2025年大模型研究系列

多模态大模型洞察

大模型向多模态发展

深入产业端垂直场景释放技术价值

企业标签：百度、腾讯、阿里云、商汤科技

AI变革行业创新发展

China Multimodal Large Model Industry

中国マルチモード モデル産業

研究框架

- ◆ **中国多模态大模型行业综述** ---------- 5
  - 定义
  - 分类
  - 发展历程
  - 市场规模
  - 政策分析
- ◆ **中国多模态大模型产业洞察** ---------- 13
  - 参与者图谱
  - 应用场景
  - 训练方式
  - 生成能力评估
  - 技术发展趋势
  - 痛点与挑战
  - 未来展望
- ◆ **方法论** ---------- 25
- ◆ **法律声明** ---------- 26

名词解释

- **多模态：**指的是能够处理和理解来自多种不同来源和形式的信息的系统，如文本、图像、音频、视频等。多模态技术使机器学习模型能够更全面地理解和表达复杂的真实世界场景。
- 模型开发流程：是指在创建和部署AI模型的过程中所涉及的步骤和阶段。这个过程通常包括问题定义、数据收集和预处理、模型选择和设计、训练、评估，以及最终的部署和维护。
- **模型训练：**是指使用大量已标记或已知结果的数据来调整和优化AI模型的参数，使其能够从数据中学到模式和规律。在训练过程中，模型通过与标签匹配的方式不断调整自身的权重，以提高在未见过数据上的表现。
- **深度学习：**是机器学习的一种分支，它通过模拟人脑的神经网络结构来实现学习和推断。深度学习的核心是深度神经网络，这种网络由多个层次的神经元组成，能够学习复杂的特征表示，广泛应用于图像识别、语音识别等领域。
- **计算机视觉：**是一门研究如何使计算机能够模拟和理解人类视觉系统的学科。它涉及图像和视频的处理，包括目标检测、图像分类、物体识别等任务。
- **机器学习：**是一种通过从数据中学习模式和规律来使计算机系统改善性能的方法。它包括监督学习、无监督学习、强化学习等不同类型，用于解决各种问题，如分类、回归、聚类等。
- **算法框架：**是一种提供了特定问题或任务解决方案基本结构和组织的软件框架。在机器学习和深度学习中，算法框架通常提供模型定义、训练、评估等一系列功能，简化模型开发的流程
- **模态对齐：**指的是在多模态数据中，不同模态的信息需要在语义层面进行对齐，确保它们能够有效地交互和融合，这对于构建有效的多模态模型至关重要
- **模态互补：**指不同模态间的信息互补特性，即一种模态可能无法完全表达的信息可以通过另一种模态得到补充，从而增强整体的信息表达能力
- **联合训练：**是指在一个统一的框架下同时训练多个模态的数据，使模型能够在学习过程中自动发现不同模态间的关联，从而提升模型的整体表现

