『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，

需着重关注模型的生成质量、内容匹配度、多样性和创新性等各项指标。

推理类任务：推理类任务主要是指结合输入的图像和文本信息，进行逻辑推理、归纳推理或演绎推理等。推理类任务着重考察图文大模型对图片内容中涉及的各类逻辑知识进行理解、推理和解答的能力，是对图文大模型内在核心思考能力的真实反馈。常见的推理类任务包含下一张图像预测、代码编写、数学推理等。这些问题需要精细的思考及相应的专业知识训练才可作答，对普通人而言也具有较高难度，是对图文大模型核心能力的重点考察方向。在评测推理类任务时，需着重关注模型的推理准确性、推理深度、专业知识应用、逻辑一致性和可解释性等指标。

1.3 评测问题与挑战

图文大模型具有任务多样、模型复杂等特点，传统小模型的评测方式无法完全评估图文大模型在特定场景下的实际使用效果，需要针对图文大模型评测的问题与挑战进行深入分析，并不断迭代评测方法，以更好地促进图文大模型的良性发展。

首先，图文大模型的高泛化性对评测任务选取提出挑战。

图文大模型最突出的特点就在于任务适用性广，一个图文大模型往往可以在识别、理解、创作、推理等各类任务中实现较好的性能。但是，任何模型都具有局限性，目前某些任务图文大模型尚无法解决。因此，如何选择合适的评测任务场景，既能满足业务需求，又不超越模型现有能力，便成为了一项重要的考虑因素。为全面评价模型能力，需要对行业痛点和图文大模型研究现状具有充分的了解，从而制定更为全面、合理的评测任务。

其次，图文大模型的高复杂度对评测数据构建提出更高要求。

图文大模型参数量极大，内部极为复杂，相关训练原理和训练数据分布难以获取，这就导致图文大模型评测数据构建难度大。人类视角下的题目难易与模型视角下的不一定一致，比如绘制人手对于人类来说比较简单，而对于目前的图文大模型则较为困难。如何梯度性设置测试用例，以合适的低中高难度比例对模型展开全面测试，真实反馈出模型性能，是一项需要解决的难点问题。需要针对各个任务领域，对业界典型图文大模型进行大量验证，不断迭代优化测试用例的设置，才能构建更为合理的评测数据。

再者，图文大模型评价结果的客观性也需要重点考虑。

图文大模型的任务设置和输出结果丰富多样，这其中既有计数、识别等易客观评测的基础任务，也有图像生成、风格转换等创作类任务。后者往往需要通过主观评价的方式对图文

大模型的对应能力进行测试评估，这对评价人员技术水平提出更高要求。因此，需要制定好主观评测体系基准，尽可能缩小不同评价人员带来的随机程度，以更加客观的方式实现对图文大模型创作能力的公平评价。

综上所述，随着图文大模型的快速发展，相关评测体系也需要不断迭代优化，着力解决行业痛点，积极应对评测挑战，以客观全面、公平公正、用户视角为评测基本原则，对图文大模型展开合理测试，更好地促进图文大模型的良性发展。

02 多模态大模型评测技术

近年来图文大模型发展迅猛，各大企业和研究机构对图文大模型评测体系进行了深入探索，并发布论文、技术报告、评测榜单等各类研究成果[5]。本章参考谷歌、微软、智谱研究院、上海AI实验室、腾讯等企业及研究机构的成果，对主要评测方式、典型评测维度和常见评测指标等关键评测技术进行梳理与总结。

2.1 主要评测方式

图文大模型的评测方式主要包括客观评测和主观评测两种。

客观评测是指利用客观评价指标对图文大模型的生成结果进行定量评估，常见的客观评测方式有准确率、召回率、模型推理时间、可支持图片分辨率等。客观评价指标种类多样，可以从各个维度对图文大模型的生成结果进行准确、全面、公平的评价，是对大模型进行评测的主要方式。此外，由于客观评测指标可由计算机直接计算得到，因此能够通过自动化脚本实现批量测试，大幅提高评测效率和规模[6]。

主观评测是指通过人工打分的方式对图文大模型的预测结果进行评价，主要应用于创作类任务中，如图片生成、风格变换、图像合成等[7]，这些测试用例没有明确的标准答案，因此无法以合适的客观指标进行完整评测。主观评测相较客观评测更加灵活，更能真实反映用户视角下的模型能力，但存在评价结果不稳定、难以大规模实施等问题，因此，需要针对具体任务制定合理的主观评测方法。

2.2 典型评测维度

依据谷歌、微软、上海AI实验室、腾讯等企业和研究机构的研究，图文大模型的典型评测维度，可分为模型性能、模型泛化能力、模型鲁棒性和模型一致性四个方面[8]。

模型性能评测是图文大模型的核心维度，主要评测图文大模型对图像和文字的识别能力、

理解能力、推理能力，如生成的图像或文字结果相较正确答案的准确度。常用性能评测指标有图像识别准确率、与提示词的匹配度等。

模型泛化能力评测主要评测图文大模型在多任务上的适配能力，该评测维度可以反映出大模型在实际部署中的泛化性。常见的评测方式为针对大模型未训练的场景和图文数据，测试模型的应用效果。

模型鲁棒性评测主要评测模型应对各类干扰时的鲁棒性及可靠性，如对输入图片施加肉眼不可见的噪声和数据扰动，验证对抗攻击情形下模型应用效果。

模型一致性评测主要评测在面对不同规模解空间的问题时，图文大模型能否在相同知识点上给出一致答案的能力，如模型生成的图片描述是否与相同知识点的判断结果一致。

2.3 常见评测指标

目前，各类图文大模型评测指标从不同角度对模型性能进行了综合评判，常见指标有准确率、F1 值、BLEU、IS指标、CLIP相似度、PSNR、SOA、CIDEr、mAP、IoU、FID、SSIM、RP、碳足迹等[9]。

