『弈衡』多模态大模型评测体系白皮书

（2024 年）

发布单位：中移智库

编制单位：中国移动通信研究院

目 录

前 言

随着人工智能技术的迅猛发展，它已成为全球科技革命的核心驱动力。特别是 2017 年Transformer模型提出后，人工智能大模型以超凡的性能和无限的可能性，迅速成为科技界的焦点。2023 年初，GPT-4[1]的问世更是在全球范围内引起了巨大反响，标志着大模型技术首次进入公众视野[2]。

随着大模型技术的不断演进，其处理能力已从单一的文字信息扩展至图像、语音等多模态数据，多模态大模型进入快速发展阶段。它们不仅在日常生活中的辅助作画、图片解读等场景中展现出应用潜力，更在视频数据分析、多目标识别等生产领域发挥着重要作用。目前典型的多模态大模型有国外的GPT-4Vision、Gemini，国内的文心一言、讯飞星火、智谱清言等[3]。这些大模型算法各异，在不同的任务场景下各有优劣，如何对这些多模态大模型开展客观、科学的评测，评估特定任务场景下的最优选择，对大模型的研发迭代以及应用落地都具有重要意义。

相比于语言类大模型，多模态大模型具备对文本、图像、视频和音频等数据进行综合处理的能力，在生产生活领域中具有广泛的应用前景。同时，多模态大模型评测面临评测数据更多样、评测任务更丰富、评测方式更复杂、评测成本更昂贵等挑战。如何应对上述挑战，构建全面、客观的多模态大模型评测体系，成为业界关注的热点问题。目前，部分业界企业和研究机构，如微软、谷歌、智源研究院、上海AI实验室、腾讯优图实验室、厦门大学、南洋理工大学等，发布了相关论文、评测报告，从性能、参数量等维度对业界主流多模态大模型进行了评测，并基于评测结果形成了榜单，如MMbench、MME等。为提升多模态大模型的实际应用效果，推动大模型与生产生活的快速结合，有必要从用户视角出发，构建一套客观全面、公平公正的多模态大模型评测体系。

中国移动技术能力评测中心作为中国移动的第三方专业评测机构，联合业界权威机构、头部企业，攻关多模态大模型评测难点技术，基于前期评测数据和评测经验积累构建“弈衡”多模态大模型评测体系，并编制本白皮书，旨在为多模态大模型的评测场景、评测指标、评测方式等提供参考基准，为评测数据和评测工具的构建提供参考指导。本白皮书聚焦于文生图、图生文、图文理解等各类应用场景，深入分析多模态大模型的应用需求，系统总结行业典型评测体系，并创新地提出“弈衡”多模态大模型评测体系，助力大模型技术与行业应用的深度融合。具体包括如下四方面内容：一是总结梳理多模态大模型的应用需求与评测挑战，将评测需求划分为识别、理解、创作、推理四种任务；二是广泛调研业界多模态大模型评测

技术和评测体系，从评测方式、评测维度和评测指标等方面进行分析总结；三是提出“弈衡”多模态大模型“2-4-6”评测框架，针对图文双模态大模型，详细阐述基础任务和应用任务两大评测场景，评测指标、评测数据等四大评测要素，以及功能性、准确性、交互性、安全性等六大评测维度；四是针对多模态大模型演进趋势，展望评测技术重点方向。

未来，中国移动将持续跟进多模态大模型发展，不断优化“弈衡”多模态大模型评测体系，与业界合作伙伴一道，共同打造评测产业标准化生态，推动多模态大模型产业成熟和落地应用，为AI+赋能千行百业贡献力量。

01 多模态大模型评测背景

1.1 多模态大模型发展现状

随着人工智能技术的快速发展，多模态大模型对图像、文本、视频和音频等信息的综合处理能力不断增强，其跨模态理解能力、高精度识别与理解能力、强大的泛化能力、丰富的表达能力、增强的交互体验，进一步推动了人工智能技术在各行业的广泛应用[4]，成为推动产业升级与生产力变革的强大引擎。目前，多模态大模型正在迅速融入到各行业的应用场景中，服务于生产生活的各方面。多模态大模型在多个领域的典型应用如下：

