arXiv:2601.00561v1 [cs.CV] 2 Jan 2026

AEGIS：探索统一多模态模型世界知识能力的极限

Jintao Lin[1]*，**Bowen Dong**[2]*，**Weikang Shi**[3],**Chenyang Lei**[4],**Suiyun Zhang**[4],**Rui Liu**[4],**Xihui Liu**[1]†

[1]香港大学 [2]香港理工大学[3]香港中文大学 [4]华为研究院



图 1. AEGIS 基准测试示意图。AEGIS 的贡献包括：1）一个全面且具有挑战性的基准测试，涵盖四种不同的视觉理解和生成任务，并覆盖广泛的知识领域（即世界知识）；2）一种基于确定性检查清单的评估协议，可获得具体的评估结果；3）对最先进的统一多模态模型及其他生成模型进行实证分析，以揭示其在世界知识和推理方面的脆弱性。

摘要

统一多模态模型（*UMM*）在多种任务中运用世界知识的能力，仍是一项亟待解决的关键挑战。现有基准测试存在不足，仅提供彼此割裂的单任务评估，诊断能力有限。为弥补这一空白，我们提出 *AEGIS*（即，***A***评估***E***编辑***G***生成***I***解释与理解——面向***S***超智能），这是一个全面的多任务基准测试，涵盖视觉理解、生成、编辑和交错生成。*AEGIS* 包含 *1,050* 道具有挑战性、经人工标注的问题，涉及 *21* 个主题（包括科学、技术、工程和数学、人文学科、日常生活等）和 *6* 种推理类型。为在不依赖模糊指标的情况下，具体评估 *UMM* 在世界知识方面的表现，我们进一步提出确定性清单式评估（*DCE*），这是一种用原子化的“是/否”判断取代模糊的基于提示的评分的协议，旨在提高评估的可靠性。大量实验表明，大多数 *UMM* 都存在严重的世界知识缺陷，且在进行复杂推理时，性能会显著下降。此外，简单的插件式推理模块能够部分缓解这些弱点，为未来研究指出了一个有前景的方向。这些结果凸显了基于世界知识的推理作为 *UMM* 关键前沿领域的重要性。

1. 引言

多模态大语言模型（MLLM）[1, 11, 24, 51, 64] 和生成模型（例如文本到图像扩散模型）[5, 27, 38, 61]的快速发展，在视觉理解、生成任务及相关应用中取得了显著成功。这些独立模型的成功夯实了人工智能的基础，并

*贡献相同，排名随机。†通讯作者：Xihui Liu < xihuiliu@eee.hku.hk>。

激励研究人员探索一种结合式框架，迈向“一个模型应对所有任务”的范式 [8, 24, 50]。因此，统一多模态模型（UMM） [6, 12, 18, 26, 47, 50, 54,55] 已成为人工智能领域的一项重要趋势。统一多模态模型能够在单一网络中同时处理多种模态和任务，因此具有很强的通用性，并能兼容各种下游应用。然而，现实应用（例如人工智能助手 [2, 36]和智能客服 [16, 48]）面临着日益多样且复杂的用户请求，要准确作答往往需要复杂的推理能力和丰富的世界知识。为了满足这些应用的需求，模型必须能够将对世界知识深刻而广博的掌握，灵活转化为高保真文本和视觉输出，从而应对涵盖广泛领域及下游任务的各种需求 [12, 37], 。这种能力对于提供准确答复、减少幻觉以及最大限度降低人工返工成本至关重要。

尽管已有大量基准测试被提出，用于评估多模态大语言模型、生成模型和统一多模态模型，但它们大多聚焦于常见任务（例如，标准视觉问答） [14, 28]或简单的物体与场景生成 [19, 22] 和编辑 [33, 59]。尽管近期研究尝试通过引入“冷门知识”主题 [4, 20] 或推理挑战 [42, 60],来提高复杂度，但仍存在三项根本性局限。1）几乎所有现有基准测试都局限于单任务评估。随着生成模型发展为统一多模态模型，这种“各自为政”的评估方式无法衡量统一多模态模型在不同任务间的关键性能差距。2）它们往往缺乏细粒度诊断能力，无法定位模型失效的根源。也就是说，我们仍难以判断表现不佳究竟源于理解（大语言模型）组件，还是生成模块。现有基准测试的难度和范围往往不足以全面探查模型世界知识的深度与广度，以及其复杂推理能力。因此，亟需一种更全面、更具挑战性且更详细的基准测试，并配备全新的评估协议，以真正评估现代统一多模态模型的能力。

为解决这些局限，如图 1所示，我们提出了 AEGIS（即，**A**ssessing（评估）、**E**diting（编辑）、**G**eneration（生成）、**I**nterpretation-Understanding（阐释与理解）for**S**uper-intelligence（超智能）），这是一个面向统一多模态模型及相关生成模型的全面且具有挑战性的多任务基准。具体而言，AEGIS 包含 1,050 个经过人工标注和验证的问题，涵盖科学、技术、工程和数学、人文学科以及日常生活中的 21 个详细主题。每个问题都配有相应的参考答案和关键词，支持深入分析。通过人工参与的数据构建、完善和标注流程，AEGIS 涵盖六种不同的推理类型，并评估四种不同的任务：视觉理解、生成、编辑以及交错生成。

基于大量的问题，我们旨在自动、准确地评估模型。然而，现有基于评分的“LLM 作为评判者”方法 [37, 42] 面临两项根本局限。其一是评分指标依赖启发式且有限；其二是大语言模型给出的分数含义模糊。尽管一系列通用评分指标（例如真实性和图像质量）已广泛用于现有基准测试，但这些指标仍显得含糊、粒度粗糙，且缺乏解释力。为解决这些问题，我们提出确定性清单式评估（DCE）。DCE 使用多模态大语言模型（例如，GPT-4o [24]）处理参考回答及其对应的关键词，并生成一系列原子化的“是/否”判断问题。每个问题都与参考回答中的一个确定性关键部分严格对应。这种方法将复杂的判断任务简化为可验证的步骤，从而提高 LLM 作为评判者范式的可靠性。随后，AEGIS 通过计算回答获得“是”判断的平均比例，有效衡量统一多模态模型的表现。

得益于我们创新的数据标注和评估协议，如表 1 所示，与现有基准测试相比，AEGIS具有显著的诊断价值。首先，AEGIS极大地促进了跨任务评估，能够分析理解、生成和编辑之间的相关性，而非将各项任务孤立评估。此外，AEGIS通过探查不同类型的推理进行深入诊断，从而揭示模型的脆弱之处并定位组件级缺陷。最后，AEGIS以现实应用为基础开展评估，对模型鲁棒性进行双重（即学术和实践）分析，以提升模型的部署就绪度。

我们在AEGIS上进行了广泛评估，涵盖了众多开源和闭源模型，系统地识别出它们各自的优势与不足。具体而言，除Gemini Nano Banana [11], 这一显著例外外，大多数统一多模态模型在世界知识方面都存在严重缺陷。此外，引入复杂推理后，所有模型的性能均显著下降。值得肯定的是，我们证明，集成简单的插件式推理模块能够在一定程度上缓解这些缺陷，这为统一多模态模型的未来发展指明了一个颇有前景的方向。总之，本文的贡献如下：

- 我们提出了 AEGIS，这是首个全面且具有挑战性的基准测试，可同时评估视觉理解、生成、编辑和交错生成任务，覆盖极其广泛的世界知识。

- 我们提出一种确定性清单式评估方法，通过一系列是非问题作为约束，评估生成回答的正确性。

表1. AEGIS 与现有世界知识评测基准的比较。任务列中的 U/G/E/I 分别表示视觉理解、生成、编辑和交错生成任务。AEGIS 在知识和推理类型覆盖范围方面更具优势，并提供了确定且可靠的评估指标。

基准测试	任务				领域			推理类型						评估	
	U	G	E	I	科学、技术、工程和数学		人文	日常生活	空间	时间	因果	比较	类比	逻辑	评估类型	具体
WISE [37]	✗	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✓		基于分数	✗
RISE [62]	✗	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗	✗	✓		基于分数	✗
KIRS-Bench [56]	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓		基于分数	✗
T2I-ReasonBench [42]	✗	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✓		基于分数	✗
R2I-Bench [7]	✗	✓	✗	✗	✓	✗	✓	✗	✗	✓	✓	✓	✗		基于分数	✗
WorldGenBench [60]	✗	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗		基于清单	✗
GIR-Bench [29]	✓	✓	✓	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓		基于分数	✗
AEGIS（我们的方法）	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		基于清单	✓

与现有方法相比，这种方法能够提供更可靠的判断。

• 在统一多模态模型和生成模型上的广泛实验结果揭示了它们在世界知识理解与生成方面存在共同的脆弱性。这些发现为未来模型的开发提供了关键洞见。

2. 相关工作

2.1. 统一多模态模型

近期的统一多模态模型（UMM）通过整合生成能力从多模态大语言模型（MLLM）发展而来，通常采用共享主干网络和外部解码器。根据架构，这些统一多模态模型可分为三大类：基于扩散的模型 [30, 40, 43, 49, 58], 纯自回归方法 [6, 18, 26, 50, 54, 55] 使用 Transformer 进行特征聚合，以及混合模型 [12, 31, 35, 41]将自回归文本生成与多步图像去噪相融合。尽管这些模型采用了多阶段训练框架，但仍有两个关键问题亟待解答：第一，理解能力究竟如何增强这些统一系统中的生成能力；第二，它们在世界知识理解 [13, 20, 21, 32]与生成 [9, 37, 42, 56]方面的实际能力究竟达到何种程度。本文旨在对这些问题展开批判性研究。

2.2. 世界知识基准测试

鉴于最先进的生成模型在处理常见视觉理解、图像生成和编辑任务方面取得了显著进展，近期研究开始评估它们在更广泛知识范围内的能力，即世界知识。为提高世界知识基准测试的难度，一些研究选择纳入专业性较强或较少涉及的主题。例如，MM-IQ 利用图形智商测试题探究多模态大语言模型在视觉理解方面的局限，而 WISE 则引入文化和自然科学领域的指令，探索智能图像生成能力。第二类研究占近期增强型世界知识基准测试的大多数，主要通过提高问题的推理难度来增强挑战性。这些基准测试在原始指令中引入复杂的推理逻辑，从而增加生成模型正确理解预期含义的难度。尽管取得了成效，现有基准测试通常存在三个关键缺陷：它们往往过度关注推理类型有限的常见问题，却忽视罕见或具有挑战性的问题；可能难以分析不同任务之间的相互影响；其基于大语言模型的评估可能采用含糊不清的指标（例如，预测真实感、一致性和质量指标的得分）。相比之下，我们的工作旨在通过涵盖多样且具有挑战性的问题，并采用基于检查清单的评估协议，克服这些具体局限。