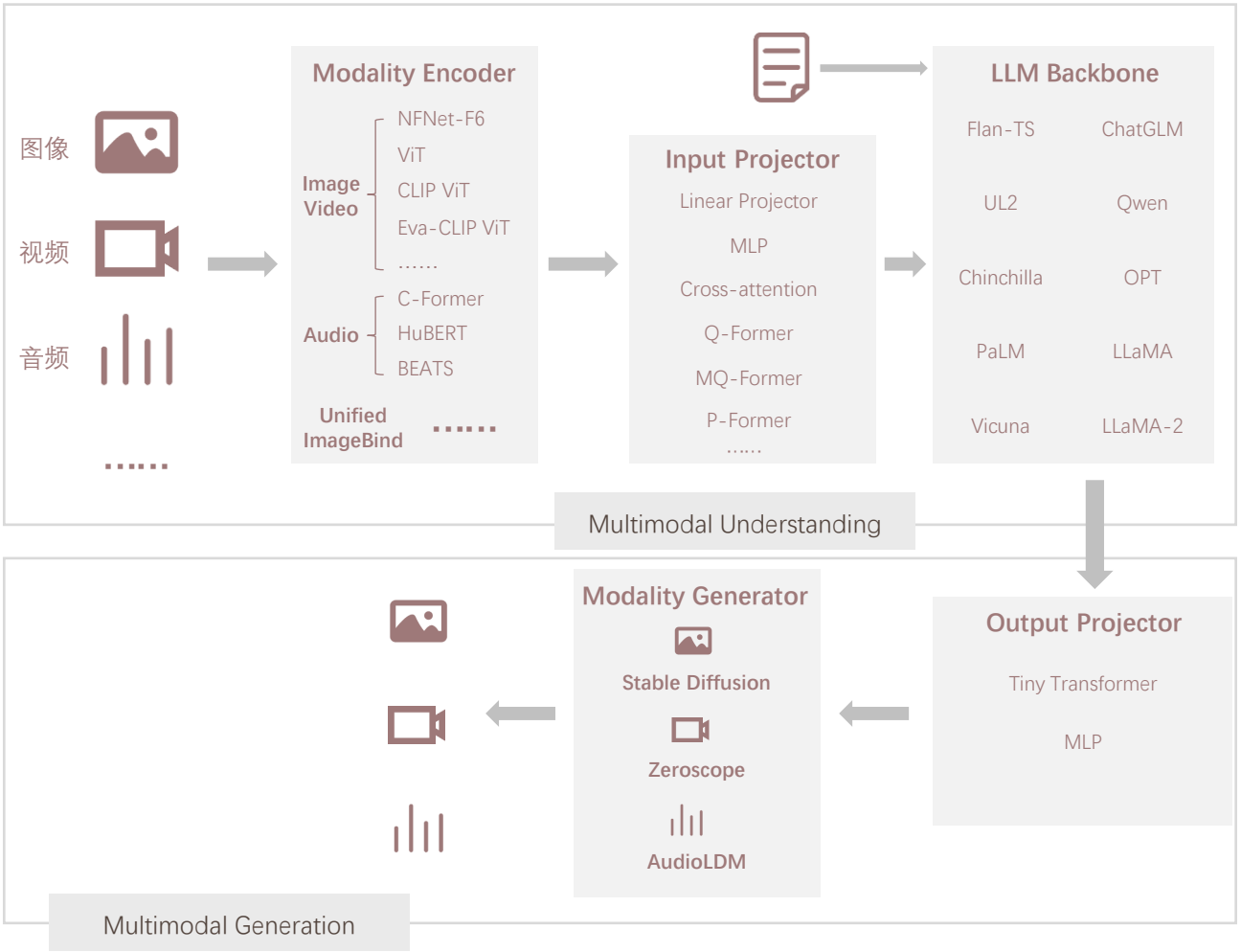
Chapter 1
多模态大模型行业综述

- 定义
- 分类
- 发展历程
- 市场规模
- 政策分析

中国多模态大模型行业综述——定义

- 多模态模型的高效运作依赖于多个组件的协同配合，具体包括模态编码器、输入投影器、大型模型基座、输出投影器以及模态生成器。这些组件共同协作，使得模型能够有效地处理并生成多种模态的数据

多模态大模型的核心组成分析



■ 模态编码器（Modality Encoder, ME）

模态编码器是多模态模型中用于处理和转换原始输入数据的关键组件，其主要功能是将不同模态的数据（如图像、视频、音频等）转换为特征表示，以便后续的处理和融合。常见的模态编码器包括：

- **图像编码器（Image Encoder）：**将图像数据转换为特征向量。常用的图像编码器包括卷积神经网络（CNN）、Vision Transformers（ViT）、CLIP ViT等。这些编码器能捕捉图像中的局部和全局特征，为后续任务提供丰富的信息。
- **视频编码器（Video Encoder）：**将视频数据转换为特征序列。视频编码器通常结合了时空特征的提取，例如3D CNN、I3D（Inflated 3D ConvNets）等。这些编码器能够捕捉视频中的动态变化和时间依赖性。
- **音频编码器（Audio Encoder）：**将音频数据转换为特征表示。常见的音频编码器包括Wav2Vec、HuBERT等。这些编码器能够捕捉音频中的频谱特征和时序信息，适用于语音识别、音乐分类等任务。

来源：企业公告，头豹研究院

定义（接上页）

■ **输入投影器（Input Projector, IP）**

输入投影器负责将不同模态的特征表示转换为统一的表示形式，以便于后续的融合和处理。常见的输入投影器包括：

线性投影器（Linear Projector）：通过简单的线性变换将特征映射到同一空间，适用于特征维度较小的情况。

多层感知器（Multi-Layer Perceptron，MLP）：通过多层非线性变换将特征映射到同一空间，适用于特征维度较大且需要复杂变换的情况。

交叉注意力（Cross-Attention）：通过注意力机制将不同模态的特征进行交互和对齐，增强特征的关联性。

Q-Former：一种基于Transformer的结构，专门用于多模态特征的对齐和融合，能够处理长依赖关系和复杂特征。

■ **大模型基座（LLM Backbone）**

大模型基座是多模态模型的核心部分，负责处理和融合来自不同模态的特征表示，生成中间表示。常见的大模型基座包括：

ChatGLM：基于Transformer的对话生成模型，能够处理文本和多模态输入，生成高质量的对话响应。

LLaMA：一种多模态预训练模型，能够处理图像、文本等多种模态数据，适用于多种下游任务。

Qwen：阿里巴巴推出的多模态预训练模型，具有强大的文本理解和生成能力，支持多模态输入。

Vicuna：一种轻量级的多模态模型，专注于模型的高效性和实时性，适用于资源受限的场景。

■ **输出投影器（Output Projector, OP）**

输出投影器负责将大模型基座生成的中间表示转换为最终的输出形式。常见的输出投影器包括：

Tiny Transformer：一种小型的Transformer模型，用于生成文本或其他模态的输出，适用于资源受限的场景。

多层感知器：通过多层非线性变换将中间表示转换为最终输出，适用于多种输出类型。

■ **模态生成器（Modality Generator, MG）**

模态生成器是多模态模型的最终输出组件，负责生成特定模态的输出。常见的模态生成器包括：

Stable Diffusion：一种基于扩散模型的图像生成器，能够生成高质量的图像，适用于图像生成和修复任务。

Zeroscope：一种视频生成器，能够生成高质量的视频片段，适用于视频合成和编辑任务。

AudioLDM：一种基于Transformer的音频生成器，能够生成高质量的音频，适用于音乐生成和语音合成任务。

来源：企业官网，头豹研究院

中国多模态大模型行业综述——分类

- 多模态大模型的分类方法，包括基于处理输入的方式（如标准定义注意力、定制层的深度融合、输入层融合、使用标记化）和根据功能与技术架构（如内容生成、内容交互、内容理解以及融合编码、分模态处理和跨模态对齐架构）

多模态大模型分类（根据模型内部层对多模态输入的处理方式）

01 基于标准交叉注意力的深度融合

这类模型通常建立在标准的Transformer架构之上，并在其内部层引入交叉注意力机制，以实现不同模态信息之间的深度融合。交叉注意力机制允许模型在处理某一模态的信息时，能够关注到其他模态的相关部分，从而增强模型的理解能力。OpenFlamingo 是这一类别中的典型例子，它能够在处理图像和文本数据时，通过交叉注意力层有效地捕捉两者之间的关联性，适用于图像描述生成、视觉问答等任务。

02 基于定制层的深度融合

这类模型采用了更为灵活的设计思路，通过自定义设计的层（如自注意力层、卷积层等）来进行模态间的融合。这种方式不仅能够针对特定任务需求进行优化，还可以支持更加多样化的模态输入。例如，LLaMA-Adapte 通过引入自适应注意力层，能够在不同模态之间建立动态连接，从而在多模态任务中表现出色。这种灵活性使得B类模型在处理复杂多模态数据时具有显著优势。

