的模型在这两方面都取得了巨大飞跃。

图像生成：以 JanusFlow 为代表的模型通过其创新的“AR + Flow”混合范式实现了在极快的速度下（通常只需 1-8 个采样步骤）生成与顶级扩散模型（如 Midjourney）质量相媲美的图像。这意味着用户几乎可以“实时”地通过文本描述来创造和修改图像极大地提升了设计师、插画师的工作效率。

视频生成：视频生成是更具挑战性的领域。早期的模型（如 Sora）虽然效果惊艳但其闭源和高昂的计算成本限制了其应用。2025 年以 Wan，NExT-OMNI 等为代表的开源模型利用统一的流模型框架实现了更高质量、更长时长的文生视频、图生视频。虽然与顶级商业模型仍有差距但其快速的迭代和开源的特性正在推动视频创作的普及化。

4.2.2 交错多模态内容的涌现：Mogao 的创新

传统的内容创作是“单点式”的即生成一张图或一段视频。而 2025 年的一个标志性进展是交错多模态内容（Interleaved Multimodal Content）的生成。Mogao 在这方面做出了开创性贡献。

Mogao 能够生成一个包含文本、图像等多种模态交错出现的长序列。这意味着 AI 可以：

创作图文并茂的文章：模型可以直接生成一篇完整的、包含标题、段落和配图的博客文章或新闻报道。

生成多模态幻灯片：根据一个主题自动生成包含标题页、目录页、带图表的正文页和总结页的完整演示文稿的“草稿”。

这种能力使得 AI 不再仅仅是一个被动执行指令的工具而更像一个能够主动规划、组织和表达复杂思想的“创作者”。这为个性化教育材料的生成、自动化新闻撰写、智能营销内容创作等领域打开了全新的想象空间。

4.2.3 交互式编辑与精细化控制

专业的内容创作不仅需要生成更需要精确的“修改”。现代多模态大语言模型正在提供越来越精细的交互式编辑能力。

像素级编辑：以 VITRON 为代表的模型由于其统一的像素级表示能够实现非常精确的图像编辑。用户可以通过简单的文本指令如“把这只猫的颜色变成蓝色”或“移除背景里的路人”来实现类似 Photoshop 的复杂操作。

风格迁移与属性控制：用户可以提供一张参考图片让模型学习其独特的艺术风格并应用到新生成的内容上。同时也可以通过文本指令精确控制生成人物的姿势、表情、服装等属性。

表 10：多模态内容创作应用与技术演进

应用方向	早期技术 (2023-2024)	2025 年技术突破	带来的变革
图像生成	扩散模型 (速度慢)	流模型 (JanusFlow)	实现高质量图像的实时、交互式生成。
视频生成	短视频、一致性差	统一流模型 (NExT-OMNI)	生成更长、更连贯的视频内容降低创作门槛。
复杂内容	单模态生成	交错生成 (Mogao)	从生成“素材”到生成“成品”AI 成为创作伙伴。
图像编辑	基于掩码的粗略编辑	像素级统一模型 (VITRON)	实现指令驱动的、Photoshop 级的精确编辑。

多模态内容创作的革命其核心是降低专业创作的门槛同时提升专业人士的创作效率。随着技术的进一步发展 AI 与人类创作者的界限将变得越来越模糊。未来的内容创作将是一种人机深度协同、相互启发的新范式这将极大地解放人类的创造力催生出前所未有的艺术形式和内容生态。

4.3 实时交互式助手：迈向“全能”个人助理

将多模态大语言模型打造成一个能够像人一样自然、流畅地与我们交流的

“全能”个人助理是该领域最激动人心的愿景之一。2025 年随着模型在实时性、多模态理解和情感表达方面取得的重大突破这一愿景正加速成为现实。其核心应用场景包括智能客服、个人 AI 助理以及为特殊人群提供辅助。

4.3.1 实时视觉-语音交互的突破

传统的多模态交互存在显著的延迟。例如用户说完一句话模型需要先通过 ASR（自动语音识别）转成文本再进行 LLM 处理最后通过 TTS（文本到语音）合成声音整个链条耗时数秒体验非常不自然。2025 年以 VITA-1.5 和闭源的 GPT-4o 为代表的模型在实时交互方面取得了革命性突破。