3. AEGIS 基准测试

3.1. 数据集概述

为全面评估统一多模态模型和其他生成模型在视觉理解、生成、编辑和交错生成任务上的表现，涵盖**广泛的**世界知识范围和**复杂推理**，我们提出了 AEGIS。如表1所示，AEGIS 涵盖三个通用领域（即科学、技术、工程和数学、人文学科和日常生活），包含 21 个不同主题。数据统计见表2。每个主题包含 15 个用于视觉理解、生成和编辑的提示，以及 5 个视觉交错生成问题，用于衡量复杂生成能力。此外，AEGIS 在大多数提示中融入了六种不同的推理类型，要求统一多模态模型具备固有的推理能力，才能完成每项请求。以下各节将详细介绍 AEGIS 的构建、标注和评估协议。

3.2. 数据构建与标注

构建一个覆盖广泛知识范围的基准测试面临着重大挑战。我们的数据构

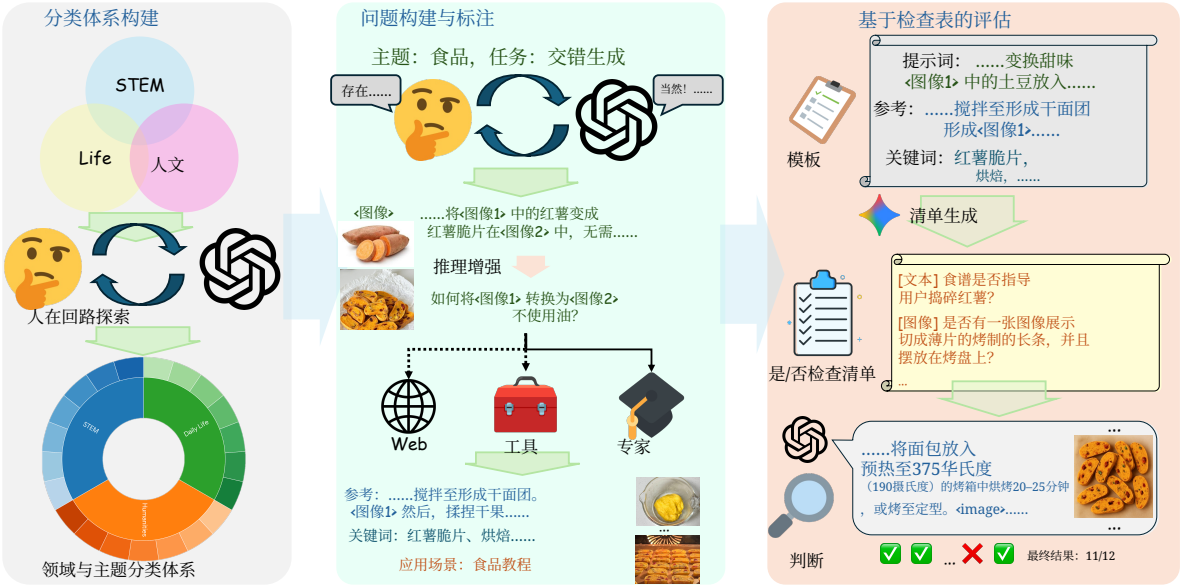


图2. 我们提出的 AEGIS 的数据构建与评估流程。AEGIS 基于人在回路探索中获得的板块分类体系，采用高质量的数据构建流程，并运用人在回路探索和最优标注方法（网络、工具或专家）来创建推理增强问题。另一项关键亮点是新颖的确定性清单式评估（DCE）：首先由多模态大语言模型根据参考答案生成一份原子“是/否”问题检查清单。随后，评判多模态大语言模型使用该检查清单，对模型预测作出清晰、具体且可靠的判断。

表2. AEGIS数据集在不同主题和任务中的数据分布，其中“U”表示视觉理解问题，“G”表示视觉生成问题，“E”表示视觉编辑问题，“I”表示视觉交错生成问题。

领域	主题	U	G	E	I	总计
STEM	生物学	15	15	15	5	50
	化学	15	15	15	5	50
	数学	15	15	15	5	50
	医学	15	15	15	5	50
	物理学	15	15	15	5	50
	天文学与地理	15	15	15	5	50
	IT	15	15	15	5	50
人文学科	农业	15	15	15	5	50
	历史	15	15	15	5	50
	电影	15	15	15	5	50
	音乐	15	15	15	5	50
	Art	15	15	15	5	50
	文化	15	15	15	5	50
	建筑	15	15	15	5	50
日常生活	活动	15	15	15	5	50
	动漫	15	15	15	5	50
	Game	15	15	15	5	50
	摄影	15	15	15	5	50
	工程	15	15	15	5	50
	Food	15	15	15	5	50
	交通	15	15	15	5	50
总计	**21 个子类别**	**315**	**315**	**315**	**105**	**1050**

建流程包括多个连续阶段，以确保数据质量高且覆盖全面。

首先，我们建立了基础分类体系。对于主题分类体系，我们使用了一个大语言模型 [24]，通过人工参与其中的流程，逐步探索并定义了 21 个不同的主题。这些主题涵盖科学、技术、工程和数学、人文学科及日常生活中的常见场景，为广泛的知识评估奠定了基础。对于推理分类体系，受近期研究 [12, 42, 56], 的启发，我们提出了六种推理类型：空间推理（分析位置）、时间推理（分析随时间发生的变化）、因果推理（理解强因果关系）、比较推理（识别差异）、类比推理（识别相似之处）和逻辑推理（分析结构化关系）。

基于这些分类体系，我们采用人工参与的方式进行提示生成。针对特定主题和随机选择的任务，我们首先提示大语言模型为该主题生成一个“明确提示”，确保其表述清晰、没有歧义，且不涉及复杂推理。如果输出内容冗余，我们会提供新的“思考方向”来引导大语言模型，直到构造出一个独特的问题。随后，我们从分类体系中选出最合适的推理类型，手动进行推理增强，将“明确提示”转化为“推理增强”提示。由此生成的提示模拟了实际应用中的模糊性和复杂性，提升了 AEGIS 在现实世界中的实用价值。

定义好提示词后，我们根据所选任务构建了必要的输入和真实标注。对于需要图像输入的提示词，我们通过三种方式构建了高质量数据集：(1) 从互联网上抓取免版权内容，(2) 使用生成模型和图像编辑工具手动创建图像，以及 (3) 委托专家制作图像。这种多管齐下的方法确保了较高的视觉保真度。对于真实答案，我们也采用同样的三种方式来构建



图 3. AEGIS 基准测试示例，红色标示问题中考查的关键点。AEGIS 涵盖四种视觉生成任务，包含来自 21 个不同主题的 1,050 个推理增强问题，可用于探索统一多模态模型和其他生成模型在广泛知识领域（即世界知识）及不同推理类型下的生成能力。

用于生成和编辑任务的参考图像。对于文本任务，大语言模型会生成一份回复草稿，随后由人类专家核验并修改。最后，为了便于评估，我们使用多模态大语言模型从明确提示及其参考答案中提取关键词。我们人工检查这些关键词，以确保其涵盖解答问题所需的所有关键要点。

3.3. 基于检查表的评估

基于细粒度标注，我们需要一种具体且可靠的评估方案。现有基于评分的“LLM 作为评判者”方法 [37, 42, 60]面临两个根本局限：（1）评分指标依赖启发式且种类有限；（2）大语言模型给出的分数含义模糊且粒度较粗。受竞赛中确定性、按分点评判方式的启发（例如国际数学奥林匹克），在这类竞赛中，完成具体且可验证的步骤即可获得相应分数，我们提出了确定性清单式评估（DCE）。该方案包含两个主要阶段：评判问题生成和模型回答判定。

判定清单的生成是基础阶段。给定一个问题（包括明确提示、可选输入图像和参考回答）及其关键词，我们利用最先进水平的多模态大语言模型（例如，Gemini [11]）提取一系列原子级判定问题。如图 2（右）所示，每个问题都可通过“是”或“否”作答，并聚焦于回答中的一个具体方面（例如，细微但可检测的修改）。与直接评分相比，这种方法降低了歧义和判定难度。为进一步提升质量，我们进行人工筛选，删除重复或冗余的问题（占原集合的 ∼20%），确保最终检查清单能够精准衡量关键要点。

这份离线生成的检查清单最终确定后，即用于模型回答判定阶段。我们通过模板（详见补充材料），将明确提示词、模型的回答和检查清单整合为统一的输入指令。随后，我们利用多模态大语言模型 [11]，针对检查清单中的每个条目生成“是”或“否”的回答，并给出相应解释。最后，DCE 通过计算模型回答在所有问题中获得“是”判定的平均比例，有效衡量统一多模态模型的表现。

3.4. 联合效用优势

借助我们提出的数据标注和评估协议，AEGIS 展现出显著的多方面效用，并凭借以下优势超越其他基准测试。首先是跨任务评估。AEGIS 不仅能够在广泛的世界知识范围内全面评估四项不同的任务，还能详细分析理解任务与其他任务之间的相关性。

表 3. 各任务与领域的综合性能比较。分数按三个领域（科学、技术、工程和数学、人文和日常生活）细分，并针对四项主要任务进行报告。“*”表示不支持交错多图像输入的模型，“#”表示不支持部分任务的模型（以“-”标注）。Nano Banana 和 GPT-4o+GPT-Image-1 在所有任务上的表现都优于其他模型。