03 输入层融合

这类模型在输入层直接对多模态数据进行融合处理，这样的设计具有高度的模块化特性，便于扩展至更多的模态类型。例如，如果一个模型最初只支持文本和图像输入，那么在后续开发中很容易加入对音频或视频的支持。这种架构下的模型通常具有良好的可扩展性和灵活性，适合于快速原型设计和迭代开发。不过，输入层的融合策略需要精心设计，以确保不同模态信息的有效结合而不失各自特色。

04 使用标记化

这类模型采用了一种不同的方法来处理多模态数据——通过标记化（Tokenization）。这种方法首先将不同模态的数据转换成统一的标记序列，然后使用同一个模型来处理这些标记。这样做的好处在于可以简化模型结构，同时保持对多模态数据的良好处理能力。但是，这种方法也面临着挑战，如需要训练一个高效的通用标记器，以及较高的计算成本。

多模态大模型分类（根据架构设计分类）

基于变换器

近年来，基于变换器架构的多模态模型因其强大的并行计算能力和对长距离依赖性的良好处理能力而受到广泛关注。这类模型通常通过自注意力机制来同时处理多种类型的输入

基于卷积神经网络

对于图像等视觉数据，CNN仍然是非常有效的处理工具。一些多模态模型会使用CNN来处理图像数据，并与其他模态的数据相结合

基于循环神经网络

虽然在处理序列数据方面不如变换器流行，但对于某些特定的应用场景（如视频中的动作识别），RNN及其变体（如LSTM、GRU）仍然具有优势

来源：专家访谈，企业官网，头豹研究院

中国多模态大模型行业综述——发展历程

- 从任务专用到通用架构的演进，体现了多模态研究不断追求更高效、更灵活解决方案的努力。随着技术的持续进步，未来多模态模型将更加智能、更加人性化，为人类社会带来前所未有的便利和创新

多模态模型发展历程分析

	阶段1：任务导向型	阶段 2：视觉—语言预训练	阶段 3：多模态语言模型
任务定义	格式 + 数据: 基准A训练集	格式 + 数据: 基准A训练集	语言指令 + ICL（上下文学习）
训练数据	基准任务的训练集	大规模图像-文本集	预训练 + 指令数据
训练目标	任务特定损失函数	自监督任务（如 MLM）	下一词生成
模型规模	<0.1B 参数	0.1B~3B 参数	1B~100B 参数
评估	基准任务的评估集	针对任务的微调	任务的零样本学习

2012　2019　2023　2024

多模态大模型研究范式演进路径

任务导向阶段

- 在多模态研究的初期，模型设计主要集中在解决具体的应用问题上。例如，在视觉问答领域，模型需要理解图像中的内容，并根据给定的问题生成答案；在图像描述中，目标是根据图像生成描述性的文本；而在图文检索中，则需要根据文本查询找到最相关的图像或反之。这些任务虽然都涉及到图像和文本两种模态的数据，但是由于每种任务的具体要求不同，导致早期的模型架构往往是针对特定任务定制的。

视觉-语言预训练阶段

- 多模态大模型研究范式演进路径中，视觉-语言预训练阶段的特点在于它开创性地将视觉和语言两种异构信息的处理融合在单一框架内，通过构建大型、多样化的图像-文本对数据集，并利用自监督学习方法，使得模型能在没有大量标注数据的情况下学习到丰富的视觉和语言表征。这一阶段强调跨模态理解与生成能力的同步提升，即模型不仅能根据给定的文字描述生成相应的图像表示，还能基于输入的图片生成准确且富有语义的文字说明。

多模态大模型阶段

- 最近几年，随着计算资源的增加和技术的发展，研究者开始尝试构建更大规模的多模态模型。这些模型借鉴了大语言模型的成功经验，试图通过增加模型的容量来提高其处理复杂多模态任务的能力。更重要的是，这些模型旨在实现一种更为灵活的交互方式——即使用自然语言作为接口，无论输入是什么形式的多模态数据，模型都能够理解并作出相应的反应。

来源：头豹研究院

中国多模态大模型行业综述——市场规模

- 2023年中国多模态大模型市场规模达到了90.9亿元，预计到2028年将增长至662.3亿元，年复合增长率达48.76%。这一快速增长主要归因于，技术创新的持续驱动，以及行业需求的强劲推动

中国多模态大模型市场规模，2023年-2028年

技术成熟度与行业渗透的加速：随着多模态大模型在图文生成、跨模态检索和视频内容分析等领域的技术突破，其行业应用价值日益显现。多模态技术的逐步成熟使得企业能够在更多场景中部署相关解决方案，如电商中的精准推荐、医疗中的多模态诊断，以及教育中的智能交互系统。

通用模型能力增强：多模态大模型正在向更通用的方向发展，通过语言、视觉、音频等模态的统一处理能力，大幅降低了模型开发和使用的门槛。这使得更多的中小企业也能负担得起技术成本，从而扩大了市场需求。

■ **政策支持与市场需求促进多模态大模型市场规模持续增长**

政府层面的支持是中国多模态大模型市场快速发展的关键驱动力之一。中国政府高度重视人工智能产业的发展，出台了一系列扶持政策，包括资金投入、税收优惠、人才培养等方面的支持措施，为相关企业提供了良好的发展环境。

市场需求的持续增长同样不容忽视。随着消费者对个性化、智能化服务需求的不断提升，各行各业都在积极探索和采用多模态大模型来提升用户体验和服务质量。无论是教育、娱乐还是金融等行业，都有望通过多模态大模型实现业务模式的创新和升级，进而带动整个市场规模的扩大。