这些模型的成功关键在于其端到端的统一架构和高效的流式处理能力：

统一建模：它们不再依赖于分离的 ASR、LLM 和 TTS 模块而是在一个统一的模型内部直接处理音频波形和视觉信息避免了模块间切换带来的延迟。

流式处理：模型能够以“流”的方式处理输入和输出。用户可以随时打断模型模型也能在思考的同时就开始生成回应从而将响应延迟降低到人类可以接受的数百毫秒级别。

实践案例：VITA-1.5 的实时交互

在 VITA-1.5 的演示中用户可以一边用手机摄像头拍摄周围的环境一边用自然的语音向模型提问：“我眼前这是什么花？”。模型几乎可以瞬间回答：“这是向日葵它的花盘会随着太阳转动。”用户可以立刻追问：“那它旁边那个红色的建筑呢？”模型也能无缝衔接继续进行对话。这种流畅的体验与此前基于“请求-等待-响应”模式的交互形成了鲜明对比。

4.3.2 情感交互与个性化

真正的智能助手不仅要“听懂”更要“听出”言外之意。2025 年的 EMOVA 模型在情感交互方面做出了开创性工作。通过其创新的语义-声学解耦的语音分词器 EMOVA 能够将语音信息分解为“内容”（说了什么）和“声学”（怎么说的）两个部分。这使得模型可以：

理解情感：从用户说话的语调、语速中判断其情绪状态（如高兴、悲伤、焦虑）。

表达情感：在生成语音回应时能够带上符合当前对话氛围的情感。例如在用户表达悲伤时用更轻柔、更富同理心的语气进行安慰。

这种情感交互能力对于智能客服、虚拟伴侣、心理咨询等应用场景至关重要。

它使得人机交互不再是冰冷的“信息交换”而更接近温暖的“情感沟通”。

4.3.3 面向特殊人群的辅助应用

实时交互式助手在为视障、听障等特殊人群提供信息无障碍服务方面具有巨大的社会价值。

视觉辅助：对于视障用户多模态助手可以成为他们的“眼睛”。用户可以通过佩戴带有摄像头的智能眼镜将看到的画面实时传输给模型。模型则通过语音为用户描述周围的环境、识别物体、朗读路牌和菜单、甚至提醒前方的障碍物。

听觉辅助：对于听障用户模型可以将周围的声音（如门铃、火警、他人的呼唤）实时地转化为文字或震动提醒。在对话中模型也可以将对方的语音实时转化为字幕显示在手机或 AR 眼镜上。

实时交互式助手的成熟预示着一个“AI-in-the-loop”时代的到来。AI 将不再仅仅存在于手机或电脑中而是以一种无处不在、时刻在线的形式深度融入我们的日常生活成为我们感知世界、与世界交互能力的自然延伸。如何确保这些助手的可靠性、安全性以及对用户隐私的保护将是其大规模普及前必须解决的核心问题。

4.4 具身智能与机器人：从虚拟走向物理

具身智能是多模态技术发展的终极前沿之一。它旨在让 AI 拥有“身体”能够通过感知物理世界并与之交互来学习和执行任务。这要求模型不仅能处理文本、图像等虚拟信息更能理解和生成在物理世界中执行的“动作”（Action）序列。2025 年随着多模态大语言模型与机器人技术的加速融合具身智能正从实验室研究快步走向现实应用其核心是构建能够连接语言指令与物理动作的“世界模型”。

4.4.1 世界模型：构建物理世界的内部模拟

“世界模型”（World Model）是具身智能的核心概念。它指的是 AI 在内部构建的一个关于物理世界如何运作的动态、可预测的模型。通过这个内部模拟 AI 可以在“脑海中”预演不同动作可能带来的后果从而做出更优的决策而无需在真实世界中进行大量昂贵的试错。

世界模型能主动地预测未来，生成一个可交互、符合物理规律的“虚拟沙盒”。这为具身智能体提供了一个近乎无限的、安全的、低成本的训练环境，从而根本上解决了数据稀缺和真实世界训练风险高的核心痛点。