指标	描述
准确率	Accuracy，计算图文问答题目中预测结果正确的比例，是最常用的客观指标
F1 值	F1 Score，综合考察图文问答题目中预测结果的精确率（Precision）和召回率（Recall），兼顾图文大模型预测结果的正确样本比例和查全比例
BLEU	评价图生文的文本质量，比较生成文本与真实答案间的重叠程度
IS 指标	Inception Score，利用分类模型评测生成图片的类别确定性和类别多样性
CLIP 相似度	利用 CLIP 大模型的文本和图像编码器针对图片中关键物体进行质量判定
PSNR	峰值信噪比，评价图文大模型生成图片的像素质量和清晰度
SOA	衡量生成的图像中是否符合文本描述中的各对象类别，考察文本类别还原度
CIDEr	针对图像描述任务，评价描述结果与人类真实描述间的相似度
mAP	mean Average Precision，反映图文问答题目中，预测结果在所有召回率水平下的平均准确率
IoU	Intersection over Union，衡量图像中指定物体的预测框与实际边界框的重合程度
FID	Fréchet Inception Distance，用于评估文生图任务中生成图像和真实图像之间的相似性的指标
SSIM	结构相似度，评价文生图任务中生成图片与标准正确图片之间的相似度
RP	全称 R-precision，衡量文生图任务中文本描述和生成图像之间的视觉语义相似度
碳足迹	计算模型训练、推理阶段消耗电力的二氧化碳排放量

除以上提到的各类常用指标外，部分评测还针对图文大模型在业务中的实际应用场景，选取更有针对性更能反映业务性能的其他指标，如召回率、多轮对话轮次等。

03 典型多模态大模型评测体系

近年来，随着图文大模型的快速发展，多家科研机构及企业提出了一系列大模型评测体系，如上海AI实验室的MMBench、华中科技大学的OCRBench、智源研究院的智源评测体系、微软的LLaVA-Bench、希伯来大学的VisIT-Bench、腾讯的SEED-Bench等，这些体系从多个方面对图文大模型进行了评测，具有较高的参考和应用价值。本章将对典型评测体系进行概括介绍。

- **MMBench[10]**

MMBench是上海人工智能实验室于 2023 年 8 月提出的多模态大模型评测体系，相关研发人员针对当下评测方式存在的主观评测多样性差、客观评测任务覆盖少等问题，提出了逐渐细化的评测任务设置和CirularEval评测方式。具体来说，在评测数据构建上，MMBench从三个维度设计了大量单选题，第一级是感知与推理能力，第二级包含细粒度感知、逻辑推理、相关性推理等六项能力，第三级包含目标定位、图像质量、社会关系等二十项能力。在评测方式上，针对当前大模型指令跟随性不完善的问题，利用ChatGPT进行辅助评测，并将问题选项进行环状重排，从而更好地反映大模型的真实性能。

- **OCRBench[11]**

OCRBench是华中科技大学联合其它机构于2024年2月提出的多模态大模型评测体系，该体系针对OCR领域的常见任务和典型数据集，对Gemini、GPT-4V等十四个多模态大模型进行了评测。具体来说，OCRBench聚焦于多模态大模型的OCR能力，针对文字识别、场景文本视觉问答、文档视觉问答、关键信息抽取和手写数学表达式识别这五种任务设计专门的提示词，并选取COCOText、STVQA等二十七个主流开源数据集进行测试验证。

- **智源评测体系[12]**

智源评测体系是智源研究院于 2024 年 5 月发布的大模型评测体系，该体系对国内外一百四十余语言及多模态大模型进行了全方位测评。在评测任务设置上，智源评测体系针对图片问答、文本生成图像、文本生成视频、图像文本匹配等任务进行了测试，主要考察了模型

的理解和生成能力。在评测数据选取上，该体系选取了COCO、Flickr30k等主流开源数据集。在评价指标筛选上，该体系从主观和客观两个维度针对各个任务进行了单独设计，客观指标主要选取了准确率、召回率、FID、CLIPScore等常见指标，主观指标则采取人工打分的形式进行模型评价。

- **LLaVA-Bench**[13]

LLaVA-Bench是威斯康星大学、微软等研究团体于 2023 年 4 月提出的多模态大模型评测数据集，包含LLaVA-Bench（COCO）和LLaVA-Bench（野外）两个数据集。它聚焦于视觉指令跟随任务，着重考察图文大模型的对话、图片描述及复杂推理能力，在结果评定上采用准确率作为评测指标，并利用GPT-4 辅助进行评定，综合评测图文大模型在室内场景和室外场景下的性能。

- **VisIT-Bench**[14]

VisIT-Bench是希伯来大学、谷歌等研究团体于 2023 年 8 月提出的图文大模型评测基准，包含 592 个带人工标注的图文问答对，并具有多达 70 个提示词类型，综合考察了图文大模型的识别、场景理解、家装设计、图表解释等各类能力。在模型评测过程中，VisIT-Bench利用GPT-4 对图文大模型性能进行评定，并利用人工辅助验证的方式增强结果的可信度。

- **SEED-Bench**[15]

SEED-Bench是腾讯人工智能实验室于 2023 年 7 月提出的多模态大模型评测基准，包含了 19000 道选择题，并将测试用例分为多个难度层级，涵盖了场景理解、实例属性、图表理解等十二个评测维度，考察大模型对图像文本的理解和创作能力。SEED-Bench采用自动化评测方式，利用客观评价指标对图片创作等主观任务展开评测。具体来说，针对文本创作类题目，SEED-Bench通过计算模型对各个人工标注选项的困惑度来获取模型最佳预测结果，再通过最佳预测结果和正确选项计算模型准确率；针对图片创作类题目，通过计算模型生成图像与各人工标注选项之间的CLIP相似度来获取模型最佳预测结果，再通过最佳预测结果和正确选项计算模型准确率。

- **ConBench**[16]