来源：头豹研究院

Chapter 2
多模态大模型产业洞察

- 参与者图谱
- 应用场景
- 训练方式
- 生成能力评估
- 技术发展趋势
- 痛点与挑战
- 未来展望

中国多模态大模型产业洞察——参与者图谱

- 2023年中国的多模态大模型发展迅速，头部企业如百度、阿里等已推出多个具备国际竞争力的模型，但与国际巨头相比，在基础架构创新和生态建设上仍有提升空间

中国与国际多模态大模型厂商图谱

首席分析师：oliver.yuan@leadleo.com

主笔分析师：qingmin.chen@leadleo.com

来源：头豹研究院

中国多模态大模型产业洞察——应用场景

- 多模态大模型应用中数字人占据了最大的份额（24%），其次是游戏（13%）和广告商拍（13%）。其他的应用场景包括智能营销、社交媒体、教学辅助、3D建模、智能驾驶等

中国多模态大模型应用场景分析

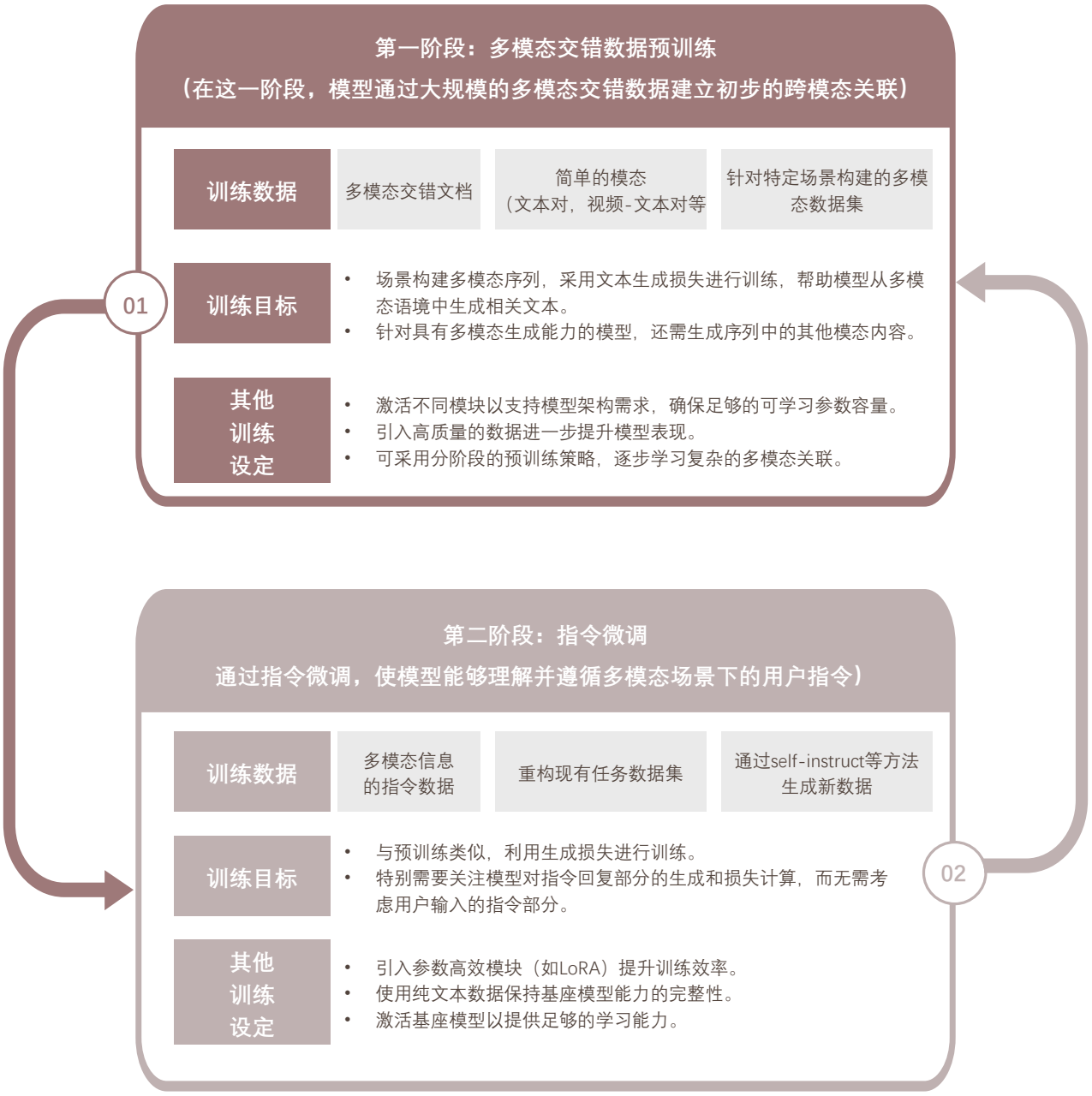
应用领域	应用占比	具体应用	
数字人	24%	虚拟主播	虚拟偶像
		虚拟客服	虚拟导游
游戏	13%	设计与开发	运维与智能客服
		美术与动画制作	游戏内对话系统
广告商拍	13%	广告视频制作	多模态内容理解
		用户画像构建	广告创意策划
智能营销	10%	市场分析	多模态营销内容
		用户画像构建	个性化推荐
社交媒体	10%	多模态内容生成	跨模态情感分析
		多模态内容推荐	多模态内容审核
教学辅助	5%	智能课件生成	个性化学习推荐
		虚拟助教	多模态评估
3D建模	5%	自动模型生成	纹理和材质优化
		交互式设计	复杂场景重建
智能驾驶	5%	环境感知	决策规划
		驾驶监控	人机交互
智能安防	5%	监控图像识别	多模态事件分析
		多模态身份验证	跨模态智能告警
智慧城市	5%	交通管理	公共安全
		城市规划	城市管理
物流机器人	5%	货物分拣	人机交互
		配送路线规划	动态避障