世界模型的核心思想是让智能体在内部构建一个关于外部世界的动态模型。智能体可以利用这个内部模型进行“心理推演”或“前瞻性规划”，在不与真实

世界交互的情况下，预测不同动作序列可能导致的结果，从而选择最优策略。2025 年，多个里程碑式的世界模型相继发布，将这一概念推向了新的高度。

2025 年代表性世界模型对比：

Google Genie 3 标志着世界模型进入了实时交互的时代。它能够从文本、图像甚至草图中生成 720p 高清、可实时探索的虚拟世界。其最令人印象深刻的能力在于长时域一致性，即使用户在虚拟世界中漫游数分钟，场景中的物体和环境依然能保持物理上的一致性，不会出现明显的漂移或错乱。此外，Genie 3 引入的“可提示的世界事件”允许用户通过文本指令动态改变世界状态(如改变天气)，为智能体训练提供了更丰富的交互维度。

腾讯混元 HY-World 1.5 是完全开源的实时交互世界模型。它以 24 FPS 的流畅帧率运行，并着重解决了“长期几何一致性”这一核心难题，确保了场景在长时间交互下的稳定性。HY-World 1.5 的开源策略意义重大，它提供了一套从数据处理到模型训练再到推理优化的完整框架，极大地降低了社区研究和开发世界模型的门槛，推动了该技术的民主化。

中科院与 CreateAI 联合发布的 NeoVerse 则在 4D 世界建模上取得了突破。它创新性地实现了仅通过采集的普通单目视频，在 30 秒内快速构建一个包含时间维度的 4D 世界表示。这项技术摆脱了对昂贵的多视角采集设备或激光雷达数据的依赖，使得大规模、低成本地构建 4D 数字孪生成为可能，在自动驾驶、城市管理和机器人导航等领域具有巨大的应用潜力。

由 AI 先驱李飞飞创立的 World Labs 推出的 Marble 是世界模型商业化的首次重要尝试。它定位为一个面向艺术家、设计师和开发者的 3D 世界创建工具，用户可以通过多模态输入（文本、图像等）快速生成高保真的 3D 世界，并进行迭代编辑和分享。Marble 的出现，标志着世界模型技术正从实验室研究走向创造经济的实际生产力工具。

这些世界模型的出现，为具身智能的发展铺平了道路。智能体可以在这些高度逼真且符合物理规律的虚拟世界中进行数百万次的试错和学习，其获得的经验和技能可以更有效地迁移到真实世界中，从而大大加速通用人工智能（AGI）的实现进程。

4.4.2 语言指令到物理动作的转化

具身智能的一个核心挑战是如何将人类用自然语言下达的模糊指令（如“把桌子收拾干净”）转化为机器人可以理解和执行的一系列精确的物理动作(如“识

别所有盘子和杯子”、“规划抓取路径”、“将它们放入水槽”）。

多模态大语言模型在其中扮演了“翻译官”和“任务规划师”的角色：

场景理解：模型首先通过摄像头“看懂”当前的物理环境识别出桌子、盘子、杯子等关键物体及其空间关系。

任务分解：LLM 骨干网络利用其强大的常识和逻辑推理能力将“收拾干净”这个高级指令分解为一系列具体的子任务。

动作生成：对于每个子任务模型会生成对应的机器人控制代码或动作原语（Action Primitives）。

4.4.3 模拟器与真实世界的鸿沟（Sim-to-Real Gap）

由于在真实世界中收集机器人训练数据的成本极高且风险巨大绝大多数具身智能模型的训练都在高逼真度的物理模拟器（如 NVIDIA Isaac Sim）中完成。然而模拟环境与真实世界之间始终存在差异（如光照、摩擦力、物体材质等）这导致在模拟器中表现完美的模型在真实机器人上部署时往往会失败。如何弥合这一“模拟到现实的鸿沟”（Sim-to-Real Gap）是具身智能落地的关键挑战。

当前的解决方案主要包括：

域随机化（Domain Randomization）：在模拟器中故意将环境的各种参数（如光照、纹理、物理参数）进行随机化从而让模型学习到对环境变化更鲁棒的策略。

少量真实世界微调：在模拟器中完成大规模预训练后再在真实机器人上进行少量的、针对性的微调以适应真实世界的特性。

具身智能是多模态技术最具想象力、也最具挑战性的应用方向。它代表着 AI 从数字世界向物理世界的终极跨越。2025 年我们看到了多模态大语言模型与机器人技术的深度融合为解决这一终极挑战带来了新的曙光。虽然距离电影中那样的通用人形机器人还有很长的路要走但技术发展的车轮已经开始加速。未来具备物理实体的 AI 将深刻地改变我们的生产方式、生活方式乃至整个社会结构。