模型	理解			生成			编辑			交错生成			总体
	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	
统一多模态模型													
Gemini Nano Banana	64.5	65.7	55.0	42.6	49.5	45.5	44.4	62.4	54.2	50.2	41.6	43.4	52.9
GPT-4o+GPT-Image-1	52.9	50.9	46.9	38.2	51.6	42.8	39.4	53.2	45.2	38.9	34.7	33.0	45.7
Bagel-7B 无思维链	25.5	26.9	19.3	12.1	20.6	15.3	15.0	17.6	21.2	13.0	11.9	8.2	18.5
Bagel-7B（思维链）	31.8	31.7	22.0	14.9	31.2	21.3	11.6	23.5	23.1	11.8	11.9	9.9	22.3
Ovis-U1*	26.3	31.0	17.1	12.7	25.0	16.3	19.3	27.2	25.9	12.2	12.5	8.4	21.2
BLIP3o*	30.8	43.3	21.7	3.7	6.3	2.7	2.6	4.4	4.5	9.4	8.3	4.0	13.7
Qwen-Image	31.4	41.2	22.7	17.9	31.9	25.4	20.7	35.4	33.4	22.0	18.7	17.9	28.0
Janus-Pro 7B#	7.9	18.0	9.7	13.7	17.0	18.2	-	-	-	2.2	6.5	2.7	-
Show-o2#	15.4	26.6	11.1	16.7	24.7	22.5	-	-	-	5.3	8.7	3.4	-
Emu-3#	3.1	8.0	2.0	8.9	19.7	14.5	-	-	-	-	-	-	-
多模态大语言模型理解													
Qwen-3-VL 8B	42.6	48.1	34.0	-	-	-	-	-	-	-	-	-	-
Kimi-VL-A3B	30.6	36.4	23.5	-	-	-	-	-	-	-	-	-	-
GPT-5	67.0	57.4	60.6	-	-	-	-	-	-	-	-	-	-
Gemini-2.5-Pro	72.1	77.3	63.3	-	-	-	-	-	-	-	-	-	-
图像生成或编辑模型													
FLUX.1-Dev*	-	-	-	15.4	29.2	16.8	-	-	-	-	-	-	-
Step1X-Edit*	-	-	-	-	-	-	19.8	31.9	37.1	-	-	-	-
Instruct-Pix2Pix*	-	-	-	-	-	-	17.3	17.6	23.5	-	-	-	-
Seedream*	-	-	-	33.9	43.7	38.6	32.8	53.0	43.0	-	-	-	-

其次是推理类型效应调查。通过比较模型对标准问题的回答与涉及不同推理类型的问题的回答，研究人员可以揭示统一多模态模型和生成模型固有推理能力中的脆弱性。这项分析还有助于调查统一多模态模型中的哪些组件对这些缺陷影响最大。最后是真实世界应用评估。AEGIS 将每个数据样本与特定的实际应用联系起来，从学术和真实世界两个视角对模型脆弱性进行双重分析。这一洞见对于增强模型的稳健性和部署就绪度至关重要。

4. 实验

4.1. 实验设置

推理时，对于黑盒模型，我们直接调用相应的 API 获取最终结果。对于开源模型，我们利用 4 个 GPU，通过 trans- formers [53] 和 diffusers [45]工具包预测结果。评估时，我们使用 Gemini-2.5-Pro [11] 作为默认评判模型。该模型具备足够精细的视觉理解能力，能够满足 DCE 的要求。我们还将在补充材料中分析评判模型的选择，以验证我们的方案。

4.2. 基线模型

统一多模态模型。我们的主要关注点是统一多模态模型的能力。因此，在实验中，我们引入了一系列黑盒或开源统一多模态模型 [6, 11, 12, 24], , 包括 Gemini Nano Banana [11]、GPT-Image-1 [24], Bagel（支持和不支持思维链） [12], 、Ovis-U1 [47], 、BLIP3o [6], 、Qwen-Image [54], 、Janus-Pro [8]、Show-o2 [57], 和 Emu-3 [50]。

其他生成模型。 除了 UMM 外，对于每项理解、生成和编辑任务，我们还评估了相应的单任务生成模型，以评估其能力。参与评估的模型包括多模态大语言模型（例如，Qwen3-VL-8B [1], GPT-5 [24], Kii-VL-A3B-Instruct [44], 和 Gemini-2.5-Pro [11]）以及图像生成或编辑模型 [3, 27, 34, 39]。

4.3. 实验结果

我们对第 4.2 节提到的模型进行了实验，并在表 3 中展示了实验结果。这些结果揭示了明显的性能层级。

统一多模态模型之间的总体比较。领先的闭源统一多模态模型 Gemini Nano Banana 和 GPT-4o+GPT-Image-1 在所有领域的所有任务上都显著优于其他模型。尤其是 Gemini NanoBanana，达到了最先进水平，在全部四项任务中都展现出独特的强大且均衡的能力，尤其在复杂的科学、技术、工程和数学理解（64.5）

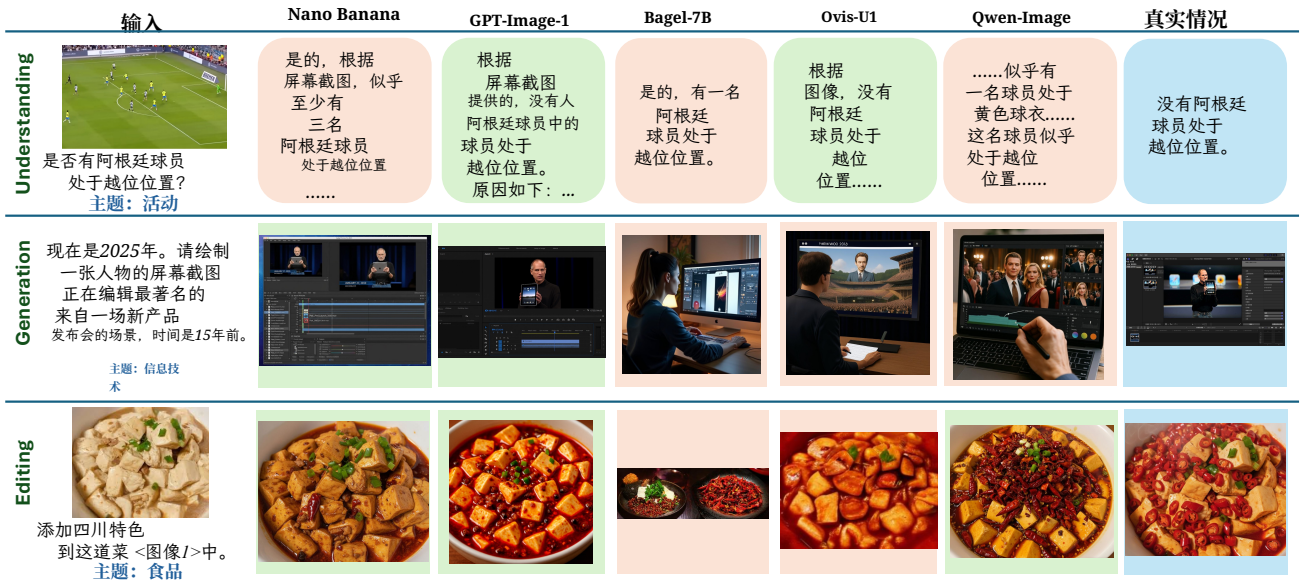


图 4. 五个最先进的统一多模态模型在理解、生成和编辑任务上的可视化结果。绿色表示回答正确，红色表示回答错误。Nana Banana 展现出良好的图像生成和编辑质量，并且在编辑任务中的主体对象一致性优于其他模型。与此同时，开源模型在这些任务上的表现并不理想。

和生成（42.6）方面表现突出。与此形成鲜明对比的是，当前开源统一多模态模型的表现并不理想。Qwen-Image 和 Bagel-7B 等模型虽代表了重要的开源努力，但与闭源模型相比仍明显落后。我们推测，这一差距主要源于两个因素：（1）这些模型的参数规模有限，限制了其存储基准测试所需大量世界知识的能力；（2）训练数据质量可能存在不足，尤其是缺乏构建稳健推理能力所必需的多图像关联数据。

推理有助于改进结果。与使用和不使用推理的 Bagel 相比，总体性能从 18.5 提升至 22.3，这表明，即使采用中等规模的开源模型，在世界知识方面的理解和生成能力也可以通过外部推理模块得到进一步增强。

更多、更高质量的训练数据有助于提升世界知识。与此同时，在理解能力方面，Gemini-2.5-Pro 和 GPT-5 在理解任务上分别取得 70.9 和 61.7 的成绩，这与相应统一多模态模型的结果一致。令人惊讶的是，仅有 8B 参数的开源 Qwen-3-VL，在科学、技术、工程和数学知识理解方面也取得了 42.6 的成绩，在人文知识理解方面取得了 48.1 的成绩。相比之下，采用 Qwen2.5-VL-7B 作为骨干网络的统一多模态模型（例如，Qwen-Image [54] 和 BLIP3o [6]）在理解任务上的表现要低得多。由于这些模型的参数量和宏观设计几乎相同，因此可以得出结论：更优的（也就是说，更细粒度、更丰富且更详尽的）多模态预训练数据有助于提升世界知识理解表现。

理解能力限制了其他任务的性能上限。此外，结果显示，不同任务的表现存在明显的不一致性。对于当前最先进的模型，难度呈现出显著趋势：理解任务的表现最佳，生成任务有所下降，编辑任务则进一步略微下降，而在要求最复杂推理的交错生成任务中，表现骤然下滑。这表明，模型更擅长解读知识，而非生成式地运用或操纵知识。如果统一多模态模型中的多模态大语言模型组件无法正确推理或理解某个问题，那么视觉生成或编辑结果也无法正确。因此可以得出结论，统一多模态模型的理解能力限制了相应视觉生成和编辑能力的上限。这些结论或可启发未来研究设计更优秀的多模态大语言模型，使未来的统一多模态模型能够在实际应用中处理复杂或模糊的指令。

4.4. 定性结果

除定量分析外，我们还对五种最先进的统一多模态模型（即，Gem- ini Nano Banana [11], GPT-Image-1 [24], Bagel-7B [12],Ovis-U1 [47], 和 Qwen-Image [54]）在理解、生成和编辑任务上的预测结果进行了可视化，以进行定性分析。相关结果如图4所示。与其他开源模型相比，Gemini Nano Banana 和 GPT-Image-1 的视觉质量和图文一致性更佳，尤其是在生成和编辑任务中。此外，Nano

表4。DCE 的人工验证一致性，其中 U/G/E/I 分别表示理解、生成、编辑和交错生成。较高的一致率验证了 DCE 的可靠性。

一致性（%）	U	G	E	I	总体
Gemini-2.5-Pro 与人工	90.2	92.5	91.7	83.9	90.7
GPT-5 与人工	90.3	85.0	54.8	74.2	73.7

Banana 的主体对象一致性和文本生成质量优于 GPT-Image-1，后者则展现出更强的生成和编辑能力。这些结果与表3中的定量结果一致。更多可视化结果将在补充材料中展示。