来源：头豹研究院

中国多模态大模型产业洞察——训练方式

- 多模态大模型的训练旨在学习和理解不同模态之间的关联性，以在多种任务中准确处理多模态信息。训练过程通常分为两个阶段：预训练和指令微调。以下是每个阶段的核心逻辑和设定

多模态大模型训练方式分析



来源：头豹研究院

中国多模态大模型产业洞察——生成能力评估

- 多模态大模型的生成能力评估不仅是技术进步的标尺，更是其商业价值和社会影响的体现。在未来的发展中，如何让生成内容更智能、更贴合用户需求，将成为这一领域的核心竞争力所在

多模态大模型生成能力评估分析（文本生成能力）

能力要求

信息捕捉与语义理解：模型需要从输入的图像中准确提取关键信息，抓住主要语义点（如物体、场景、动作等）

语言生成的连贯性与自然性：模型不仅要生成准确的描述，还需确保语法正确、语义连贯，避免生成不流畅或逻辑混乱的文本

关键挑战

场景复杂性：当图像包含多主体、多事件或背景信息复杂时，模型可能会遗漏重要细节，或生成的描述过于笼统

多样性与创新性：模型在文本生成中是否能展现出多样化的表达，而非单一、模板化输出，这也是衡量生成质量的重要指标

评估维度

场景准确性：生成文本是否与输入内容的语义一致

细粒度：描述是否覆盖了图像的细节，例如颜色、位置和动作

自然性：文本是否流畅、符合语言习惯

创新性：生成的表达是否具有多样性，避免重复

多模态大模型生成能力评估分析（视觉内容生成能力）

能力要求

语义理解与视觉转化：根据文本或其他模态输入，模型需深刻理解描述的语义内容，并通过视觉化表现出来。例如，将“蓝天白云下的一座红色桥”生成符合场景描述的高质量图像。

视觉生成质量：模型需确保生成图像具有高分辨率、清晰的纹理和准确的细节表现，并尽量避免出现失真或不合理的视觉元素。

关键挑战

复杂内容生成：对于多目标、多层次的场景描述，模型需要在保持全局一致性的同时，生成局部细节清晰的视觉内容。

风格与创新：模型是否能生成特定风格化的作品，并保持创新性，例如艺术绘画或科幻场景。

评估维度

视觉质量：图像的分辨率、清晰度、颜色表现是否达到预期。

语义一致性：生成内容是否与输入描述的语义相符。

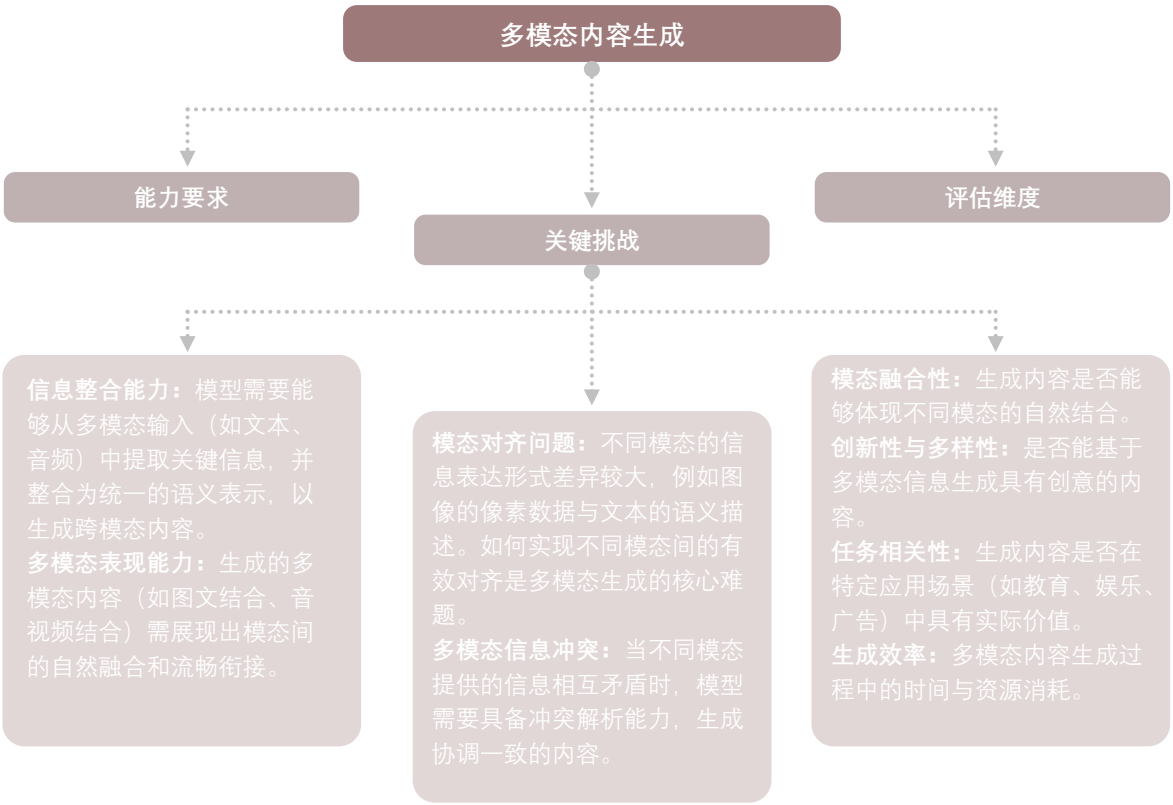
风格匹配：当要求生成特定风格时，是否符合目标风格特征。

多样性：在不同输入下，模型生成的内容是否具有足够的多样性和变化。

来源：头豹研究院

生成能力评估（接上页）

多模态大模型生成能力评估分析（多模态内容生成）



多模态大模型生成能力评估测评方法

能力维度	测评方法
文本生成	**标准化评估：**如BLEU、ROUGE等语言生成指标
	人类评价：让人类评估生成文本的语义准确性与语言流畅性
	任务驱动测试：将生成文本用于实际任务（如场景描述），观察其效果
视觉内容生成	**视觉质量指标：**如FID（Fréchet Inception Distance）评估生成图像质量
	人类评价：通过用户调研判断生成视觉内容的质量与一致性
	多样性分析：测试同一文本输入下生成的不同图像是否足够丰富
多模态内容生成	**多模态数据集测试：**利用涵盖多模态任务的数据集对模型进行全面评估
	用户体验调查：测试用户对多模态内容的接受程度和满意度
	应用案例分析：将生成内容应用于具体场景，测试其效果，如生成的图文广告是否吸引用户点击