行业	领域	应用
企业应用	内容创作与审核领域	用于图片创作、图片内容理解、图形合成修改等任务。
	教育科技领域	利用图文数据为教育领域提供智能化支持。
	金融风控领域	根据签字等图像数据辅助金融机构提高决策效率。
	医疗健康领域	利用内置摄像头进行辅助诊断，协助医生提高医疗效率。
	智能制造领域	进行缺陷图片检测，助力工厂实现智能化生产、降本增效。
	软件开发领域	根据现有图形界面，辅助提升开发人员的软件开发效率。
	市场分析领域	帮助企业洞察市场动态，优化产品、提供更加安全的服务。
	法律领域	用于文书识别等法律相关任务，降低法律服务成本。
	媒体与娱乐领域	为画师、视频创作者等相关从业者提供创意灵感，提高创作效率。
	人力资源领域	实现人脸识别等人力资源智能管理功能。
	客服领域	应用于智能客服助手等任务，实现图形理解，提高客服效率。
	公共服务领域	利用摄像头等终端识别提高政府服务效率，优化公共资源配置。
个人应用	旅游领域	提供景点照片匹配等个性化的旅行建议和服务。
	个人金融业务领域	用户人脸识别、收支明细预测等个人金融业务。
	教育辅导领域	针对题目进行智能搜索、解答等教育辅导工作。
	数据搜索领域	实现拍图识别、搜索等智能搜索功能。
	图像修复领域	针对老照片、不完整照片等图像进行智能修复与补全。

多模态大模型中，图文双模态大模型发展尤为迅速，它在处理图像与文本及其复杂交互关系上取得了显著成果，为内容创作、信息检索、智能决策等多个应用场景带来了革命性的变化，应用范围不断拓宽，影响力日益增强。鉴于图文双模态大模型的重要性和广泛应用前

景，本白皮书主要聚焦图文大模型评测，深入分析评测需求以及面临的问题和挑战，系统讨论关键评测技术，旨在为业界提供一套科学、系统、可操作的图文双模态大模型评测框架，促进技术的健康发展与广泛应用，进一步加速人工智能技术在各行各业的深度融合与创新实践。

1.2 评测需求

图文大模型相较于传统视觉模型和大语言模型，在图像识别、图文深度理解与推理以及图片创作等复杂图文交互任务中展现出了显著的优势。由于不同图文大模型在处理应用场景时各有专长，因此选择适合各行业特定应用需求的模型变得尤为重要。在对图文大模型进行评测时，需面向不同任务类型，从各个维度进行综合全面的评测，以评估图文大模型的真实性能和用户体验。目前，对图文大模型的评测需求包括但不限于以下几类任务：

识别类任务：识别类任务主要是指对图片中的特定事物进行识别、计数等工作。识别类任务主要可分为基础任务和应用任务两类。其中基础任务包含实例识别、颜色识别、手势识别、目标检测等基础场景；应用任务则包含商品识别、垃圾满溢识别、道路安全识别、智慧养殖等更加复杂的端到端场景。识别类任务作为目前最广泛应用的任务之一，是衡量图文大模型性能的重要场景，具有极高的评测价值。在评测识别类任务时，需着重关注模型的准确性、鲁棒性、实时性和泛化能力等指标。

理解类任务：理解类任务主要是指针对输入图片进行内容理解，并回答对应问题。理解类任务也可分为基础类及应用类两种。基础类理解任务侧重于考察图文大模型的通用能力，而不过分强调某一特定应用场景中的实际能力。常见的基础类任务包含场景理解、实例属性、空间关系、字幕匹配、图像质量分析等底层核心场景；而应用类任务则着重考察图文大模型在专一领域的实际能力，与目前具有智能化需求的场景结合更加紧密，如活体检测、人像属性、人脸属性、口罩检测、舞蹈艺考评分等。理解类任务相较识别类任务，不仅仅考察模型对某一特定事物的特征识别能力，更要求图文大模型对图像整体场景及各事物之间关系进行精准把控，并依据提问内容进行匹配跟踪，相较识别任务难度更大。在评测理解类任务时，需着重关注模型的准确性、上下文感知、通用性与专一性以及语义一致性等指标。

创作类任务：创作类任务主要是指通过给定的文字或图像提示信息进行图片创作或图像修改。常见的创作类任务包含图像生成、图像风格转换、图像合成等，图文大模型根据要求生成相应图片，图片需要在美观上符合人类需求，在逻辑上符合基本的事物原理，在匹配度上完全实现提示词或提示图片中的内容要求。创作类任务综合考察了图文大模型的文字图像理解和图像创作能力，是目前应用最为广泛关注度最高的任务之一。在评估创作类任务时，