4.5. 实证分析

人工验证 DCE。为衡量我们提出的 DCE 的可靠性，我们从 AEGIS 中抽取 10% 的判断题，并由多位专家对评估结果进行人工验证。随后，我们计算与人工验证结果具有相同决策的一致性判断题所占的百分比。评估结果如表 4 （上）所示，总体一致性达到 90.7%，三个多数任务（即理解、生成和编辑）中的一致性也均高于 90%。这些令人鼓舞的结果表明了 DCE 的可靠性。需要注意的是，交错生成任务中的一致性相对较低。通过人工分析这些问题，我们发现，该任务中复杂的“图像-文本一致性检查”判断题提高了评估难度。这一发现为我们未来的研究方向提供了启发。

评判模型选择。此外，我们也对 DCE 中所使用的评判模型选择感到好奇。具体而言，我们使用两个广泛采用的最先进的多模态大语言模型，即 Gemini-2.5-Pro [11] 和 GPT-5 [24], , 使用相同的判断题评估抽取的回答，然后分别计算人工一致性率。评估结果如表 4所示。根据结果，尽管 GPT-5 在理解（90.3 对 90.2）和生成（85.0 对 92.5）任务上的判断一致性相同或相近，但其在编辑和交错生成任务上的判断一致性仍远低于 Gemini-2.5-Pro。一个可行的解释是，Gemini-2.5-Pro 具有更强的推理能力，这表明它能够更准确地捕捉预测结果与参考结果之间的视觉细节和差异。因此，Gemini-2.5-Pro 在编辑和交错生成任务上的一致性远高于 GPT-5。这些结果验证了我们在 DCE 设计中的选择，也为复杂视觉关系分析的未来研究方向提供了启发。

最先进的统一多模态模型是否获得了“世界知识”？ 在验证 DCE 的有效性后，可以进一步分析这些优秀模型是否具备世界知识。如表5所示，我们通过在四项任务中将含糊的、推理密集型提示替换为清晰提示，开展了消融研究：理解（U）、生成（G）、编辑（E）和交错（I）。这一调整消除了模型执行最具挑战性的推理步骤的需要，但表5中的结果显示，所有任务的性能均得到了一致提升。这种增强可归因于模型固有世界知识的激活；在提供清晰提示时，这些知识得到了更充分的利用。有关模块特定问题的进一步分析将在补充材料中详细介绍。

表5。Gemini Nano Banana 与 GPT-4o 在四项任务——理解（U）、生成（G）、编辑（E）和交错（I）——中使用和不使用明确提示时的性能比较。

性能（%）	U	G	E	I	总体
Gemini Nano Banana	61.7	45.9	53.7	45.1	52.9
+ 明确提示	72.9	61.3	64.3	56.4	65.2
GPT-4o+GPT-Image-1	50.2	44.2	45.9	35.5	45.7
+ 明确提示	63.2	61.7	57.7	52.4	60.0

表6. 三种统一多模态模型在不同推理类型任务上的性能比较，即空间、时间、因果、比较、分析、逻辑和无推理任务。

推理类型	Gemini Nano Banana	GPT-Image-1	Bagel w/ CoT
空间	51.2	43.9	25.4
时间	57.3	53.0	25.9
因果	56.2	47.1	25.5
比较	60.9	47.5	24.2
类比	46.4	43.6	23.9
逻辑	49.9	42.5	18.3
无推理	52.8	50.9	21.6

对不同推理类型的分析。如表 6 所示，不同类型的推理也会导致问题难度有所差异。消融实验结果凸显了涉及世界知识的推理任务在空间、时间、因果、比较、类比、逻辑及无推理等类型上的难度差异。Gemini Nano Banana 始终表现最佳，其次是 GPT-Image-1，而 Bagel w/ CoT 在所有类别中均明显表现不佳。时间推理和因果推理是最具挑战性的任务，这反映了编码序列和关系的复杂性；而无需推理的任务则最简单。这些结果强调，需要改进多模态推理框架，以应对处理复杂推理类型时的弱点。

5. 结论

我们提出 AEGIS，这是一个全面的基准测试，用于评估统一多模态模型（UMM）和生成模型在广泛的世界知识和推理类型范围内的视觉理解、生成、编辑和交错生成能力。通过人机协同构建与标注策略，AEGIS 展现了跨任务评估、推理类型影响研究和现实世界应用评估的优势。为确保评判细致且具体，

我们提出一种确定性清单式评估方法，该方法利用一系列原子级“是/否”判断题评估答案中的确定性关键部分，从而提高大语言模型裁判框架的可靠性。大量实验表明，大多数统一多模态模型都存在严重的世界知识缺陷，并且在复杂推理方面面临巨大困难。然而，我们也发现，简单的插件式推理模块可以在一定程度上缓解这些缺陷，为开发更稳健的未来模型提供了一个有前景的方向。

参考结果

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu 和 Junyang Lin。Qwen2.5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025年。1，6[2] Maciej Besta, Lorenzo Paleari, Jia Hao Andrea Jiang, Robert Gerstenberger, You Wu, J ˇAłn Gunnar Hannesson, Patrick Iff, Ales Kubicek, Piotr Nyczyk, Diana Khimey 等。基于思维知识图谱的经济型人工智能助手。arXiv 预印本 arXiv:2504.02670，2025年。2[3] Tim Brooks, Aleksander Holynski 和 Alexei A Efros。InstructPix2Pix：学习遵循图像编辑指令。在 IEEE/CVF 计算机视觉与模式识别会议论文集中，第 18392–18402 页，2023 年。6[4] Huanqia Cai, Yijun Yang 和 Winston Hu。MM-IQ：多模态模型类人抽象与推理能力基准测试。arXiv 预印本 arXiv:2502.00698，2025年。2，3[5] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu 等。混元图像 3.0 技术报告。arXiv 预印本 arXiv:2509.23951，2025年。1[6] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese 等。BLIP3-O：全开放统一多模态模型系列——架构、训练与数据集。arXiv 预印本 arXiv:2505.09568，2025年。2，3，6，7[7] Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang 和 Lifu Huang。R2i-bench：推理驱动的文本生成图像基准测试。arXiv 预印本 arXiv:2505.23493，2025年。3[8] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu 和 Chong Ruan。Janus-pro：通过数据与模型扩展实现统一多模态理解与生成。arXiv 预印本 arXiv:2501 17811，2025 2，6。。[9] Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song 和 Jiahao Zhang。T2vworldbench：用于评估文本生成视频中世界知识的基准测试。arXiv 预印本 arXiv:2507.18107，2025年。3

[10] Yew Ken Chia、Vernon Toh、Deepanway Ghosal、Lidong Bing 和 Soujanya Poria。《Puzzlevqa：借助抽象视觉模式诊断语言模型的多模态推理挑战》。载于《计算语言学协会 ACL 2024 研究成果》，16273，2024，16259，第 3 页–。[11] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、I nderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。《Gemini 2.5：以高级推理、多模态性、长上下文和新一代智能体能力推动前沿》。arXiv 预印本 arXiv:2507.06261，2025。1、2、5、6、7、8、13、14[12] Chaorui Deng、Deyao Zhu、Kunchang Li、Chenhui Gou、Feng Li、Zeyu Wang、Shu Zhong、Weihao Yu、Xiaonan Nie、Ziang Song 等。《统一多模态预训练中的新兴特性》。arXiv 预印本 arXiv:2505.14683，2025。2、3、4、6、7、13[13] Bowen Dong、Minheng Ni、Zitong Huang、Guanglei Yang、Wangmeng Zuo 和 Lei Zhang。《Mirage：评估多模态大语言模型推理链中的幻觉》。arXiv 预印本 arXiv:2505.24238，2025。3[14]Chaoyou Fu、Peixian Chen、Yunhang Shen、Yulei Qin、Mengdan Zhang、Xu Lin、Jinrui Yang、Xiawu Zheng、Ke Li、Xing Sun 等。《MME：多模态大语言模型综合评估基准测试》。arXiv 预印本 arXiv:2306.13394，2023。2、3[15] Chaoyou Fu、Yi-Fan Zhang、Shukang Yin、Bo Li、Xinyu Fang、Sirui Zhao、Haodong Duan、Xing Sun、Ziwei Liu、Liang Wang 等。《MME-survey：多模态大语言模型评测综合综述》。arXiv 预印本 arXiv:2411.15296，2024。3[16] Min Fu、Jiwei Guan、Xi Zheng、Jie Zhou、Jianchao Lu、Tianyi Zhang、Shoujie Zhuo、Lijun Zhan 和 Jian Yang。《ICS-Assist：面向大型电子商务企业在线客户服务的智能客户咨询解决方案推荐》。载于《国际面向服务计算会议》，第 370–385 页。Springer，2020。2[17] Xingyu Fu、Muyu He、Yujie Lu、William Yang Wang 和 Dan Roth。《常识文本生成图像挑战：文本生成图像模型能否理解常识？》。arXiv 预印本 arXiv:2406.07546，2024。3[18] Zigang Geng、Yibing Wang、Yeyao Ma、Chen Li、Yongming Rao、Shuyang Gu、Zhao Zhong、Qinglin Lu、Han Hu、Xiaosong Zhang 等。《X-Omni：强化学习让离散自回归图像生成模型重获新生》。arXiv 预印本 arXiv:2507.22058，2025。2、3[19] Dhruba Ghosh、Hannaneh Hajishirzi 和 Ludwig Schmidt。《GenEval：用于评估文本与图像对齐的以物体为中心的框架》。神经信息处理系统进展，36：52132–52152，2023。2、3[20] Meng-Hao Guo、Jiajun Xu、Yi Zhang、Jiaxi Song、Haoyang Peng、Yi-Xuan Deng、Xinzhi Dong、Kiyohiro Nakayama、Zhengyang Geng、Chen Wang 等。《R-Bench：用于大语言模型和多模态大语言模型复杂推理评测的研究生级跨学科基准测试》。arXiv 预印本 arXiv:2505 02018，2025，2、3。。[21] Xiaotian Han、Quanzeng You、Yongfei Liu、Wentao Chen、Huangjie Zheng、Khalil Mrini、Xudong Lin、Yiqi Wang、Bo-