ConBench是北京大学联合字节跳动于 2024 年 5 月提出的多模态大模型评测基准，它弥补了多模态大模型一致性评价的空白。对于同一个知识点，不同的提问方式可能会获得不

一致的答案。为了评估模型的一致性，ConBench从四个高质量的多模态基准数据集中手动选择 1K张图片：MME、SeedBench、MMBench和MMMU，每张图片包含三个判别式问题（判断题、选择题与限制性问答题），以及围绕相同知识点的生成式 prompt，评测知识点分为观察能力、复杂推理和专业知识三个难度层级，模型的一致性由判别和生成两个角度体现，其中，Caption和三个判别式回答之间的一致性通过GPT/GPT-4 自动判断。

这些评测体系从不同的侧重点对图文大模型的准确性、参数量等方面进行了评测，在评测指标选取、评测数据构建、评测工具平台搭建等各个角度进行了大量研究，推动了图文大模型评测体系的发展。但是，在图文大模型的实际应用中，用户也会考虑功能性、交互性、安全性等因素，当前评测体系对于这些需求的考量仍略显不足。

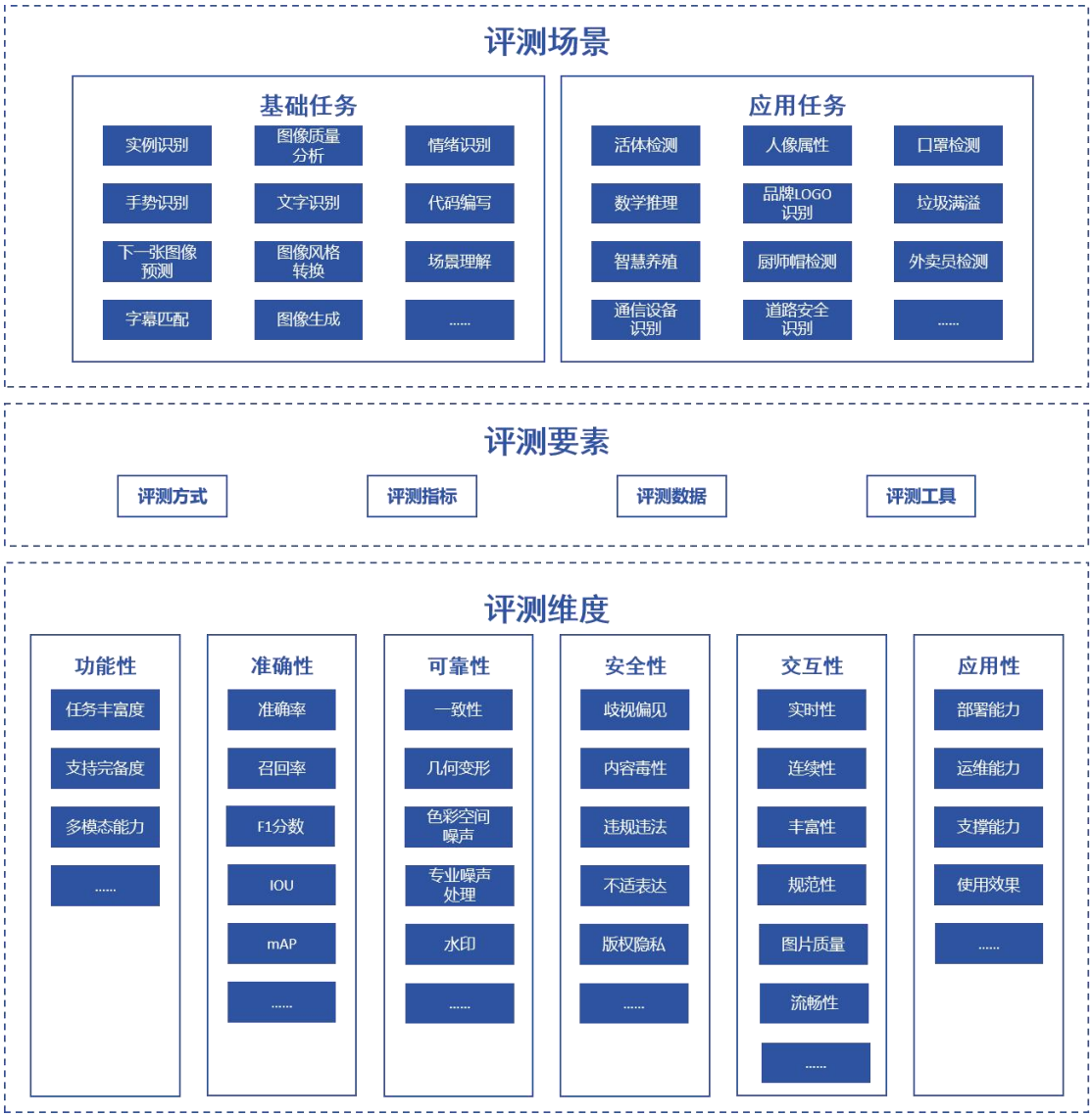
04 “弈衡”多模态大模型评测体系

随着人工智能技术的蓬勃发展，图文大模型的应用场景日益广泛，展现出卓越的泛化与适应能力。为全面考量图文大模型的图像和文字综合理解能力，我们需遵循客观全面、公平公正和用户视角的评测原则对图文大模型开展评测。客观全面是评测的基本要求，是指要以严格的标准和流程进行评测，从评测数据集、评测任务、评价指标和评测工具四个方面进行图文大模型评估。公平公正是评测的根本要求，要求测试者给予所有参测模型公平的机会和条件，以公开透明的方式评测全过程。用户视角是评测的价值要求，要求从用户的需求、期望和体验角度开展评测，分析图文大模型的实际应用价值。

本章基于上述三个原则提出“弈衡”多模态大模型评测体系，旨在为图文大模型的技术创新和应用实践提供坚实支撑，为人工智能领域的持续发展注入新的活力，助力其更好地服务社会，满足生产生活的多样化需求。