来源：头豹研究院

中国多模态大模型产业洞察——技术发展趋势

- 多模态模型的未来发展需要在生成一致性、上下文学习能力、复杂推理技术以及实际应用解决方案上不断创新。多模态幻觉、多模态上下文学习、多模态思维链和LLM辅助视觉推理等研究方向，构成了多模态技术在生成AI时代实现突破的关键路径

多模态大模型技术发展趋势

多模态幻觉

多模态幻觉指的是多模态模型生成的回答与所呈现的图片或其他视觉内容之间出现不一致、不匹配甚至矛盾的问题。这种现象不仅降低了用户对多模态模型的信任度，还可能在关键应用场景中（如医学影像诊断或自动驾驶）引发严重后果。因此，深入研究并解决多模态幻觉问题对于提升多模态模型的可靠性和实用性至关重要。

- **视觉信息Token化方法：**当前的视觉信息Token化方法直接影响模型对视觉内容的理解与表达。例如，视觉特征的提取粒度、编码方法（如Patch Embedding vs. Object-Centric Encoding）以及Token间的关联性都会影响模型生成的多模态对齐效果。改进视觉Token化方法，可以更精确地捕捉图像的语义信息，从而减少多模态幻觉。
- **多模态对齐范式：**多模态对齐范式（如对比学习、交叉注意力机制）直接决定了模型如何融合视觉和文本信息。研究更高效的对齐范式，尤其是探索动态对齐策略以适应复杂场景变化，能够有效提高多模态生成的内容一致性。未来可能需要引入更多基于人类认知过程的对齐机制，以增强模型的感知与决策能力。
- **生成质量评价与纠偏：**通过设计自动化的评价指标（如图文一致性评分）和后处理纠偏模块，可以在生成阶段动态调整模型输出，从而减少幻觉问题。例如，结合图像的视觉显著性区域与语言描述的关键实体，实时检测并修正潜在的不一致之处。

多模态上下文学习

多模态上下文学习通过少量问答样例（即Few-shot Prompting）来提示模型，使其能够泛化到新问题上。然而，与单模态文本上下文学习相比，多模态上下文学习需要同时关注图像和文本内容，并在两者之间建立动态关联，这使其面临更高的复杂性和挑战。

- **上下文提示优化：**针对多模态任务设计更有效的提示策略。例如，如何在Prompt中嵌入视觉内容的核心信息，并将其与语言提示自然结合，从而帮助模型快速适应新问题模式。
- **多模态上下文注意力机制：**当前模型在Few-shot学习中难以充分利用上下文信息，原因在于其对多模态上下文的注意力分布可能存在偏差。研究基于多模态的联合注意力机制，可以引导模型更精准地识别和使用关键上下文信息。
- **泛化与鲁棒性：**Few-shot场景下，样本数量有限，模型往往表现出过拟合或欠拟合的倾向。通过增强多模态上下文的泛化能力（如增加样本多样性、引入对抗训练等），可以显著提升Few-shot学习性能。

来源：头豹研究院

技术发展趋势（接上页）

模态思维链是一种解决复杂任务的推理方法，将复杂问题分解为多个简单的子问题分别求解并汇总结果。相比于纯文本推理，多模态思维链需要在分解问题的同时，融合并对齐多模态信息，从而显著增加了其实现的技术复杂度。

多模态思维链

- **任务分解方法：**在多模态场景中，如何有效地分解复杂任务是关键。例如，设计基于任务语义的自动分解策略，可以将整体任务分割为相互独立但相辅相成的子问题。
- **推理路径规划：**当前研究大多集中于线性推理路径，但在多模态场景中，问题的解答路径可能是非线性的。探索动态推理路径规划算法，能够提高多模态思维链的灵活性与效率。
- **复杂多模态任务数据集：**目前多模态思维链的研究受到数据集的限制。构建覆盖更多领域和任务类型的高质量多模态推理数据集，将有助于推动这一领域的发展。

LLM（大型语言模型）的内嵌知识和强大的推理能力为视觉推理任务提供了新的可能性。通过将LLM与视觉信息结合，可以设计更高效的视觉推理系统，用以解决从简单问答到复杂任务规划的现实问题。

LLM辅助的视觉推理

- **LLM与视觉模型的协作机制：**研究如何构建LLM与视觉模型之间的高效交互框架。例如，通过知识查询接口实现语言模型对视觉信息的动态补充或解释。
- **基于知识的视觉推理：**LLM的优势在于其广泛的背景知识库，研究如何在视觉推理任务中有效调用这些知识，并将其与视觉感知数据结合，从而提高推理的准确性和广度。
- **现实问题解决能力：**设计面向实际应用场景的视觉推理系统，例如智能医疗影像分析、自动驾驶决策支持等。重点在于通过LLM的语义理解能力弥补当前视觉模型的推理不足。
- **可解释性与用户信任：**在多模态推理场景中，结合LLM可以显著增强系统的可解释性。通过生成自然语言形式的推理过程说明，可以帮助用户更好地理解模型的决策逻辑，从而提高信任度。