第五章：当前挑战与未来展望

在经历了 2025 年的爆发式增长后多模态大语言模型技术在取得辉煌成就的同时也逐渐暴露出其在技术、应用和伦理层面面临的一系列深层次挑战。对这些挑战的清醒认识是确保该领域能够持续、健康发展的必要前提。本章将首先系统性地剖析当前多模态技术在计算资源、数据、模型能力和安全伦理四个维度所面

临的核心挑战。在此基础上本章将结合前沿的探索性研究对多模态技术的未来演进方向进行展望重点探讨其在构建更通用世界模型、迈向自主智能以及与其他AI技术融合方面的巨大潜力。通过对“挑战”与“机遇”的辩证分析本章旨在为读者提供一个关于多模态技术未来走向的、更为审慎和前瞻的视角。

5.1 当前挑战：通往通用智能之路的障碍

尽管多模态大语言模型已经展现出惊人的能力但距离实现真正通用、可靠、安全的人工智能仍然面临着诸多严峻的挑战。

5.1.1 计算与资源的“诅咒”

模型规模的持续膨胀带来了对计算资源无止境的需求这已成为限制技术发展和普及的最大障碍之一。

高昂的训练成本：训练一个如Qwen3-Omni级别的先进模型需要数千甚至上万个顶级GPU进行数周乃至数月的计算。这使得只有少数科技巨头和国家级实验室能够参与到这场“军备竞赛”中极大地限制了学术界和中小型企业的创新空间。

严峻的推理挑战：即使模型训练完成其巨大的尺寸也给部署和推理带来了巨大压力。要在个人电脑甚至移动设备上运行这些模型需要进行复杂的模型压缩和量化而这往往会以牺牲性能为代价。如何开发出更高效的模型架构和推理引擎是决定多模态技术能否真正“飞入寻常百姓家”的关键。

算泥社区定位为“AI 大模型开发服务+算法+算力”三位一体的 AI 开发者社区，提供国产异构算力服务，使其有可能在这一趋势中扮演重要角色。通过整合英伟达、寒武纪等多种 AI 芯片，并利用异构计算技术，社区为开发者提供了一种稳定、高效的算力资源选择。这在开发者训练或微调模型、应对高昂推理成本等现实挑战中，提供了一个规避“卡脖子”风险的潜在解决方案。未来，这类平台与国产硬件厂商的深度合作，以及对异构调度能力的持续优化，将是其发展的关键观察点。

5.1.2 数据的“瓶颈”与“偏见”

数据是AI的燃料而当前多模态模型正面临着数据上的双重困境。

高质量视频与交错数据的稀缺：虽然我们拥有海量的网络图文数据但高质量、多样化的视频数据特别是包含精确时间戳和动作标注的视频仍然极度稀缺。此外用于训练多模态生成的图文音混合长序列数据几乎完全依赖于成本高昂的人工

构建或合成。这种数据瓶颈是限制模型在动态理解和复杂内容创作方面能力提升的主要原因。

数据偏见与公平性：网络数据不可避免地反映了现实世界中存在的社会、文化和历史偏见。在这些数据上训练出的模型可能会无意中学习并放大这些偏见例如在生成图像时表现出刻板印象或在理解不同口音的语音时存在能力差异。如何检测、量化并缓解这些数据偏见是一个亟待解决的技术和伦理难题。

5.1.3 模型能力的“幻觉”与“脆弱”

尽管模型在许多基准上取得了超人的表现但其能力边界仍然存在明显的“软肋”。

幻觉问题（Hallucination）：这是所有大模型都存在的通病。在多模态场景下幻觉表现为模型“看到”或“听到”了不存在的东西或者对图像内容进行与事实完全不符的描述。虽然 MME 等基准的设计在一定程度上可以抑制幻觉但如何从根本上让模型“知之为知之不知为不知”仍然是一个开放性问题。