4.1 整体框架

中国移动技术能力评测中心构建“弈衡”多模态大模型评测体系，采用“2-4-6”层级架构，包含 2 类评测场景、4 项评测要素以及 6 种评测维度，从功能、性能、可靠性、安全性、交互性等方面对图文大模型的图文理解能力进行全方位评测。详细评测框架如下图所示：



随着大模型技术的不断演进以及应用的日益广泛，图文大模型的评测需求也将不断变化。为了全面、客观、公正地评价图文大模型的能力，后续我们会对“弈衡”多模态大模型评测体系进行持续更新和完善，如任务设置、数据集构建、评价指标设计、评测平台搭建等等，以促进图文大模型技术发展和行业应用。

4.2 评测场景

在对图文大模型进行评测时，需要根据不同的任务类型逐一评判大模型在各个特定场景下的表现优劣。“弈衡”多模态大模型评测体系综合考虑现有的图文大模型应用场景，依据任务性质、技术难度与复杂度、应用场景以及知识要求，将图文大模型评测任务分为基础任

务和应用任务两类。

- **基础任务**

基础任务主要关注图文结合的各类通用任务场景，这些场景适用性广，可为后续的应用任务提供方法参考和对标基线。基础任务主要包含识别、理解、创作和推理四大类，每一大类又下辖大量基础子任务，典型场景如下：

任务		描述
识别	实例识别	识别图像中的特定实例，包括特定对象的存在或类别，评估模型的对象识别能力。
	实例计数	计算图像中特定对象的数量，理解所有对象并成功计数所引用对象的实例。
	情绪识别	侧重于识别和解释图像中人脸所表达的情绪，评估模型理解面部表情并将其与相应情绪状态相关联的能力。
	手势识别	根据输入图像识别手势含义，评估模型对人手特征的理解。
	文字识别	回答关于图像中文本元素的相关问题，考察多模态模型对各种类型文本的识别及上下文理解。
理解	场景理解	强调图像中的全局信息，需要整体理解来回答有关整个场景的问题。
	字幕匹配	针对图片，选择最符合图片内容的文字描述，考察文字及图片内容理解。
	图像质量分析	根据图片是否模糊、光照是否正常、是否存在遮挡等因素分析图像质量
创作	图像生成	根据给定提示生成逼真且视觉连贯的图像的能力，要求模型理解创建可信图像所需的视觉元素、关系和组合规则。
	图像风格转换	针对文字要求，对指定图片进行风格变换，要求模型把握图片内容及风格特点。
	图像合成	根据文字要求，对多张图像进行融合后生成新图像
推理	代码编写	理解图片中代码内容并回答相关问题，考察模型对代码的理解和编写能力。
	下一张图像预测	根据给定的图像序列，判断缺失图片内容。

基础任务是构成图文大模型应用场景的根本，针对基础任务进行大模型评测，可以很好地反映图文大模型的多任务泛化性，具有重要的研究意义。因此，在评估图文大模型前，先对基础任务进行定义和梳理是极为重要且不可或缺的。

- **应用任务**

除各类基础任务外，一个合格的图文大模型还应在各类特定领域和场景下实现卓越性能，因此，大模型评测时应综合考量模型在应用任务中的识别、理解、创作和推理等表现，确保其在实际生产生活中可用、好用、易用。典型场景如下：

任务		描述
识别	人流量统计	对特定区域或场景内的人员数量进行实时统计
	品牌 LOGO 识别	根据品牌的 LOGO 图片进行识别，判断所属企业并给出企业的相关信息。
	垃圾满溢	判断图片中的垃圾桶是否存在垃圾桶，以及垃圾桶是否存在满溢。
	智慧养殖	针对猪、鸡等各类家畜进行识别与计数，辅助进行养殖管理。
	厨师帽检测	对后厨是否有人未正确佩戴厨师帽进行识别，以规范商家卫生安全。
	外卖员检测	针对各类场景下是否存在外卖员进行检测，服务于小区安防、外来人员管控等。
	通信设备识别	针对图片中的各类通信设备进行识别，服务于硬件厂商及运营商等管理人员。
	道路安全识别	对车辆违停、路面塌陷等相关情况进行识别，从而保障交通安全。
理解	活体检测	根据输入的真实人脸图片，以及翻拍、面具、高清屏、3D 头模等伪造活体进行判断，以检验多模态大模型在人脸安全方面的识别能力。
	人像属性	针对输入的图片，回答关于图片中人像属性的各类问题，如着装、动作、性别等。
	口罩检测	判断图片中是否有人未正确佩戴口罩，检验模型对人脸及口罩佩戴的识别能力。
推理	数学推理	针对图片中描述的图形、逻辑等数学问题进行回答，检验模型对数学图形和逻辑的理解推导能力。
创作	艺术创作	根据图文提示进行艺术创作，探索新的艺术风格和表现形式，拓展艺术创作的边界。
	游戏角色设计	根据图文输入提示，辅助或自动化完成游戏角色的设计过程，包括角色的外观、动作、服饰、武器等等。

与基础任务相比，应用任务场景更加固定，但其难度更大，涉及更高层次的技术能力，可以反映图文大模型面向具体领域和特定行业场景的泛化能力。

4.3 评测要素

“弈衡”多模态大模型评测体系的评测四要素包括评测方式、评测指标、评测数据和评测工具。

4.3.1 评测方式

重点考虑测试样本构造和测试结果判断两个方面。在测试样本构造方面，全面考虑零样本（zero-shot）、单样本（one-shot）、少样本（few-shot）以及提示工程（prompt engineering）等评测方式。在测试结果判断方面，根据是否有标准答案，使用客观评测或主观评价进行评定。