来源：头豹研究院

中国多模态大模型产业洞察——痛点与挑战

- 多模态大模型在长上下文处理、复杂指令理解、上下文学习与思维链、全面能力整合、安全性以及数据共训等方面仍面临诸多挑战。通过引入分布式推理框架、增强鲁棒性机制、优化模态交互策略等创新方法，能够推动多模态模型更广泛地适应现实应用需求

多模态大模型发展痛点与挑战

01 长上下文处理的局限性

挑战点

- **时间连续性与因果关系缺失：**当前的多模态大模型在处理长视频时，往往无法捕捉视频中复杂的时间轴结构及因果关联。例如，在分析一段包含多个人物互动的长视频时，模型可能忽略前后事件之间的关键因果链条，从而导致理解片面或错误。
- **图文内容的异质性整合困难：**图文内容交错的场景中，文字和图像往往传达不同层次的信息。现有模型在将两者进行语义整合时，可能因模态间表示方法的差异性而失去全局一致性。

应对策略

- **分段式理解与时间建模：**引入分段理解策略，将长视频切分为多个短片段进行独立处理，再结合因果推理模块重构时间序列上的连续性。
- **跨模态联合注意力机制：**针对图文交错内容，设计增强型跨模态注意力机制，使模型能够在图文信息间高效捕捉相关性，提升语义整合的能力。

02 复杂指令理解的障碍

挑战点

- **多条件约束的推理瓶颈：**对于包含多个条件的复杂指令，模型在逐步解析时可能难以保持逻辑一致。例如，面对“从图中找出同时满足红色、圆形且旁边有数字5的物体”这种任务时，模型可能会遗漏某一条件或在步骤间丢失上下文。
- **多步推理能力不足：**当前模型在复杂指令推理中容易陷入局部优化，缺乏全局视野。例如，面对需要结合多模态信息（如图像中物体的空间位置和文字描述）进行逐步推导的问题，模型可能仅完成部分推理链条。

应对策略

- **模块化推理框架：**构建基于分步执行的模块化推理框架，将复杂指令分解为独立的子任务并逐步完成。
- **任务记忆与状态跟踪：**增强模型的任务记忆能力，通过状态跟踪模块记录推理过程中的关键信息，确保逻辑链条完整。

03 上下文学习与思维链的初级阶段

挑战点

- **新情境泛化能力不足：**对于现有模型难以将已有知识灵活迁移到新的情境中。例如，模型可能擅长处理特定领域的任务，但在面对领域外的相似任务时表现大幅下降。
- **思维链的局限性：**当前的思维链推理在多模态场景中应用较少，尤其是在同时涉及图像分析和文本推理的复杂任务时。

应对策略

- **领域无关的上下文提示设计：**引入通用型Prompt设计策略，提升模型在不同情境下对任务的快速适应能力。

来源：头豹研究院

痛点与挑战（接上页）

04 智能体全面能力的欠缺

挑战点

- **能力孤立现象：**当前多模态模型的感知、推理和规划能力相互独立，导致模型无法在实际任务中形成闭环。例如，在无人机路径规划中，模型可能能够感知障碍物但无法结合全局环境做出合理决策。
- **动态任务适应性不足：**面对实时变化的环境，模型缺乏整合多能力应对动态任务的能力。

应对策略

- **全栈融合架构设计：** 开发感知—推理—规划一体化的多模态架构，利用统一的学习机制协调不同模块间的协作。
- **多模态任务模拟环境：** 构建多模态任务模拟平台，训练模型在复杂动态环境中完成任务，从而提高综合能力。

05 安全性的隐患

挑战点

- **恶意攻击风险：**多模态大模型容易受到对抗攻击（如投毒数据或对抗样本），可能被用来生成误导性内容或窃取隐私数据。
- **训练数据偏差导致的歧视性结果：**由于模型依赖于大量的训练数据，训练数据中固有的偏见（如性别、种族偏见）会被模型放大并体现在生成结果中。

应对策略

- **对抗鲁棒性增强：** 通过对抗训练和鲁棒优化方法，提高模型抵御恶意攻击的能力。
- **偏见检测与纠正机制：** 引入训练数据的多模态偏差检测方法，并在模型生成阶段动态校正偏见

06 数据共训的影响不明

挑战点

- **模态间负迁移问题：**在多模态与单模态数据共同训练时，不同模态间可能产生负迁移效应，即一种模态的训练特性干扰另一种模态的学习。
- **交互机制的不确定性：**不同模态间信息交互的具体机制尚未完全明确，可能导致模型在优化时效率低下或结果不稳定。

应对策略

- **分步训练与联合优化策略：**先分别对单模态进行优化，再在多模态融合阶段引入模态自适应权重，以减少负迁移。
- **模态交互研究：**深入探索不同模态间的特性差异和交互机制，为数据共训提供理论指导。

来源：头豹研究院

中国多模态大模型产业洞察——未来展望

- 通过不同策略构建多模态输入输出空间、设计对齐架构与训练策略、进行全面可靠评测，以及将输入输出扩展框架应用于具身智能场景，最终目标是构建具有一般性能力的世界基座模型

多模态大模型未来展望分析（将离散或连续模态表示与文本空间结合）

主流策略

- **构建混合空间：**以连续形式整合模态信息，并与离散文本对齐。
- **统一离散表示：**使用离散化的形式建模多种模态，确保生成与理解任务的统一处理。

优势对比

- **混合空间：**在理解类任务上表现出色，但在多模态生成任务中支持有限。
- **统一离散表示：**适配生成与理解任务，在理解任务上的效果略逊。

问题 1：如何通过不同方式，将离散或连续模态表示与文本空间结合，构建多模态的输入输出空间？

未来展望

- 学术社区正在积极探索更高效的模态表示方法和编码器，旨在兼顾生成和理解任务。
- 离散与连续模态表示的选择和优化目标密切相关，两者在设计思路和训练方法上存在互相借鉴和促进的可能性。