han Zhai、Jianbo Yuan、Heng Wang 和 Hongxia Yang。《Infimm-Eval：多模态大语言模型的复杂开放式推理评测》，2023。3[22] Xiwei Hu、Rui Wang、Yixiao Fang、Bin Fu、Pei Cheng 和 Gang Yu。《Ella：为扩散模型配备大语言模型，以增强语义对齐》。arXiv 预印本 arXiv:2403.05135，2024。2、3[23] Yuzhou Huang、Liangbin Xie、Xintao Wang、Ziyang Yuan、Xiaodong Cun、Yixiao Ge、Jiantao Zhou、Chao Dong、Rui Huang、Ruimao Zhang 等。《SmartEdit：探索借助多模态大语言模型进行复杂的指令式图像编辑》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 8362–8371 页，2024。3[24] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。《GPT-4o系统卡》。arXiv 预印本 arXiv:2410.21276，2024。1、2、3、4、6、7、8、13、14[25] Dongzhi Jiang、Renrui Zhang、Ziyu Guo、Yanwei Li、Yu Qi、Xinyan Chen、Liuhui Wang、Jianhan Jin、Claire Guo、Shen Yan 等。《MME-CoT：对大型多模态模型中的思维链进行基准测试，评估推理质量、鲁棒性和效率》。arXiv 预印本 arXiv:2502.09621，2025。3[26] Yang Jiao、Haibo Qiu、Zequn Jie、Shaoxiang Chen、Jingjing Chen、Lin Ma 和 Yu-Gang Jiang。《UniToken：通过统一视觉编码协调多模态理解与生成》。载于《计算机视觉与模式识别会议论文集》，第 3600–3610 页，2025。2 、 3[27] Black Forest Labs 、Stephen Batifol、Andreas Blattmann、Frederic Boesel、Saksham Consul、Cyril Diagne、Tim Dockhorn、Jack English、Zion English、Patrick Esser 等。《Flux.1 Kontext：潜在空间中的上下文图像生成与编辑流匹配》。arXiv 预印本 arXiv:2506 15742，2025，1、6。。[28] Bohao Li、Rui Wang、Guangzhi Wang、Yuying Ge、Yixiao Ge 和 Ying Shan。《SEED-Bench：通过生成式理解对多模态大语言模型进行基准测试》。arXiv 预印本 arXiv:2307.16125，2023。2、3[29] Hongxiang Li、Yaowei Li、Bin Lin、Yuwei Niu、Yuhang Yang、Xiaoshuang Huang、Jiayin Cai、Xiaolong Jiang、Yao Hu 和 Long Chen。《GIR-Bench：用于推理生成图像的通用基准测试》。arXiv 预印本 arXiv:2510.11026，2025。3[30] Zijie Li、Henry Li、Yichun Shi、Amir Barati Farimani、Yuval Kluger、Linjie Yang 和 Peng Wang。《用于统一图像生成与理解的双重扩散》。载于《计算机视觉与模式识别会议论文集》，第 2779–2790 页，2025。3[31]Chao Liao、Liyang Liu、Xun Wang、Zhengxiong Luo、Xinyu Zhang、Wenliang Zhao、Jie Wu、Liang Li、Zhi Tian 和 Weilin Huang。《Mogao：用于交错式多模态生成的全能基础模型》。arXiv 预印本 arXiv:2505.05472，2025。3[32] Haowei Lin、Xiangyu Wang、Ruilin Yan、Baizhou Huang、Haotian Ye、Jianhua Zhu、Zihao Wang、James Zou、Jianzhu Ma、Yitao Liang。大型语言模型中复杂推理的生成式评估。arXiv 预印本 arXiv:2504.02810，2025年。3[33] Shiyu Liu、Yucheng Han、Peng Xing、Fukun Yin、Rui Wang、Wei Cheng、Jiaqi Liao、Yingming Wang、Honghao Fu、Chun-rui Han 等。Step1x-edit：通用图像编辑的实用框架。arXiv 预印本 arXiv:2504.17761，2025年。2、3[34] Shiyu Liu、Yucheng Han、Peng Xing、Fukun Yin、Rui Wang、Wei Cheng、Jiaqi Liao、Yingming Wang、Honghao Fu、Chun-rui Han 等。Step1x-edit：通用图像编辑的实用框架。arXiv 预印本 arXiv:2504.17761，2025年。6[35]Yiyang Ma、Xingchao Liu、Xiaokang Chen、Wen Liu、Chengyue Wu、Zhiyu Wu、Zizheng Pan、Zhenda Xie、Haowei Zhang、Xingkai Yu 等。Janusflow：协调自回归与整流流，以实现统一的多模态理解与生成。载于《计算机视觉与模式识别会议论文集》，第7739–7751页，2025年。3 – 。[36] Gr´egoire Mialon、Cl´ementine Fourrier、Thomas Wolf、Yann LeCun 和 Thomas Scialom。Gaia：通用人工智能助手的基准测试。载于第十二届国际学习表征会议，2023年。2[37] Yuwei Niu、Munan Ning、Mengren Zheng、Weiyang Jin、Bin Lin、Peng Jin、Jiaqi Liao、Chaoran Feng、Kunpeng Ning、Bin Zhu 等。WISE：用于文本生成图像的世界知识驱动语义评估。arXiv 预印本 arXiv:2503.07265，2025年。2、3、5、14[38]Dustin Podell、Zion English、Kyle Lacey、Andreas Blattmann、Tim Dockhorn、Jonas M¨uller、Joe Penna 和 Robin Rombach。SDXL：改进潜在扩散模型，实现高分辨率图像合成。arXiv 预印本 arXiv:2307.01952，2023年。1[39] Team Seedream、Yunpeng Chen、Yu Gao、Lixue Gong、Meng Guo、Qiushan Guo、Zhiyao Guo、Xiaoxia Hou、Weilin Huang、Yixuan Huang 等。Seedream 4.0：迈向新一代多模态图像生成。arXiv 预印本 arXiv:2509.20427，2025年。6[40]Qingyu Shi、Jinbin Bai、Zhuoran Zhao、Wenhao Chai、Kaidong Yu、Jianzong Wu、Shuangyong Song、Yunhai Tong、Xiangtai Li、Xuelong Li 等。Muddit：借助统一离散扩散模型，实现超越文本生成图像的生成。arXiv 预印本 arXiv:2505.23606，2025年。3[41] Weijia Shi、Xiaochuang Han、Chunting Zhou、Weixin Liang、Xi Victoria Lin、Luke Zettlemoyer 和 Lili Yu。LMFusion：将预训练语言模型用于多模态生成。arXiv 预印本 arXiv:2412.15188，2024年。3[42] Kaiyue Sun、Rongyao Fang、Chengqi Duan、Xian Liu 和 Xihui Liu。T2I-ReasonBench：推理驱动的文本生成图像基准测试。arXiv 预印本 arXiv:2508.17472，2025年。2、3、4、5[43] Alexander Swerdlow、Mihir Prabhudesai、Siddharth Gandhi、Deepak Pathak 和 Katerina Fragkiadaki。统一多模态离散扩散。arXiv 预印本 arXiv:2503 20853，2025年 3 。 。[44] Kimi Team、Angang Du、Bohong Yin、Bowei Xing、Bowen Qu、Bowen Wang、Cheng Chen、Chenlin Zhang、Chen-zhuang Du、Chu Wei 等。Kimi-VL 技术报告。arXiv 预印本 arXiv:2504.07491，2025年。6

[45]Patrick von Platen、Suraj Patil、Anton Lozhkov、Pedro Cuenca、Nathan Lambert、Kashif Rasul、Mishig Davaadorj、Dhruv Nair、Sayak Paul、William Berman、Yiyi Xu、Steven Liu 和 Thomas Wolf。Diffusers：最先进的扩散模型。https://github.com/huggingface/ diffusers，2022年。6[46] Chenglin Wang、Yucheng Zhou、Qianning Wang、Zhe Wang 和 Kai Zhang。ComplexBench-Edit：通过组合依赖关系对复杂指令驱动图像编辑进行基准测试。arXiv 预印本 arXiv:2506.12830，2025年。3[47] Guo-Hua Wang、Shanshan Zhao、Xinjie Zhang、Liangfu Cao、Pengxin Zhan、Lunhao Duan、Shiyin Lu、Minghao Fu、Xiao-hao Chen、Jianshan Zhao 等。Ovis-U1 技术报告。arXiv 预印本 arXiv:2506.23044，2025年。2、6、7[48]Haoxin Wang、Xianhan Peng、Xucheng Huang、Yizhe Huang、Ming Gong、Chenghan Yang、Yang Liu 和 Ling Jiang。Ecom-Bench：大语言模型智能体能否解决现实世界中的电子商务客户支持问题？arXiv 预印本 arXiv:2507.05639，2025年。2[49] Jin Wang、Yao Lai、Aoxue Li、Shifeng Zhang、Jiacheng Sun、Ning Kang、Chengyue Wu、Zhenguo Li 和 Ping Luo。Fudoki：通过动能最优速度实现基于离散流的统一理解与生成。arXiv 预印本 arXiv:2505.20147，2025年。3[50] Xinlong Wang、Xiaosong Zhang、Zhengxiong Luo、Quan Sun、Yufeng Cui、Jinsheng Wang、Fan Zhang、Yueze Wang、Zhen Li、Qiying Yu 等。Emu3：你只需要下一词元预测。arXiv 预印本 arXiv:2409.18869，2024年。2、3、6[51] Haoran Wei、Yaofeng Sun 和 Yukun Li。Deepseek-OCR：上下文的光学压缩。arXiv 预印本 arXiv:2510.18234，2025年。1[52] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示能够激发大型语言模型的推理能力。《神经信息处理系统进展》，35：24824–24837，2022年。13[53] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chau-mond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、R´emi Louf、Morgan Funtowicz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander M. Rush。Transformers：最先进的自然语言处理技术。载于《2020年自然语言处理实证方法会议论文集：系统演示》，第38–45页，线上，2020年。计算语言学协会。6[54] Chenfei Wu、Jiahao Li、Jingren Zhou、Junyang Lin、Kaiyuan Gao、Kun Yan、Sheng-ming Yin、Shuai Bai、Xiao Xu、Yilei Chen 等。Qwen-Image 技术报告。arXiv 预印本 arXiv:2508.02324，2025年。2、3、6、7[55] Chenyuan Wu、Pengfei Zheng、Ruiran Yan、Shitao Xiao、Xin Luo、Yueze Wang、Wanli Li、Xiyan Jiang、Yexin Liu、Junjie Zhou 等。OmniGen2：探索先进的多模态生成。arXiv 预印本 arXiv:2506.18871，2025年。2、3[56] Yongliang Wu、Zonghui Li、Xinting Hu、Xinyu Ye、Xian-fang Zeng、Gang Yu、Wenbo Zhu、Bernt Schiele、Ming-Hsuan Yang、Xu Yang。Kris-bench：下一代智能图像编辑模型的基准测试。arXiv 预印本 arXiv:2505.16707，2025年。3，4[57] Jinheng Xie、Zhenheng Yang 和 Mike Zheng S hou。Show-o2：改进的原生统一多模态模型。arXiv 预印本 arXiv:2506.15564，2025年。6[58] Ling Yang、Ye Tian、Bowen Li、Xinchen Zhang、Ke Shen、Yunhai Tong 和 Mengdi Wang。Mmada：多模态大型扩散语言模型。arXiv 预印本 arXiv:2505.15809，2025年。3[59] Yang Ye、Xianyi He、Zongjian Li、Bin Lin、Shenghai Yuan、Zhiyuan Yan、Bohan Hou 和 Li Yuan。Imgedit：统一的图像编辑数据集与基准测试。arXiv 预印本 arXiv:2505.20275，2025年。2，3[60] Daoan Zhang、Che Jiang、Ruoshi Xu、Biaoxiang Chen、Zi-jian Jin、Yutian Lu、Jianguo Zhang、Liang Yong、Jiebo Luo 和 Shengda Luo。Worldgenbench：面向推理驱动的文本生成图像的世界知识集成基准测试。arXiv 预印本 arXiv:2505.01490，2025年。2，3，5[61] Shihao Zhao、Dongdong Chen、Yen-Chun Chen、Jianmin Bao、Shaozhe Hao、Lu Yuan 和 Kwan-Yee K Wong。Uni-controlnet：面向文本到图像扩散模型的一体化控制。神经信息处理系统进展，36：11127–11150，2023年。1[62] Xiangyu Zhao、Peiyuan Zhang、Kexian Tang、Xiaorong Zhu、Hao Li、Wenhao Chai、Zicheng Zhang、Renqiu Xia、Guang-tao Zhai、Junchi Yan 等。超越像素的构想：推理引导的视觉编辑基准测试。arXiv 预印本 arXiv:2504.02826，2025年。3[63] Pengfei Zhou、Xiaopeng Peng、Jiajun Song、Chuanhao Li、Zhaopan Xu、Yue Yang、Ziyao Guo、Hao Zhang、Yuqi Lin、Yefei He 等。Opening：用于评判开放式交错图文生成的综合基准测试。载于《计算机视觉与模式识别会议论文集》，第56–66页，2025年。3[64] Jinguo Zhu、Weiyun Wang、Zhe Chen、Zhaoyang Liu、Shenglong Ye、Lixin Gu、Hao Tian、Yuchen Duan、Weijie Su、Jie Shao 等。Internvl3：探索开源多模态模型的高级训练与测试时方案。arXiv 预印本 arXiv:2504.10479，2025年。1