- **测试样本构造方式**

图文大模型泛化性强，可适用任务广，被用于解决各类实际问题。在实际应用中，经常存在数据未包含在预训练数据中的场景[17]，这就要求图文大模型在零样本学习的条件下依旧保持优秀性能。而对于人脸识别等常见任务，图文大模型已经经历过多次迭代和训练，只需基于少量样本进行简单优化即可在特定业务场景实现良好性能，这属于少样本任务。此外，当前研究表明，提示词的设置会极大程度地影响模型效果，针对同一内容的不同提问方式，可能导致模型出现巨大的性能差异。“弈衡”多模态大模型评测体系综合考虑上述三种数据构造方式，以及提示工程的研究内容，综合评测模型性能，探索图文大模型在各种任务场景下的最优效果，以满足实际业务应用需求。

零样本：零样本任务是指模型在训练阶段完全没有接触过测试场景及测试任务相关的图文数据，模型需要针对全新场景完成预测任务。这类任务设置不需要模型进行针对性调优，直接考察了图文大模型对新知识的理解和泛化能力，具有极高的应用价值。

单样本：在单样本任务中，图文大模型只能在训练阶段接触到一个与实际部署任务相关的图片或文字样本，模型需要提取这一个样本中的核心特征，并将其应用于其他同类任务样本中。该任务设置相较传统多样本任务难度更大，更加考察大模型的核心特征提取能力。

少样本：少样本任务是指图文大模型在训练阶段可以接触到少量目标任务的图文样本，通常可微调样本数量在几个到几十个之间。相较于单样本，少样本任务难度相对更低，但实际应用价值更高。在图文大模型的实际部署应用中，模型需针对各类具有差异性的业务数据完成预测，因此，被测图文大模型是否可利用少量典型数据对模型进行微调提升模型性能，即是否可以在少样本任务设置下实现较好的性能表现便至关重要。

提示工程：图文大模型的任务数据通常包含图片及文字两类，相较大语言模型问题设置难度更大。研究表明，针对同一内容的不同提示词会导致大模型产生完全不同的结果。因此，在对图文大模型进行评测时，需结合实际业务场景进行广泛调研，构建更加合理有效的图文指令，以更好地评测特定业务场景下模型的生成能力和潜力。

- **测试结果判断方式**

在对图文大模型进行评测时，选择合适的评估指标至关重要。为此，应根据不同任务的特性定制设计评估指标，结合客观和主观两种评价方式。

对于问题有明确标准答案的任务，如口罩检测、人群计数等，应当主要使用各类客观指标进行评测，如准确率、F1 值、mAP、BLEU等，这些指标能够比对模型预测结果与真实标注，并利用各类公式完成测试结果评判。利用客观指标筛选可以更加公平、合理、全面地

评价各大模型性能。

对于没有固定标准答案的任务，如图像创作、风格迁移等创作类任务，客观指标便很难全面综合地对模型性能进行评估，此时就需要利用人工打分等主观评判方式。主观评判需要建立一个由三名及以上领域专家组成的评审团，其中，评审员不仅需要对图文大模型的发展现状及相关技术有广泛了解，还需要对模型评测具有丰富的实践经验，以此更加精准地评估图文大模型的回答质量。评审团需针对特定任务设置评分标准，如针对图像创作任务可从美观性、逻辑性、匹配度等角度进行衡量，并对模型预测结果进行独立评判，最终再通过计算平均值等统计学手段统计评测结果。相较客观评价方式，主观评价具有灵活性高以及与实际部署场景贴近等优势。

4.3.2 评测指标

在构建图文大模型评测体系时，需根据任务特性将评测指标分为客观和主观两大类。客观类指标的主要特征是确定性和可量化性，主要适用于评测有明确答案的任务，如识别图片中行人的数量。该类指标的评估结果易于量化和比较，可为图文大模型的评估提供一个稳定且一致的衡量标准。

主观类指标主要用于评估没有固定标准答案的开放性问题，如文生图和风格迁移等创作型任务，在评估时需采取更为灵活的方法，通常可通过人工打分综合评价图文大模型的应用效果。虽然主观类指标相较于客观类指标存在一定的不确定性，但优势在于它更加灵活，更能从用户视角反映模型的实际表现。

- **客观类**

为确保评测的客观性、全面性和公正性，降低主观评测对评估结果的影响，需要利用准确率、召回率等客观性评价指标完成对模型的综合考量。

客观指标通常可应用于评估识别、理解和推理任务的准确性。对于识别任务，如实例识别、手势识别、垃圾满溢、品牌LOGO识别等，由于模型推理结果通常为单一数值，因此可根据分类任务的标准，选取准确率（Accuracy）、精确度（Precision）、召回率（Recall）等指标进行评测。对于理解任务，如口罩位置检测、场景理解等，则侧重于考察大模型对整张图片内容的全面理解，这其中可能涉及目标物体的位置信息，因此常使用交并比（IoU）、CIDEr等评测指标。而对于推理任务，如下一张图像预测，着重考查图文大模型的逻辑理解能力，可以利用FID、SSIM等图像类评价指标对模型预测结果进行客观评测。

除准确性外，实时性、连续性等功能指标也是评价图文大模型的重要维度。其中，实时

性主要考察图文大模型推理的时延，在实际测试时需要根据任务特定要求，分别统计模型在处理短文本问答、长文本问答、单图片问答和多图片问答等任务场景下的响应时间，并进行综合比对。连续性着重考察图文大模型的记忆能力，可通过模型支持的问答最大连续轮次等指标进行评测。这些客观指标全面反映了图文大模型的综合能力，在实际应用中具有重要价值。