对微小扰动的脆弱性：研究表明许多多模态模型对输入中微小的、人眼难以察觉的扰动（即“对抗性攻击”）非常敏感。例如在图像中加入一些精心设计的噪声就可能让模型将其完全识别成另一种物体。这种脆弱性使得在自动驾驶、医疗诊断等高风险安全攸关领域的应用充满了隐患。

物理世界常识的缺乏：尽管模型学习了海量的知识但它们对物理世界的基本规律（如物体恒存性、因果关系、力的作用）的理解仍然非常肤浅。这在具身智能领域表现得尤为突出机器人常常会犯一些在人类看来匪夷所思的“低级错误”。

5.1.4 安全与伦理的“红线”

随着模型能力的日益强大其被滥用的风险也与日俱增对社会安全和伦理规范构成了前所未有的挑战。

深度伪造（Deepfake）与信息操纵：高质量的图像、视频、音频生成技术可能被用于制造虚假新闻、进行身份欺诈、或传播恶意政治宣传对社会信任和公共安全构成严重威胁。

隐私泄露：实时交互式助手需要持续地访问用户的摄像头和麦克风这带来了巨大的隐私泄露风险。如何确保这些敏感数据在设备端得到妥善处理防止被滥用或泄露是赢得用户信任的前提。

责任界定：当一个由 AI 驱动的自动驾驶汽车发生事故或一个 AI 辅助诊断

系统出现误诊时责任应该如何界定？是算法的设计者、训练数据的提供者还是最终用户？缺乏清晰的法律和伦理框架将阻碍这些技术的进一步应用。

表 11：多模态大语言模型当前面临的核心挑战

挑战维度	具体挑战	潜在影响
计算与资源	训练成本高昂推理部署困难	创新门槛提高技术普及受阻
数据	高质量视频/交错数据稀缺数据偏见	模型能力提升遭遇瓶颈公平性问题凸显
模型能力	幻觉对抗性攻击缺乏物理常识	可靠性不足在高风险领域应用受限
安全与伦理	深度伪造隐私泄露责任界定困难	社会信任危机技术滥用风险法律法规滞后

对这些挑战的克服不可能一蹴而就。它需要算法、硬件、数据、法律和伦理等多个层面的协同努力。只有正视这些障碍并以负责任、审慎的态度去解决它们多模态技术才能真正行稳致远最大限度地释放其造福社会的潜力。

5.2 未来展望：迈向更通用、更自主的智能

尽管挑战重重但多模态大语言模型展现出的巨大潜力依然让我们对人工智能的未来充满期待。站在 2025 年的时间节点上我们可以预见未来的技术演进将主要围绕着构建更通用的“世界模型”、追求更高级的“自主智能”以及与其他 AI 技术进行更深度的“融合创新”这三大主线展开。

5.2.1 世界模型：从“感知”到“理解”物理世界

如第四章所述构建能够模拟和预测物理世界动态的“世界模型”是具身智能的终极目标也是多模态技术未来最重要的发展方向之一。未来的世界模型将呈现以下趋势：

更丰富的模态融合：当前的世界模型主要融合视觉、语言和动作。未来的模型将进一步整合触觉、力觉、声音等更丰富的传感器信息以构建对物理世界更全面、更细致的内部表征。

从模拟器到真实世界的飞跃：解决“Sim-to-Real”的鸿沟将是未来数年的研究重点。通过开发更高逼真度的模拟器、更有效的域适应技术以及更高效的真实世界数据收集策略（如让机器人在安全的环境中自主探索和学习）模型将逐步摆脱对模拟器的依赖。

涌现的物理常识：随着模型在更大规模、更多样化的物理交互数据上进行训练我们有望看到模型“涌现”出对物理规律更深层次的、符合直觉的理解。这将

使其能够在全新的、未曾见过的场景中做出更合理、更安全的决策。

5.2.2 自主智能：从“执行者”到“规划者”

当前的多模态模型在很大程度上仍是一个被动的“执行者”需要人类给出明确的指令。未来的发展方向是赋予模型更高层次的自主性（Autonomy）使其能够成为一个主动的“规划者”和“决策者”。

主动学习与探索：未来的模型将具备主动探索未知环境、识别自身知识盲点并提出有价值问题的能力。例如一个家庭服务机器人在遇到一个不认识的物体时可以主动向用户提问：“这是什么？我应该如何处理它？”。这种主动学习的能力将使其能够持续地、终身地获取新知识。