在第A节中，我们阐述了 AEGIS 所涵盖主题类型的详情。在第B节中，我们说明了 AEGIS 中各类推理类型的具体情况。在第C节中，我们进行了更深入的分析，以揭示世界知识评估中的关键结果。而在第D节中，我们介绍了 AEGIS 中使用的核心提示词。

A. AEGIS 主题类型说明

AEGIS 将现实世界知识划分为三个领域（科学、技术、工程和数学、人文学科、日常生活），并将每个主题进一步细分为更具体的子主题，以评估其互补的不同方面。

A.1. 科学、技术、工程和数学

科学、技术、工程和数学主题用于评估科学、技术、工程和数学方面的能力，重点考察定量推理、物理和数学原理的应用，以及基于形式化方法的问题解决能力。包括：

• **生物学**评估与生物学常识相关的知识，包括代表性物种及其生态特征、基本生命过程，以及具有文化相关性的生物学概念。示例：请画出 <图像*1*> 中动物在亚洲的现代女性亲属。

•**化学**关注化学物质、日常化学现象，以及融入传统工艺的化学知识，涵盖常见物质的用途、安全意识，以及不同文化和行业中广为人知的过程。示例：已知 <图像*1*>和 <图像*2*>中所示的两种化合物，在浓硫酸中预计会生成什么反应产物？

•**数学** 评估对基础数学概念、常见几何图形和日常应用的理解，重点关注算术、测量和形状识别，这些内容通常在不同文化中均有教授。示例：图像 <图像*1*>所示代表一个三次函数。若三次项的系数为 *1/2*，线性项的系数是多少？

• **医学** 考察基本的医疗与健康素养，包括疾病预防、急救基础知识，以及传统和现代医疗保健实践中广泛认可的概念或工具。示例：水杨酸 <图像*1*>如何增强治疗效果？请画出改良药物的键线式。

• **物理学**聚焦日常物理现象和基础概念——力学、热力学、电磁学、光学和声学，并着重介绍直观的现实应用及其解释。示例：根据薄膜干涉原理，请为图中的空白区域上色<图像*1*>。

• **天文学与地理**评估对典型天体和地理特征的识别能力，包括肉眼可见的天空现象、季节知识与方向知识，以及具有文化象征意义的地标和生物群系。示例：<图像*1*> 展示了北纬*60*度某地在*1908*年之前的景象。请画出同一地点在*1908*年之后的景象。

• **信息技术** 关注常见的数字素养和信息技术概念，包括基础计算与网络知识、日常数据和安全实践，以及广泛使用的软件和硬件术语。示例： <图像*1*>展示了 *CPU* 架构示意图。请使用相同的配色方案，绘制另一种常见计算芯片的架构示意图。

A.2. 人文学科

人文学科主题评估对人类社会、文化和创造性表达的理解，重点考察诠释性推理、历史分析、语境理解，以及对人工制品和实践的批判性评估。包括：

• **农业**评估对不同地区农业实践、农作物、工具和食物体系的知识，包括传统和现代方法及其文化与经济意义。示例：请根据图像中提供的标签 <图像*1*>，按照杂交水稻种植面积占水稻总种植面积的比例为地图着色 <图像*2*> 。

• **历史**考察对重大历史事件、历史时期、历史人物和人工制品的辨识，重点关注年代顺序、因果关系和文化影响。示例：清朝时期的哪一事件与这张图像中所示的事件相似<图像*1*>?

• **电影**评估对有影响力的电影、类型、导演、标志性场景和具有重要文化意义的电影符号的熟悉程度。示例：*1942*年奥斯卡最佳原创剧本奖获奖作品中的麦高芬是什么？

• **音乐**关注音乐传统、乐器、类型以及知名作曲家或表演者，着重展现风格特征与文化背景。示例：生成一份《一闪一闪小星星》*C*大调的简易乐谱。

• **艺术**评估对视觉艺术、风格、技法、艺术运动以及跨文化和不同时代的经典作品或艺术家的理解。示例：编辑 <图像*1*>，让蒙娜丽莎移开视线，露出轻蔑的表情，并举起一块牌子，表示她不想被拍照。

• **文化**评估界定集体身份认同和社会生活的广泛文化实践、规范、遗产项目与象征。示例：将 <图像*1*> 中间三个糕点的外皮替换为华北风格。

•**建筑**测试对建筑风格、结构特征、地标建筑以及建成环境的历史与技术背景的识别能力。示例：与 <图像*1*>所在城市相同的另一个宗教观光景点是什么？

A.3. 日常生活

日常生活主题评估人们在常见现代情境中的实用知识和日常推理能力，重点考察情境理解、日常决策，以及对日常环境中接触到的工具、活动和媒体的识别。

• **活动**评估对常见日常活动和休闲活动的熟悉程度，包括其典型工具、场景和操作步骤。例如：根据这张屏幕截图 <图像*1*>，是否有阿根廷球员处于越位位置？

• **动漫**评估对知名动漫系列、角色、视觉惯例和风格惯例的识别能力，以及对动画媒体中具有文化意义的符号的辨识能力。例如：在向女帝告别之后、目睹兄长身亡之前，《海贼王》的主角去了哪里？

• **游戏** 侧重于对电子游戏和桌面游戏的理解，包括知名作品、类型、游戏玩法元素，以及独特的游戏内物品或界面。例如：*2022*年*TGA*年度移动端游戏获奖作品的团队负责人还制作过哪款知名游戏？

•**摄影** 考察图像采集和编辑流程中所使用的摄影设备、技法、类型及视觉惯例等方面的知识。 示例：将女孩的照片从插图 <图像*1*>绘制成世界上最著名动漫展中的角色扮演照片。但要保留原始背景。

• **工程** 评估与日常环境相关的日常工程器具、机械装置、家用设备以及基本技术操作方面的实践理解。 示例：向<图像*1*> 中添加适当的材料，使其成为一个简单的蒸馏器。

• **食品** 评估对日常饮食中常见的食材、菜肴、烹饪方法、饮食习俗及营养概念的识别能力。 示例：为这道菜添加四川特色 <图像*1*>。

• **交通** 测试识别城市出行中所使用的交通方式、道路标志、交通规则和导航惯例的能力。 示例：在<图像*1*>中绘制从塔楼位置到红点位置的最快轨道交通路线。

B. AEGIS推理类型描述

除了通用世界知识外，AEGIS 还通过评估大语言模型的底层推理能力，进一步探究其遵循混淆指令的能力——具体而言，AEGIS 将推理分为六种类型：

• **空间推理**评估在二维/三维空间中推断位置、距离、方向、包含关系以及部分与整体布局之间关系的能力，占整个基准测试的 10.9%。示例：给定一个 *3D* 对象的正视图 <图像*1*>、俯视图 <图像*2*>以及右侧视图 <图像*3*> ，绘制其等轴测投影图。

• **时间推理**评估对时间顺序、持续时间、并发情况和日程安排的理解，包括先后关系和时间线一致性，占整个基准测试的 12.2%。示例：对其进行编辑，以展示 <图像*1*> 今天的样子。

•**因果推理**考察识别因果关系、必要/充分条件，以及干预或反事实变化所产生结果的能力，占整个基准测试的 12.2%。示例：根据 <图像*1*>和 <图像*2*>推断出一个统一的天文事件。

• **比较推理**涉及沿一个或多个维度对两个或更多实体进行比较，并根据它们的相对差异或排名得出结论，占整个基准测试的 15.4%。示例： <图像*1*> 和 <图像*2*>，哪个需要更多烹饪步骤？

•**类比推理**评估将已知情境中的关系结构映射到新情境的能力，识别比例或功能类比，占整个基准测试的 9.4%。示例：正如<图像*1*> 对应于其动漫作品一样，在火影忍者中，哪个角色处于类似的位置？

• **逻辑推理**强调在日常情境中，根据给定的事实、规则或约束得出连贯的结论，占整个基准测试的 36.3%。示例：绘制一个包含三个相交集合 A、B 和 C 的维恩图，并将对应于 $(A \cap B) \cup C$ 的区域涂阴影。