- **主观类**

从用户视角全面评估模型的实际应用能力，除采用客观指标外，还须通过主观指标对模型展开评测。主观评测主要集中在创作类任务中，如图像创作、风格变换、图像合成等，这些任务往往需要模型发挥创造性，开放性地生成预测结果，因此没有标准答案。在进行主观评测时，首先需要组建评审专家团，并由评审团制定评分标准。评分标准需综合考察图文大模型能力，以尽可能全面的角度进行评测，在构建评分标准时，需从各个维度对评测任务进行剖析，分维度制定评测指标。除图片美观性、文字优美性等纯主观维度外，还需关注图片内容的正确性、文字的语病错字、与提示词要求的匹配程度等相对客观的评测维度。如在图像创作任务中，可从创作图像的美观程度、逻辑正确性、图像中要素与关键词的匹配程度三个方面评价模型，并分别从各个方面制定打分标准，比如在关键词匹配程度上，可以根据匹配度的百分比进行打分，在逻辑正确性上，可从各事物本身正确性和各事物间相对关系正确性两个方面进行打分。

在采用主观指标进行评估时，首先，需制定合理全面的评价标准；其次，需由专家团中各位专家依据既定标准对模型表现独立评分；最后，采用内部一致性检验、加权平均统计等多种方法统计评估结果，在综合不同专家意见的同时，确保评分一致性，降低人为因素导致的误差，最大程度提高评测结果的稳定性和可信度。

4.3.3 评测数据

构建评测数据需要以任务为导向，覆盖基础场景和实际应用场景，综合考察图文大模型在各种任务下的泛化能力与实际应用效果。在数据构建时，一方面，应尽量避免使用知名的开源数据集，因为这些数据往往会出现在图文大模型的训练集中，无法真实考察模型性能。另一方面，应注意梯度性构建评测用例，合理设置难易比例，不过分脱离当前业界模型的能力范围，同时有效区分各模型的能力水平。

- **数据集构造原则**

在构建评测数据时，须遵循丰富性、公平性和准确性三项核心原则，全面考察图文大模型的综合能力，客观评估其真实能力。

丰富性：在构建评测数据时，需要涵盖业界各种应用场景，真实反映图文大模型的实际应用表现。在测试用例题目设置上，需要采取多元化形式，包括简答、选择、定向回答、图片生成等多种形式进行评测，同时设置不同难度等级的用例。

公平性：构建评测数据时需要确保数据分布在语言、文化等方面具有公平性，并确保不同国家和地区的研究者可以在相同的任务设置下完成评测。

准确性：在构建评测数据时必须确保准确性。题目设计应避免歧义，确保其逻辑严密，能够被不同评测专家一致理解和认可。答案设计应与人类的常识和认知相符，并在测试过程中不断检测和修正可能出现的错误，以确保评估结果的准确性和可靠性。

- **数据集构造方法**

为了更加客观全面地构建评测数据，以真实反映图文大模型的实际应用能力，“弈衡”多模态大模型评测体系从用户视角出发，以丰富性、公平性和准确性为原则，分别面向基础任务和应用任务探索评测数据构造策略，综合评价图文大模型性能。典型构造方法如下：

基础任务数据集构造：在各类识别、检测、计数等基础任务中构建评测数据时，需优先确保全面性。一方面，广泛选取各种任务场景下的图像及文字数据。如在实例识别任务中，综合考察图文大模型对动物、载具、衣着、家具、食物、植物、个人物品等各类生活中常见类别的识别能力，并根据难易度进行梯度设置，简单题目应选取目标物体的典型照片，特征明显清晰，而困难题目则应相对违反常识，以更具迷惑性的方式进行数据构造，如画在墙面上的树木。另一方面，在提示词上应从问题形式上确保全面性，构造选择、简答、判断等各类题目，兼顾中文、英文等语种。此外，还应考虑为数据增加视觉提示，如在图片中添加箭头、圆圈、方框等标记作为会话辅助，与文字提示词一起作为大模型输入，然后要求图文大模型回答视觉提示物体的类别、数量等问题，以增加题目难度。如上，在基础任务的评测数据构造中，需要设置丰富多样的题目，全方位测试模型对典型场景的识别、理解、推理和创作能力。

应用任务数据集构造：应用任务应更加注重从业务场景出发，考察图文大模型在特定场景下的实际应用能力，相较于基础任务偏向广度考察，应用任务的数据构造则着重体现大模型能力的深度考察。需面向部署场景，发掘任务需求，确保评测数据能够更好地反映模型的鲁棒性和可用性。如在口罩检测任务中，不仅仅考察图片中是否有人未佩戴口罩，还应询问大模型是否有人未正确佩戴口罩，从而识别出口罩未覆盖鼻子、嘴部等错误的佩戴方式，测

试模型在实际部署中的可用性；在活体检测任务中，须深入研究并借鉴业界在构造非活体数据方面的各种方法，包括通过照片翻拍、屏幕翻拍、使用面具等手段来生成数据，确保评估数据集更贴近实际应用场景。

4.3.4 评测工具

为全面解决图文大模型评测在技术验证、质量控制、风险管理和合规性等多个层面上的需求，同时规范模型评测，克服当前评测过程中存在的速度慢、不全面、不稳定等局限性问题，中国移动技术能力评测中心构建了“弈衡”大模型评测平台，该平台以智能化自动化、灵活可扩展性、交互体验设计为原则，提供标准化、公正、安全且易于操作的评测服务，推动图文大模型技术的持续创新和应用拓展。具体相关能力如下：

- **数据与模型管理**

数据与模型管理能力包括数据管理、模型管理等功能，主要作用为帮助用户更好地构建数据集，并完成对模型的启停管理。相关功能具体描述如下：

数据管理：提供标准化的数据存储、访问和预处理能力，包括清洗、去重、去噪和异常值处理等核心功能。

模型管理：提供全面的模型接入支持，能够实现自动化模型配置，并广泛兼容各类开源模型，确保了评测平台的开放性和灵活性。

- **评测流程管理**

为提升图文大模型评测效率，评测平台具有完整的评测流程管理功能，可涵盖数据构建、任务下发、任务监控、任务审核等大模型评测的关键环节，为用户提供全自动评测服务。相关功能如下：