长期任务规划：模型将能够理解和规划需要数小时、数天甚至更长时间才能完成的复杂任务。例如用户可以给出“帮我策划一个为期一周的巴黎旅行”这样的高级指令模型能够自主地完成信息检索、行程规划、酒店预订、撰写攻略等一系列子任务。

AI Agent 的兴起：多模态大语言模型将成为驱动各种 AI Agent（智能代理）的核心大脑。这些 Agent 将在数字世界（如自动完成在线购物、管理日程）和物理世界（如自动驾驶、机器人）中代表人类自主地执行任务成为无处不在的“数字员工”和“物理劳工”。

5.2.3 融合创新：与其他 AI 技术的协同进化

多模态大语言模型并非孤立发展的技术其未来的巨大潜力在于与其他 AI 技术的深度融合。

与强化学习的融合：强化学习（Reinforcement Learning，RL）提供了让模型通过试错来学习最优策略的强大框架。将 LLM 的通用知识和推理能力作为强化学习智能体的“先验知识”可以极大地提升其学习效率和泛化能力这在游戏 AI 和机器人控制领域已经展现出巨大潜力。

与知识图谱的融合：知识图谱（Knowledge Graphs）能够以结构化的方式存储和表示世界知识。将多模态模型与知识图谱相结合一方面可以利用知识图谱为模型提供更准确、更可解释的知识来源缓解“幻觉”问题；另一方面也可以利用模型从非结构化的多模态数据中自动抽取和更新知识构建更全面、更及时的知识图谱。

与脑机接口的融合：这是一个更具科幻色彩但潜力巨大的远景。通过脑机接

口（Brain-Computer Interface，BCI）人类的思想可以直接与多模态大语言模型进行交互。这不仅可能为残障人士提供前所未有的交流和控制能力更有可能从根本上重塑人机交互的定义实现真正意义上的“人机共生”。

5.3 结语

2025 年我们正站在一个由多模态大语言模型开启的、波澜壮阔的新时代的入口。技术的发展正以前所未有的速度将曾经只存在于科幻作品中的想象变为现实。从“感知”到“理解”从“执行”到“规划”从“工具”到“伙伴”多模态技术正引领人工智能迈向一个更通用、更自主、更深度融入人类社会的新纪元。

这条道路上机遇与挑战并存。技术的突破需要仰望星空的想象力更需要脚踏实地的严谨和对伦理红线的敬畏。我们有理由相信在全世界研究者、开发者和决策者的共同努力下多模态人工智能技术必将行稳致远成为推动人类文明向前发展的、一股强大而向善的力量。本报告的撰写正是希望为这一伟大的历史进程提供一份系统性的思考和一份审慎的展望。

参考文献表

ViLBERT：Pretraining for Grounded Vision-Language Tasks. Advances in Neural Information Processing Systems，32. https：//arxiv.org/abs/1908.02265

LXMERT：Learning Cross-Modality Encoder Representations from Transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). https：//arxiv.org/abs/1908.07490

Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning，PMLR 139. https：//arxiv.org/abs/2103.00020

Visual Instruction Tuning. Advances in Neural Information Processing Systems，36. https：//arxiv.org/abs/2304.08485

LLaMA：Open and Efficient Foundation Language Models. arXiv preprint arXiv：2302.13971. https：//arxiv.org/abs/2302.13971

Chameleon：Mixed-Modal Early-Fusion Foundation Models. arXiv preprint arXiv：2405.09818. https：//arxiv.org/abs/2405.09818

VITRON：A Unified Pixel-level Vision LLM. arXiv preprint. https：//haofe

i.vip/downloads/papers/Skywork_Vitron_2024.pdf

Show-o：One Single Transformer to Unify Multimodal Understanding and Generation. arXiv preprint arXiv：2408.12528. https：//arxiv.org/abs/2408.12528

Janus：Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. arXiv preprint arXiv：2410.13848. https：//arxiv.org/abs/2410.13848

JanusFlow：Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation. arXiv preprint arXiv：2411.07975. https：//arxiv.org/abs/2411.07975