C. 补充实验

在本节中，我们基于 Gemini Nano Banana [11]（简称 Gemini）和搭载 GPT-Image 的 GPT-4o 1 [24] （简称 GPT）开展补充实验，以考察具体程度不同的提示词如何影响各项任务的表现。我们还通过受控消融实验，将常识知识与统一多模态模型分离，以隔离模块特定问题并量化其影响。此外，我们还通过评估最先进的 Gemini-3-Pro（即Nano Banana Pro）来探究统一多模态模型的性能上限。

C.1. 基于 UMM 改写提示词的评估

除了研究外部推理模块 [52] 对 Bagel [12]，的影响外，我们还进行了消融研究

表 7. Gemini Nano Banana（简称 Gemini）与 GPT-4o 搭配 GPT-Image-1（简称 GPT）在不同类型的提示和外部网络搜索工具下的综合性能比较。

模型	理解			生成			编辑			交错生成			总体
	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	科学、技术、工程和数学	人文	生命	
Gemini	64.5	65.7	55.0	42.6	49.5	45.5	44.4	62.4	54.2	50.2	41.6	43.4	52.9
Gemini（使用网络搜索）	64.9	69.2	60.0	-	-	-	-	-	-	-	-	-	-
Gemini（使用 GPT 提示词）	63.8	64.7	56.2	37.8	46.0	44.3	41.0	61.4	51.3	47.3	43.6	43.3	51.1
Gemini（使用 Gemini 提示词）	63.4	72.9	57.6	45.3	57.9	52.2	42.2	63.9	47.8	50.8	49.6	47.2	55.2
使用明确提示的 Gemini	72.4	74.3	72.1	53.6	67.6	62.6	54.2	68.4	70.4	53.1	62.0	54.0	65.2
使用明确提示和网络的 Gemini	72.8	80.8	73.3	-	-	-	-	-	-	-	-	-	-
GPT	52.9	50.9	46.9	38.2	51.6	42.8	39.4	53.2	45.2	38.9	34.7	33.0	45.7
使用 GPT 提示词的 GPT	48.4	52.0	43.4	35.6	52.8	41.9	38.1	54.9	46.0	33.5	33.6	34.7	44.7
使用 Gemini 提示词的 GPT	57.0	66.7	53.2	42.6	57.3	47.0	39.0	59.3	45.4	44.6	36.1	41.8	50.8
使用清晰提示词的 GPT	61.6	65.4	62.7	52.1	71.5	61.6	49.2	66.9	56.9	51.0	57.2	49.2	60.0
Gemini-3-Pro	77.7	79.3	70.4	62.2	64.4	58.2	64.1	67.8	58.5	42.6	40.2	38.5	64.3

以进一步单独考察推理质量的影响。具体而言，我们采用了一种“自我推理”策略：模型首先改写原始提示词，生成一条“明确提示”，从而通过消除歧义、识别实体并明确隐含语境，减少后续推理的需求。出乎意料的是，尽管人工核验的明确提示通常能带来显著提升，但我们发现**自我改写产生了截然不同的效果**：经 GPT-4o 改写的提示词 [24] 与原始输入相比导致性能下降，而经 Gemini Nano Banana 改写的提示词 [11] 则提升了性能。为验证这一差异，我们通过互换改写后的提示词进行了跨模型评估，也就是说，将 Gemini 改写的提示词输入 GPT-4o，反之亦然。结果一致：GPT 生成的改写内容会导致各模型性能下降，而 Gemini 生成的改写内容则始终带来提升。这些发现有力地表明，在消除指令歧义方面，Gemini Nano Banana 的推理能力优于 GPT-4o。因此，利用具备高级推理能力的大语言模型，有望缓解统一多模态模型中模糊指令或需要大量推理的指令所带来的挑战，从而使广泛的世界知识领域中的各类任务受益 [37]。

C.2. Gemini-3-Pro 的世界知识更丰富

如第 4节所述，引入质量更高、领域方面更丰富的训练数据，可以提升模型对世界知识的理解能力。然而，其他任务是否也能从中受益尚未得到验证。因此，我们在 AEGIS 基准测试上评估了 Gemini-3-Pro，即 Gemini Nano Banana 的扩展版本。如表 7底部所示，即使推理时使用了推理增强提示，Gemini-3-Pro 的总体表现（64.3）仍远高于 Gemini（52.9），且与使用清晰提示的 Gemini 表现相当。

这些令人鼓舞的结果表明，更好的预训练数据有助于提升世界知识能力，也说明 Gemini-3-Pro 是在世界知识理解和生成方面达到最先进水平的统一多模态模型。

C.3. 网络搜索评估

在研究如何消除复杂指令的歧义之后，我们进一步探索了通过融入新的外部知识来减少预测错误和幻觉的方法。直观来看，整合搜索引擎应能提供准确作答所需的关键信息和最新信息。为评估这一点，我们开展了一项消融研究，评估网络搜索增强对模型性能的影响。具体而言，我们为 Gemini（具体来说是 Gemini-2.5-Flash-Image）启用了 Google 搜索工具，使其能够依据时事和可验证的网络事实生成回答。出乎意料的是，启用网络搜索后，模型在理解任务上的性能仅略有提升。尤其是在科学、技术、工程和数学主题的问题上，性能仅提升了 0.4。一个合理的解释是，尽管搜索工具能够有效获取外部世界知识——这对于核实事实或检索近期事件至关重要——但它们并不能从根本上增强模型的核心推理能力。这一发现与人类解决问题的行为相符：有效的解决方案很少来自将复杂、未经处理的问题直接输入搜索引擎。相反，成功的问题解决通常需要先进行推理，形成清晰且有针对性的查询，然后再查阅外部资源。为验证这一假设，我们进一步在 Gemini 中集成网络搜索工具并提供清晰提示，以探究更清晰、更有效的问题描述是否能带来更精确的搜索结果，作为辅助知识。如表 7所示，通过使用清晰提示简化问题，理解任务的表现大幅提升。尤其是，人文和生命理解问题的性能均提升了 10 以上

，这表明，对于问题描述清晰的情况，网络搜索工具能够提升统一多模态模型的世界知识能力。这些结果也表明，在推理过程中，统一多模态模型固有的推理能力至关重要。

C.4. 对模块特定瓶颈的调查

此外，我们旨在确定哪个组件是统一多模态模型世界知识能力的主要限制因素：大语言模型组件还是视觉解码器组件。为找出错误来源，我们分析了 Gemini 的失败案例。具体而言，我们利用 Gemini-2.5-Pro 通过自我反思过程改写原始提示词，确保所有隐含知识都得到明确表达。随后，我们将这些改写后的提示词重新输入 Gemini 进行生成。图5 展示了根据原始提示、改写提示词和经验证的清晰提示生成的图像对比。值得注意的是，我们发现，大语言模型组件能够在改写后的文本中成功表达关键视觉属性（例如，正确识别“胡桃”或“米开朗基罗”）。然而，视觉解码器未能始终如一地呈现这些概念，其结果偏离了清晰提示生成的图像和真实情况。这种差异表明，视觉解码器限制了统一多模态模型的世界知识能力，这可能是因为解码器本身的知识编码不足，或对输入措辞极其敏感。

为严格验证这一假设，我们使用极其详尽的描述开展了一项后续实验。我们使用 Gemini-2.5-Pro 生成全面的视觉描述，明确列出生成或编辑所需的每个关键点，从而有效绕过模型回忆视觉属性的需求。随后，我们将这些描述输入 Gemini。如图 8 所示，尽管大语言模型提供了高度准确且详尽的视觉指令，只有第一个示例呈现出合理的结果（形状略有偏差），其余结果仍不正确。这些结果明确证实，瓶颈在于视觉生成模块，表明模型较强的文本理解能力与较弱的视觉生成能力之间存在不匹配，这与 AEGIS 中观察到的性能差距一致。

D. AEGIS 中使用的必要模板

最后，我们提供 AEGIS数据集标注中使用的模板，包括图 9 中的清单生成提示，以及图 10中的评估提示。



图5. 使用原始提示词和经大语言模型改写提示词的失败案例可视化。我们用红色突出显示答案中的关键点。尽管外部推理模块（例如 Gemini）可以通过改写复杂提示词来降低生成难度，但在涉及世界知识各个方面的多样化任务中，推理能力仍存在不足。

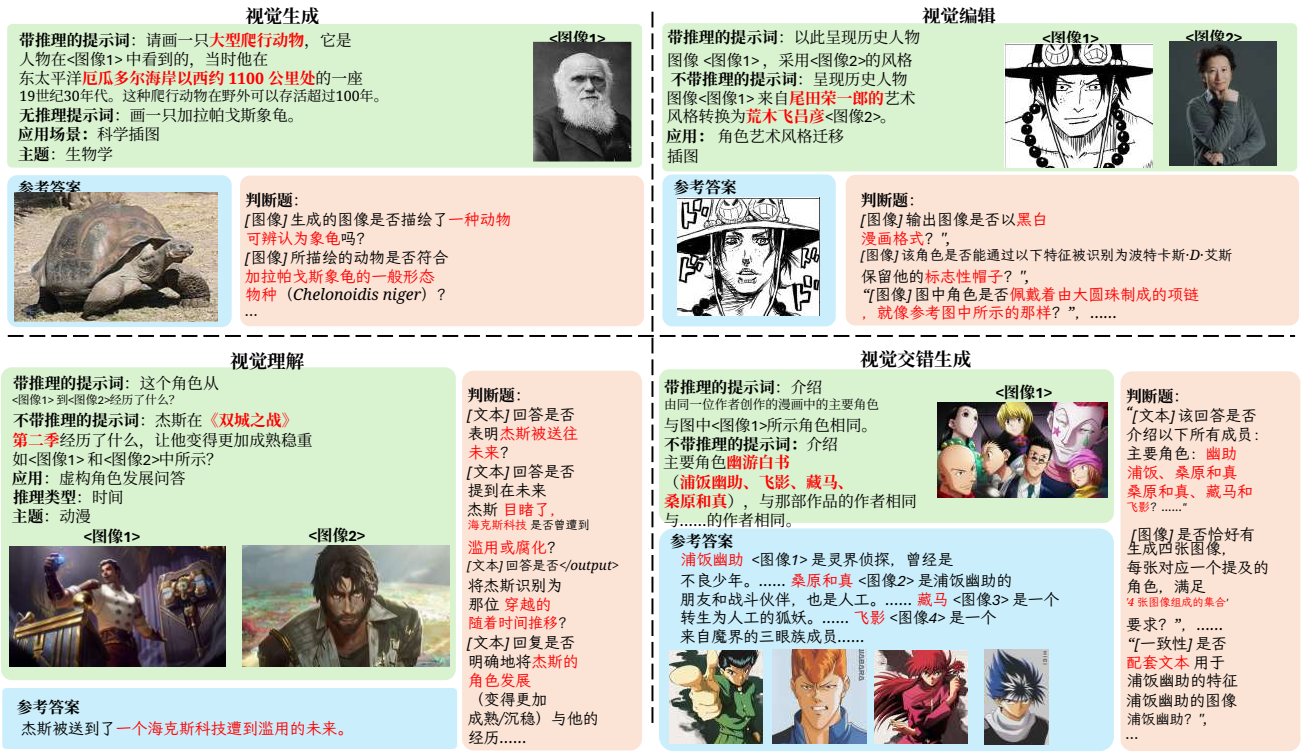


图6. AEGIS 基准测试中更多问题的可视化。

空间推理



<图像1>

哪些农作物适合种植在
<图像1>所描绘的阴凉区域?