评测数据构建：用户可根据评测任务自主设计数据集和选择评测指标，实现数据预处理，并提供多样化指标模板，满足用户的评测需求，增强评测的灵活性和实用性。

评测任务下发：评测任务下发是评测平台高效自动化特性之一，用户无需深入了解不同模型的接口细节，只需在平台上选定评测对象和相应的数据集，即可通过一键式操作快速下发评测任务，从而简化评测流程，减少人工设置和干预，提升图文大模型评测的效率和准确性，并确保了评测的一致性和可复现性。

评测任务监控：用户可通过用户界面，对图文大模型评测进度进行直观跟踪，实时监控评测任务的执行状态，包括当前的进度、已处理的数据量等。该能力有助于及时发现并解决评测过程中可能出现的问题，确保图文大模型评测的顺利进行。

评测任务审核：评测任务审核功能允许专业人员对平台自动生成的评测结果进行人工核查，以确保评测结果的准确性。在评测结束后，平台会进行自动判卷，此时人工可进行再次核查，为评测的精确性和权威性提供额外保障，增强评测结果的可信度和实用性。

- **结果分析与展示**

评测平台除了各项自动化能力，还可对评测结果进行分析与展示，计算各参测模型的综合得分并进行排名，梳理并总结各图文大模型的综合能力水平。具体相关功能如下：

专家评分：对于图片创作等生成类任务，常规的客观指标很难对图文大模型的真实能力进行综合评判，评测平台提供专家评分功能，对模型能力进行主观评价。

榜单生成：评测平台可依据模型的自动化评测结果和专家评分，自动整理图文大模型在不同指标上的表现，一键生成模型综合能力排名，帮助用户快速了解模型能力水平。

榜单图形化展示：评测平台可通过图形化界面，清晰展示各图文大模型的综合排名，将模型在关键性能指标上的相对排名直观展示给用户，帮助用户快速甄选优秀模型、及时发现模型性能瓶颈，为用户选择和优化模型提供支持。

智能分析与报告：评测平台可通过AI技术，深度挖掘评测数据，精准捕捉并总结模型能力，自动编制评测报告，呈现图文大模型的性能指标及排名，全面评估和比较不同模型的性能表现。

“弈衡”大模型评测平台为用户提供了一个全面、高效、智能的评测解决方案，具有“2-4-6”多维度评测体系、业界领先的自动化评测能力、用户友好的“一键测试”功能、高可拓展性等多项优势，可广泛应用于图文大模型评测，大幅提高评测效率和准确性，对于图文大型模型的评测和优化具有重要意义。

4.4 评测维度

为全面评估和综合测试图文大模型在识别、理解、推理、创作等各类任务中的能力，确保覆盖各类任务类型和应用场景，应从功能性、准确性、可靠性、安全性、交互性、应用性

六大维度对大模型进行评测。具体如下：

功能性：此维度主要关注图文大模型解决多种任务的能力，包含任务丰富度、多模态能力和支持完备度三类，其中任务丰富度是指大模型支持任务类型的数量，多模态能力是指对文生图、图生文等五种多模态输入输出类型的支持程度，支持完备度包含语种支持度、最大输入文本长度、最高图片分辨率等七项指标，主要考察图文大模型在输入输出设置上的支持程度。

准确性：此维度主要关注图文大模型执行各类任务的性能。在评估图文大模型准确性时，需要针对不同类型的任务，选择最合适的评价指标。针对实例识别、口罩检测、人群计数等具有明确标准答案的任务，要优先选择准确率、召回率等客观评价指标，而针对风格变换、图像合成等创作类任务时，应选择主观评价方式，更加全面地反映图文大模型在用户视角下的真实性能。

可靠性：此维度主要关注大模型的抗噪声能力，以及对同一问题多次输出结果的一致性。抗噪声测试中，对测试数据集进行几何变形、色彩空间噪声、专业噪声处理和水印等处理后，重新输入大模型进行评测，全面考察图文大模型对各种图片噪声的抗干扰能力。一致性测试中，评测人员针对同一个问题，对图文大模型进行连续多次问答，关注多次问答的评测结果是否一致。

安全性：此维度主要考察图文大模型生成结果的毒害性和公平性，包括歧视偏见、内容毒性、违规违法、不适表达和版权隐私五类。其中每一类又包含多种测试角度，比如歧视偏见中包含种族歧视、性别歧视、年龄歧视等，内容毒性包含不实信息、毒性内容、敏感话题等。安全性评估在确保生成内容合法合规、防止歧视偏见、维护社会道德等方面具有重要作用，是保障大模型技术健康发展的关键评测维度。

交互性：此维度主要关注用户使用图文大模型时的交互体验。在评估交互性时，着重考察实时性、连续性、丰富性和规范性，此外如果应用场景为生成图片任务，还考察清晰度、色彩等图片质量指标；如果应用场景包含文本生成，则考察表达的流畅度。其中，实时性是指图文大模型生成结果的速度，连续性是指支持问答的最大连续轮次，丰富性是指生成图片的多样性或生成文本的长度，规范性则是指生成图片和文字的合理合规性。

应用性：此维度主要关注图文大模型产品或系统在现实应用场景中的部署、运维、支撑能力和使用效果，旨在全面审视基于图文大模型的产品在各方面的实用性。在部署能力方面，关注系统兼容性、快速部署、可扩展性等情况；在运维能力方面，关注系统稳定性、故障告警及恢复等情况；在支撑能力方面，关注定制化、业务整合等情况；在使用效果方面，关注用户体验、应用成效等情况。此维度中大部分评价指标很难通过自动化的客观指标来衡量，往往需要借助人工主观评估、访谈调研等方式进行考察。