多模态大模型未来展望分析（设计模型架构与训练策略）

问题 2：如何设计模型架构与训练策略，完成多模态输入输出空间的对齐？

01

❑ **模型架构**

- 根据输入空间设计对应的输入端对齐模块，根据不同模态的输出形式设计输出对齐模块。
- 引入额外的内部模块，以更好地建模跨模态的交互。

02

❑ **训练策略**

- 经历预训练和指令微调两个阶段，前者用于对齐多模态表示，后者学习多模态场景下的指令遵循能力。
- 根据应用和优化的目标选择和混合训练数据，确保数据的丰富度、质量和规模。
- 根据模型架构设定合适的参数，如可训练参数等。

03

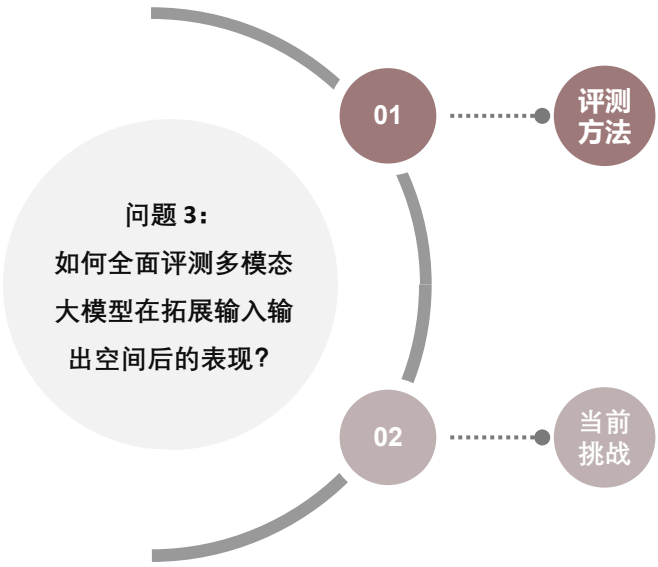
❑ **挑战与机遇**

- 模型架构的设计存在多种选择，且不同设计间存在tradeoff，需要通过实证性分析获取相对较优的设定。
- 训练策略的选择同样重要，需要平衡数据丰富度、质量和规模，以及模型架构的复杂性。

来源：头豹研究院

未来展望（接上页）

多模态大模型未来展望分析（全面评测多模态大模型）



构建理解类和生成类的Benchmark：

理解类Benchmark：设计一系列测试集来评估模型对不同模态数据的理解能力，例如通过提供带有文本描述的图像，要求模型准确地识别出图像中的对象或场景，并能回答关于图像的问题。

生成类Benchmark：构建测试集以评估模型根据给定的多模态输入生成相应输出的能力，比如基于一段文字和一张图片生成一个连贯的故事或者描述。

评估指标与实际用户感知的差异：传统的评估指标可能无法完全反映模型在真实世界中的性能。例如，高准确率并不总是意味着更好的用户体验。这是因为用户不仅关心模型的准确性，还关注其响应的速度、相关性和自然性等方面。

任务与需求的差异：现有的一些评测标准可能过于理论化，未能充分考虑用户的具体需求和使用场景。例如，一个在学术论文摘要生成任务中表现优异的模型，在撰写创意故事方面表现不佳。

评估指标与人类偏好的差异：即使是在相同的任务上，不同的用户也可能有不同的偏好。因此，单一的评估指标难以全面衡量模型的好坏。

多模态大模型未来展望分析（具身智能场景下）

具身智能的深化

环境感知的增强

- **多传感器融合：**在具身智能中，智能体需要具备高度发达的感知能力。这可以通过集成多种传感器（如摄像头、麦克风、触觉传感器、激光雷达等）来实现，从而全面获取环境信息。
- **跨模态感知：**不同模态的数据（如视觉、听觉、触觉）可以相互补充，提高感知的准确性和鲁棒性。例如，结合视觉和听觉信息可以更准确地定位声源位置。

动作决策的优化

- **强化学习：**利用强化学习算法，智能体可以根据环境反馈不断优化其动作决策。通过试错和奖励机制，智能体可以学会在复杂环境中完成任务。
- **策略规划：**除了简单的运动控制，智能体还需要具备高级的策略规划能力。例如，能够制定长期目标并分解成短期任务，逐步实现目标。

输入输出空间拓展

多模态信号的融合

- 统一表示：设计一种通用的数据表示方法，将不同类型的输入信号（如文本、图像、声音、视频、电磁波等）转换成统一的格式，便于模型处理。
- 多模态预训练：通过大规模多模态数据的预训练，使模型能够学习到不同类型信号之间的关联和依赖关系，提高其综合处理能力。

通用接口的设计

- **模块化架构：**设计一个模块化的系统架构，每个模块负责处理特定类型的输入信号，并输出相应的结果。模块之间可以通过标准化接口进行通信，便于扩展和维护。
- **灵活配置：**允许用户根据具体应用场景灵活配置输入输出模块，以满足不同任务的需求。

来源：头豹研究院

业务合作

会员账号

可阅读全部原创报告和百万数据，提供PC及移动端，方便触达平台内容

定制报告/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

对产业及细分行业进行现状梳理和趋势洞察，输出全局观深度研究报告

招股书引用

研究覆盖国民经济19+核心产业，内容可授权引用至上市文件、年报

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

行研训练营

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

报告作者

袁栩聪
首席分析师
oliver.yuan@leadleo.com

陈庆民
行业分析师
qingmin.chen@leadleo.com

业务咨询

- **客服电话：400-072-5588**
- **官方网站：www.leadleo.com**

深圳办公室

广东省深圳市南山区粤海街道华润置地大厦E座4105室

邮编：518057

上海办公室

上海市静安区南京西1717号会德丰国际广场 2701室

邮编：200040

南京办公室

江苏省南京市栖霞区经济开发区兴智科技园B栋401

邮编：210046

方法论

- 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明