NExT-OMNI：Towards Any-to-Any Omnimodal Foundation Models with Discrete Flow Matching. arXiv preprint arXiv：2510.13721. https：//arxiv.org/abs/2510.13721

VITA-1.5：Towards GPT-4o Level Real-Time Vision and Speech Interaction. arXiv preprint arXiv：2501.01957. https：//arxiv.org/abs/2501.01957

Qwen3-Omni Technical Report. arXiv preprint arXiv：2509.17765. https：//arxiv.org/html/2509.17765v1

Mogao：An Omni Foundation Model for Interleaved Multi-Modal Generation. arXiv preprint arXiv：2505.05472. https：//arxiv.org/abs/2505.05472

Denoising Diffusion Probabilistic Models. Advances in Neural Information Processing Systems，33. https：//arxiv.org/abs/2006.11239

MME：A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. Advances in Neural Information Processing Systems，36. https：//arxiv.org/abs/2306.13394

Video-MME：A Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). https：//arxiv.org/abs/2405.21075

Attention is All You Need. Advances in Neural Information Processing Systems，30. https：//arxiv.org/abs/1706.03762

Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems，33. https：//arxiv.org/abs/2005.14165

BLIP：Bootstrapping Language-Image Pre-training for Unified Vision-Langu

age Understanding and Generation. Proceedings of the 39th International Conference on Machine Learning，PMLR 162. https：//arxiv.org/abs/2201.12086

BLIP-2：Bootstrapping Language-Image Pre-training with a Frozen Image Encoder and a Frozen Large Language Model. Proceedings of the 40th International Conference on Machine Learning，PMLR 202. https：//arxiv.org/abs/2301.12597

InstructBLIP：Towards General-purpose Vision-Language Models with Instruction Tuning. Advances in Neural Information Processing Systems，36. https：//arxiv.org/abs/2305.06500

MiniGPT-4：Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv：2304.10592. https：//arxiv.org/abs/2304.10592

GPT-4V(ision) System Card. https：//cdn.openai.com/papers/GPTV_System_Card.pdf

Gemini：A Family of Highly Capable Multimodal Models. arXiv preprint arXiv：2312.11805. https：//arxiv.org/abs/2312.11805

Visual ChatGPT：Talking，Drawing and Editing with Visual Foundation Models. arXiv preprint arXiv：2303.04671. https：//arxiv.org/abs/2303.04671

HuggingGPT：Solving AI Tasks with ChatGPT and its Friends in Hugging Face. Advances in Neural Information Processing Systems，36. https：//arxiv.org/abs/2303.17580

M2-Omni：A Unified Framework for Any-to-Any Modality Conversion. arXiv preprint arXiv：2402.19288. https：//arxiv.org/abs/2402.19288

An Image is Worth 16x16 Words：Transformers for Image Recognition at Scale. arXiv preprint arXiv：2010.11929. https：//arxiv.org/abs/2010.11929

Mini-Gemini：Mining the Potential of Multi-modality Vision Language Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). https：//arxiv.org/abs/2403.18814

Taming Transformers for High-Resolution Image Synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). https：//arxiv.org/abs/2012.09841

ChartMoE：A Mixture-of-Experts Model for Chart Understanding. Internatio nal Conference on Learning Representations(ICLR) Oral. https：//proceedings.icl r.cc/paper_files/paper/2025/file/c33cd281f8cd784626a57de340e43fe4-Paper-Confere nce.pdf

EMOVA：Empowering Language Models to See，Hear，and Speak with E motion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition(CVPR). https：//openaccess.thecvf.com/content/CVPR2025/papers /Chen_EMOVA_Empowering_Language_Models_to_See_Hear_and_Speak_with_C VPR_2025_paper.pdf

Microsoft COCO：Common Objects in Context. European Conference on C omputer Vision. https：//arxiv.org/pdf/1405.0312.pdf

Visual Genome：Connecting Language and Vision Using Crowdsourced De nse Image Annotations. International Journal of Computer Vision，123. https：// arxiv.org/pdf/1602.07332.pdf

Im2Text：Describing Images Using 1 Million Captioned Photographs. Adva nces in Neural Information Processing Systems，24. https：//proceedings.neurips. cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