上述六大评测维度相互独立，覆盖产品应用中用户端到端业务全流程的各环节，可真实评估图文大模型实际应用中的能力表现，具有全面性、科学性和客观性的特点。

05 多模态大模型评测展望

随着近年来人工智能技术的快速演进，各类多模态大模型（如GPT-4V、文心一言、讯飞星火等）取得显著进展，逐渐成为人工智能领域的核心技术之一，引起国内外产业界和学术界的广泛关注。

这些大模型以其卓越的多模态理解、推理能力，为自然语言处理、计算机视觉等领域带来了革命性的变革。为了全面评估这些图文大模型的性能和潜力，我们致力于构建一个科学、客观、公正的评测基线。本白皮书的贡献可总结为以下三点：**一是**深入剖析了图文大模型在多个领域的典型应用，根据评测的实际需求，精准划分出四大类任务：识别类任务、理解类任务、创作类任务以及推理类任务。这一分类不仅体现了模型功能的多样性，也为后续的评测提供了明确方向。**二是**全面梳理了当前主流的评测方式、评测维度和常见评测指标，深入分析了业界具有代表性的图文大模型评测体系。**三是**基于上述调研与分析，提出了“弈衡”多模态大模型评测体系，该体系通过多元化的评测方式，对基础任务与应用任务进行全面评测，从功能性、准确性、可靠性、安全性、交互性和应用性六个维度，全面评估图文大模型的综合能力，覆盖多个关键评测指标。**此外**，我们还为评测数据集与评测平台的构建提供了具有指导意义的范式，为未来的评测工作奠定了坚实的基础。

然而，目前图文大模型仍面临许多亟待解决的问题。首先，**图文大模型在特定领域的准确性仍有待进步。**虽然图文大模型在多任务泛化性上表现出了明显的优势，但在针对大规模人群计数、多图像内容合成、小物体识别等难度较高的任务中，图文大模型表现尚不如经过针对性训练的图像小模型，这使得它们在实际业务场景中应用仍然受限。其次，**图文大模型的实时信息更新慢。**由于图文大模型实现良好性能需要针对性数据进行预训练，所以对于新兴任务及特殊场景，数据更新慢，能力有待提高，难以实际应用。

针对以上问题，图文大模型评测体系发展也需要深入思考，更好地规范大模型良性发展。未来评测技术的研究重点可能聚焦于以下两个方面：

一是针对特定业务场景开展评测。

在对图文大模型进行评测时，不仅要考察常规物体的识别和理解能力，更要针对实际业务场景开展评测，尤其要在复杂任务上评估模型能力边界，确保对大模型进行深度与广度上的全面测试，真实反映其应用能力。

二是跟踪技术演进优化评测体系。

应实时掌握多模态大模型发展现状，及时把握前沿应用场景，进一步拓展评测模态范围，不断更新评测数据，优化评测指标，丰富评价维度，迭代评测工具，衡量模型对新数据、新场景的适应能力，提升模型应用能力与部署的鲁棒性。

中国移动技术能力评测中心作为中国移动集团级第三方专业评测机构，多年来深入开展评测技术研究，积累了丰富的产品技术能力评测经验。经过广泛调研与实践，针对图文等多模态大模型，构建了“弈衡”多模态大模型评测体系，一方面，可为中国移动工业、政务、金融、交通、安全等十余个行业大模型的全面客观评测提供标准基线，助力中国移动AI+重塑千行百业；另一方面，可为业界大模型评测提供参考依据，为业界合作伙伴提供一站式大模型评测服务，推动国产大模型产业成熟和落地应用。我们希望与产业界相关企业和研究机构一道，继续攻关大模型评测关键技术，不断完善多模态大模型评测体系，共同构建评测产业标准化生态，促进大模型技术的健康快速发展。

参考文献

[1]OpenAI. "GPT-4 Technical Report" arXiv preprint arXiv:2303.08774v3 (2023).

[2]Bubeck, Sébastien, et al. "Sparks of artificial general intelligence: Early experiments with gpt-4." arXiv preprint arXiv:2303.12712 (2023).

[3]哈尔滨工业大学. ChatGPT 调研报告。[R/OL]. (2023-03-06)

[4]亚信科技&清华大学. AIGC 赋能通信行业应用白皮书（2023）。[R/OL]. (2023-03)

[5]IDC 国际数据公司. 2022 中国大模型发展白皮书。[R/OL]. (2023-02)

[6]Liang, Percy, et al. "Holistic evaluation of language models." arXiv preprint arXiv:2211.09110 (2022).

[7]Gili, Kaitlin, Marta Mauri, and Alejandro Perdomo-Ortiz. "Evaluating generalization in classical and quantum generative models." arXiv preprint arXiv:2201.08770 (2022).

[8]Strubell Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[9]Anil, Rohan, et al. "Palm 2 technical report." arXiv preprint arXiv:2305.10403 (2023).

[10]Liu Y, Duan H, Zhang Y, et al. Mmbench: Is your multi-modal model an all-around player?[J]. arXiv preprint arXiv:2307.06281, 2023.

[11]Liu Y, Li Z, Li H, et al. On the hidden mystery of ocr in large multimodal models[J]. arXiv preprint arXiv:2305.07895, 2023.

[12]He Z, Wu X, Zhou P, et al. CMMU: A Benchmark for Chinese Multi-modal Multi-type Question Understanding and Reasoning[J]. arXiv preprint arXiv:2401.14011, 2024.

[13]Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.

[14]Bitton Y, Bansal H, Hessel J, et al. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use[J]. arXiv preprint arXiv:2308.06595, 2023.

[15]Li B, Wang R, Wang G, et al. Seed-bench: Benchmarking multimodal llms with generative comprehension[J]. arXiv preprint arXiv:2307.16125, 2023.

[16]Zhang, Yuan, et al. "Unveiling the Tapestry of Consistency in Large Vision-Language Models." arXiv preprint arXiv:2405.14156 (2024).

[17]中国信息通信研究院. 大规模预训练模型技术和应用评估方法。[R/OL]. (2022-06-01)