参考文本：玉米和大豆

时间推理

维多利亚时代英国情侣会在
二月中旬进行的一项习俗。



<参考_图像>

因果推理

使用万圣节食品作为基础
用于感恩节食品。



<参考_图像>

比较推理





<图像1> **<图像2>**

<图像1> 和 <图像2>，这需要
更多烹饪步骤?

参考文本：第一张图像

类比推理







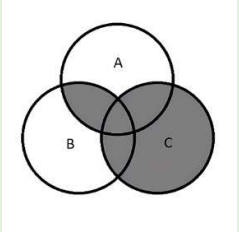
<图像1> **<图像2>** **<图像3>**

根据<图像1> 和<图像2>，识别
与<图像3>对应的角色。

参考文本：空条承太郎

逻辑推理

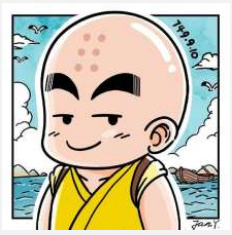
绘制一个包含三个
相交集合 A、B 和 C 的维恩图，并将
对应于 (A ∩ B) ∪ C 的区域标灰。



<参考_图像>

图7. 不同推理类型问题的可视化。AEGIS的问题涵盖多种推理类型，涉及实际应用中的常见场景。

<图像1>



无需推理的提示词：绘制一幅加拉帕戈斯象龟的图片<图像>的主人居住的地方。

参考图像



大语言模型预期：一座小巧、色彩鲜艳的粉红色房屋，标有“龟仙屋”坐落在一座热带岛屿上，周围环绕着清澈湛蓝的海水，伴有 棕榈树。

统一多模态模型输出



<图像1>



无推理提示词：根据<电视剧画出Gatanothor。1>

参考图像



大语言模型的预期：一个庞大而怪诞的生物，模样像是一只巨大的贝壳长着触手般的附肢，闪烁着红光眼睛，并且从水中浮现出一种黑暗而诡异的气息。

统一多模态模型输出



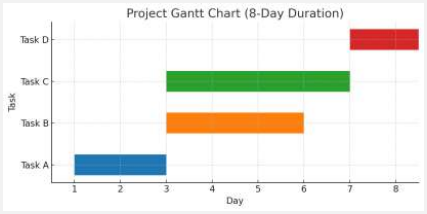
纯文本

无推理提示词：

创建一张甘特图，表示一个总工期为 8 天的项目：

任务 A：第 1–2 天
任务 B：第 3–5 天
任务 C：第3天至第6天
任务 D：第7天至第8天

参考图像



大语言模型预期：任务 A：一条从第1天开始，并于第2天结束的条形。任务 B：一条从第3天开始，并于第5天结束的条形。任务 C：一条从第3天开始，并于第6天结束、与任务 B 部分重叠的条形。任务 D：一根从第7天开始、到第8天结束的横条。

统一多模态模型输出



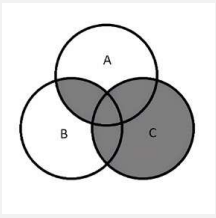
纯文本

无推理提示词：

绘制三个相互重叠的圆，分别标记为 A、B 和 C，默认不填充颜色。

将圆 A 和圆 B 的重叠区域涂成灰色，并将圆 C 的整个区域填充为灰色。

参考图像



大语言模型预期：一幅维恩图，其中三个相互重叠的轮廓圆分别标记为 A、B 和 C，A–B 重叠部分涂为灰色，整个圆 C 填充为灰色。

统一多模态模型输出

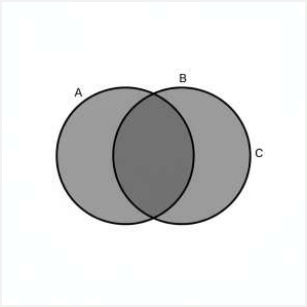


图8. 使用 Gemini 对清晰提示进行润色后的大语言模型描述所对应的失败案例可视化。可以发现，即便这些描述准确地阐明了答案，视觉解码器通常仍难以生成正确答案。

任务：全面的任务感知评估问题生成

目标：
你的任务是担任评估专家。你的目标是生成一份详细、客观且可验证的问题检查清单，用于评估多模态模型的输出。这些问题必须以所提供的参考材料（`参考文本` 和 `参考图像`）为依据，同时确保涵盖 `关键词` 和 `类别`所界定的核心概念。

输入：
- **任务类型：** {task}（任务的类别，决定输出模态：`理解`、 `编辑` ， `生成`或 `交错`。）- **关键词：** {keywords}（必须涵盖的一组核心概念。）- **类别：** {category}（用于提供上下文关联的更广泛主题。）- **明确提示：** {clear_提示词}（提供给待评估模型的原始指令。）- **参考文本：** {ref_text}（文本事实的**主要事实来源**。）- **参考图像：** <ref_images_placeholder_list> （视觉事实的**主要事实来源**。）

核心指令：

1. **根据任务类型确定输出模态：**这是第一步，也是最重要的一步。它决定了你可以使用哪些问题标签。* **

If `Task 类型` is `understanding` (仅文本 ly 输出):** 你必须 **ONLY** 使用 the `[文本]` 标签 .
* **如果 `任务类型` 为 `编辑` 或 `生成` （仅图像输出）： ** 你必须**仅**使用 `[图像]` 标签。
* **如果 `任务类型` 为 `交错` （文本和图像输出）： ** 你可以使用 `[文本]`、 `[图像]`和 `[一致性]` 标签。`[一致性]` 问题在此至关重要。

2. **以参考材料为依据提出问题：** 所有问题都必须源自`参考文本` 和/或`参考图像`中具体且可验证的细节。不要提出无法通过参考材料回答的问题。

3. **利用关键词和类别聚焦问题：** 将`关键词` 和`类别` 作为聚焦方向。优先针对参考材料中与这些关键词及总体主题最相关的细节提出问题。例如，如果某个关键词是“鸡蛋”，则根据`参考文本`的描述，针对鸡蛋的使用方式（例如，分离、打发）提出具体问题。

4. **提出具体、客观且带标签的问题：** * 每个问题前都必须加上模态标签（`[文本]`、 `[图像]`、 `[一致性]`）。* 问题必须客观且基于事实。**避免**对质量、风格、语气、主题或艺术价值进行主观评判。* **错误（主观）：** `[图像]` 这张照片在美学上是否令人愉悦？ * **错误（模糊）： ** `[文本]` 这段文本是否谈到了这些关键词？* **正确（具体且客观）： ** `[文本]` 食谱是否注明要以 150°C 烘烤蛋糕 60 分钟？

5. **确保覆盖全面且无冗余：** 最终检查清单应涵盖与关键词相关的所有关键方面，避免针对同一细节反复提问。

输出格式：
注意，输出应为一个字符串，可通过 json.loads() 函数直接转换为 Python 列表；其值应为一个字符串数组，其中每个字符串都是一个带标签的评估问题。

现在为提供的输入生成检查清单。

图9. AEGIS 基准测试中的清单生成提示。我们将 LLM 作为评判者的评估设计为一系列原子“是/否”问题，以避免产生模糊的判断。

任务：全面的任务感知评估

目标：
你是一位基于世界知识进行评估的专家。你的任务是核验输出图像或文本是否
符合一系列检查清单，并提供评估推理。

输入格式：
你将获得以下字段：
- **任务类型**：任务的类别，用于决定输出模态：`理解`、 `编辑`、
`生成`或`交错`。
- **类别**：用于体现上下文相关性的更广泛主题。
- **明确提示**：向模型发出的原始指令，用于输出图像、文本或两者。
- **参考文本**：文本事实的**主要事实来源**。
- **参考图像**：视觉事实的**主要事实来源**。
- **输出文本**：需要由你评估的输出文本。
- **输出图像**：需要由你评估的输出图像。
- **检查清单**：针对输出的一系列检查清单。每个检查清单项目都包含以下标签之一：`[文本]`,
`[图像]`，以及 `[一致性]`。

对于带有标签 `[文本]` 的项目，请重点分析 **输出文本** 是否符合
检查清单项目的要求。你可以将 **参考文本** 视为完美符合所有检查清单项目要求的参考。
不过，输出内容不必与**参考文本**完全相同。只要符合
检查清单项目的要求，即可标记为通过（Y）。
对于带有标签 `[图像]` 的项目，请重点分析**输出图像**是否符合
检查清单项目的要求。你可以将**参考图像**视为完全符合所有检查清单项目要求的参考。
不过，输出内容不必与**参考图像**完全相同。只要符合
检查清单项目的要求，即可标记为通过（Y）。
对于带有 `[一致性]` 标签的条目，请重点分析 **输出文本** 和 **输出
图像** 是否符合检查清单条目的要求。你可以将 **参考文本**
和 **参考图像** 视为完全符合所有检查清单条目的参考内容。不过，输出不
必与 **参考文本** 和 **参考图像** 完全相同。只要符合检查清单条目的要求，
即可标记为通过（Y）。

任务与输出要求：
你的输出必须是单个有效的 JSON 对象。该 JSON 对象应为包含以下键的字典：
- **答案列表**：检查清单各项的答案列表。答案列表中的每一项对应
检查清单中的相应条目，顺序一致。每项为“Y”或“N”，分别表示“是”或“否”。
- **理由列表**：评估的推理过程。理由列表中的每一项都说明
答案列表中相应“Y”或“N”的原因。

根据上述所有要求评估输出。确保输出为有效的 JSON。

图 10. AEGIS 基准测试中的 DCE 评估提示。我们对所有原子判断问题预测“是/否”，并将“是”的比例作为最终得分。