LAION-400M：Open-Source Image-Text-Pairs in English. arXiv preprint ar Xiv：2111.02114. https：//arxiv.org/pdf/2111.02114.pdf

LAION-5B：An Open Large-Scale Dataset for Training Next Generation Im age-Text Models. Advances in Neural Information Processing Systems，35. http s：//proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb34c91a364b45209 9000f6a-Paper-Datasets_and_Benchmarks.pdf

DataComp：In search of the next generation of multimodal datasets. Advan ces in Neural Information Processing Systems，36. https：//proceedings.neurips.c c/paper_files/paper/2023/file/2042259174df491c33b76437d23e5939-Paper-Datasets_ and_Benchmarks.pdf

ShareGPT4V：A Large-Scale Dataset of Images and High-Quality Captions from GPT-4V. arXiv preprint arXiv：2311.12793. https：//arxiv.org/pdf/2311.127 93.pdf

The LLM-as-a-Judge Score for Unveiling the Vision-Language Capability o

f MLLMs. arXiv preprint arXiv：2311.17907. https：//arxiv.org/pdf/2311.17907.pdf

ChartQA：A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. Findings of the Association for Computational Linguistics：ACL 2022. https：//aclanthology.org/2022.findings-acl.177.pdf

M-VQA：A Multimodal Question Answering Benchmark for Document Images. Proceedings of the IEEE/CVF International Conference on Computer Vision. https：//openaccess.thecvf.com/content/ICCV2021/papers/Hu_UniT_Multimodal_Multitask_Learning_With_a_Unified_Transformer_ICCV_2021_paper.pdf

ActivityNet-QA：A Dataset for Understanding Localized Narratives in Videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https：//openaccess.thecvf.com/content_CVPR_2019/papers/Yu_ActivityNet-QA_A_Dataset_for_Understanding_Localized_Narratives_in_Videos_CVPR_2019_paper.pdf

MM-Vet：A Comprehensive Benchmark for Evaluating Vision-Language Models. arXiv preprint arXiv：2308.02490. https：//arxiv.org/pdf/2308.02490.pdf

SEED-Bench：A Benchmark for Multimodal Foundation Models. arXiv preprint arXiv：2307.16125. https：//arxiv.org/pdf/2307.16125.pdf

MathVista：Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). https：//openaccess.thecvf.com/content/CVPR2024/papers/Lu_MathVista_Evaluating_Mathematical_Reasoning_of_Foundation_Models_in_Visual_Contexts_CVPR_2024_paper.pdf

DocVQA：A Dataset for VQA on Document Images. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. https：//openaccess.thecvf.com/content/WACV2021/papers/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.pdf

Towards VQA Models That Can Read. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. https：//openaccess.thecvf.com/content_CVPR_2019/papers/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.pdf

OpenVLA：An Open-Source Vision-Language-Action Model. arXiv preprint arXiv：2406.09246. https：//arxiv.org/pdf/2406.09246.pdf

Qwen3-VL：Towards Long-Context Multimodal Understanding. arXiv preprint arXiv：2511.21631. https：//arxiv.org/abs/2511.21631

DeepSeek-OCR：Optical Compression for Efficient Document Understanding. arXiv preprint arXiv：2510.18234. https：//arxiv.org/abs/2510.18234

文心 5.0：https：//qianfan.cloud.baidu.com/qianfandev/topic/687501

Emu3.5：Large-Scale Multimodal World Model with Discrete Diffusion Adaptation. arXiv preprint arXiv：2510.26583. https：//arxiv.org/abs/2510.26583

Genie 3：A General-Purpose World Model for Real-Time Interaction. Google DeepMind Blog. https：//deepmind.google/blog/genie-3-a-new-frontier-for-world-models/

HY-World 1.5：Open-Source Real-Time Interactive World Model. arXiv preprint. https：//github.com/Tencent-Hunyuan/HY-WorldPlay

NeoVerse：Fast 4D World Reconstruction from Monocular Videos. arXiv preprint arXiv：2601.00393. https：//arxiv.org/abs/2601.00393

Marble：Commercial 3D World Generation Platform. World Labs. https：//marble.worldlabs.ai/

主编单位：中科算网算泥社区
网址：sumw.com.cn
邮箱：zhusiliang@sumw.com.cn

算泥社区

大模型交流群