大型多模态模型在材料科学中的跨尺度感知能力基准测试

Yuting Zheng[1,2],Zijian Chen[1],[2]*, Jia Qi[1]上海人工智能实验室[2]上海交通大学 邮箱：{zhengyt058,zijian.chen}@sjtu.edu.cn, jiaqi@pjlab.org.cn

arXiv:2603.19327v1 [cs.DL] 18 Mar 2026

摘要——**揭示层次化的结构—性能关系是材料科学的核心挑战，这需要解读涵盖从微观到宏观的广阔物理尺度的数据。尽管大型多模态模型（大型多模态模型）迅速融入科学工作流程，现有科学基准测试主要侧重于通用图表解读或孤立的常识推理，未能评估模型跨越复杂物理维度的推理能力。为此，我们推出了CSMBench，这是一个由截至2025年9月发表于顶级期刊的1,041幅高质量图像组成的数据集。CSMBench 将数据划分为四种科学上迥异的尺度：原子、微观、介观和宏观尺度，并严格遵循材料研究中的重点与定义。我们通过开放式图描述和多项选择式图注匹配任务，评估当前最先进的开源和闭源模型。分析发现，由于不同物理尺度具有迥异的视觉特征，模型在各尺度上的表现差异显著；这凸显了当前通用模型的局限性，并指出了实现材料研究中层次化且准确理解的关键方向。CSMBench 已公开发布于：**
https://huggingface.co/datasets/lututu/CSMBench

索引词—**大型多模态模型，材料科学，视觉问答，评估**

I. 引言

材料科学是现代技术进步的主要引擎，支撑着高效储能、航空航天结构完整性以及下一代半导体领域的创新 [1]。这一领域从根本上由物质的层次化特性所定义：材料的宏观性能，是多个物理尺度上结构共同作用的结果。从单个原子的排列，到微米尺度的晶粒分布，再到最终的块体几何形态，这些特征都通过复杂且协同的相互作用而演化。因此，跨尺度推理能力——将微观观察与宏观结果联系起来——是材料研究领域专业知识的显著标志。

随着大型多模态模型不断从通用理解迈向领域专属知识， [2], [3],科学发现已成为评估并拓展其能力的关键前沿领域 [4]。近期，大型多模态模型在多种基准测试中取得了卓越表现，但其在材料这一专业领域的有效性仍未得到充分验证。现有科学基准测试（如 ScienceQA） [5] 评估的是一般科学主题或大学水平推理，且往往依赖教科书和学术资料等二手来源 [6]。因此，这些通用基准测试无法体现现实材料研究的多尺度特性。它们并未评估模型能否跨越不同物理尺度，从埃尺度晶格延伸至厘米级块体材料，因而在细粒度衡量模型跨层级物理尺度区间的科学认知能力方面，仍留下一个根本性挑战。

为弥合这一关键差距，我们推出 CSMBench，即面向大型多模态模型的**跨**尺度**材**料科学**基**准**测**试，旨在通过四个物理尺度区间评估大型多模态模型的层次化理解能力。原子尺度考察模型辨识晶格排列和单个原子缺陷的能力。微观尺度体现模型解读纳米至微米特征（如析出物和位错）的能力。介观尺度体现模型分析微米至毫米结构（包括晶界和织构）的能力。宏观尺度则体现模型对厘米尺度块体材料及最终部件几何形状的理解能力。

我们使用 CSMBench 对最先进的专有和开源大型多模态模型进行基准测试，如表 I 所示。这些模型在我们的框架上表现欠佳，表明 CSMBench 为科学领域大型多模态模型的发展构成了具有挑战性的前沿，也揭示了视觉感知与深入物理推理之间的差距。我们的核心贡献总结如下：

1）我们从最新的2025年同行评审文献中精选出包含1,041幅图的无污染数据集，确保模型在训练截止时间之后的新颖数据上接受评估。2）我们将数据集划分为原子、微观、中观和宏观四种物理尺度，并设计了开放式图像描述和多项选择式图注匹配两项任务，以衡量识别精度和推理深度。3）我们全面评估了十种最先进的大型多模态模型，结果表明，由于视觉特征各不相同，模型在不同物理尺度上的表现存在显著差异。

* 通讯作者。

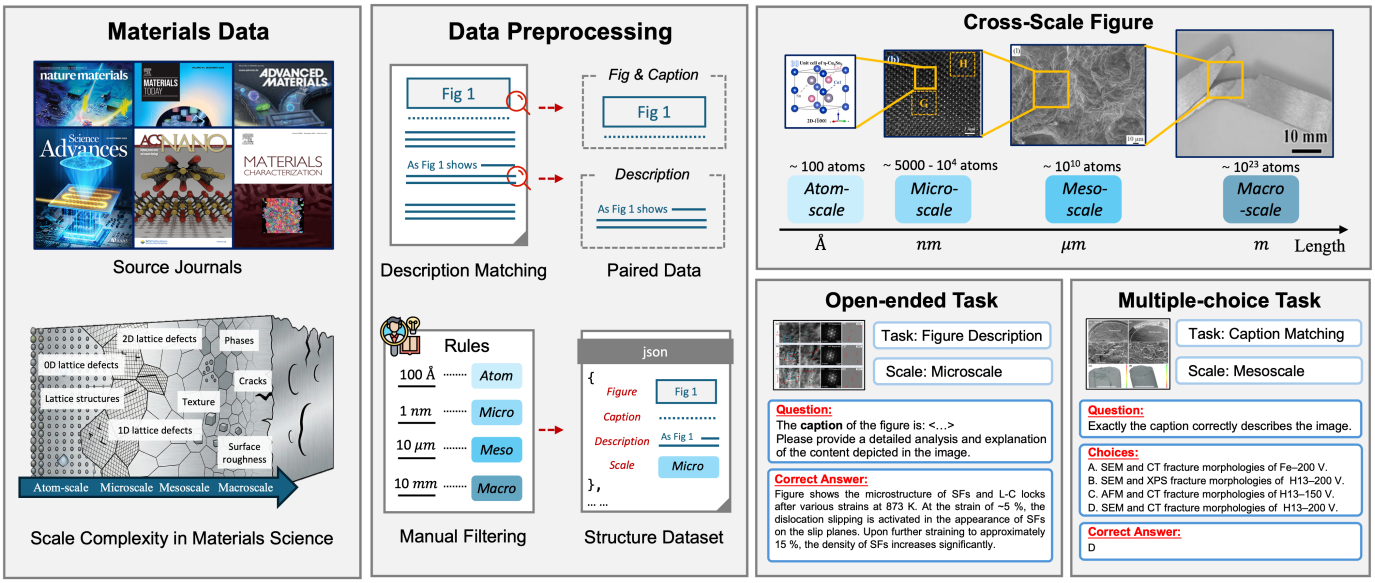


图 1：CSMBench 概览，由三个主要阶段组成：源数据收集、数据处理和任务设计。该流程汇集了权威材料学期刊中涵盖四种物理尺度的图像，范围从原子级到宏观级，并严格遵循材料科学的基本定义。左下方的示意图[1] 突显了不同尺度的内在特征，为我们的尺度分类提供了理论依据。通过结合自动描述匹配与专家人工筛选的混合工作流程，原始数据被整理为高质量的结构化数据集。最后，该框架设立了两项评估任务，以评估跨尺度感知能力：开放式图像描述和多项选择图注匹配。

II. 相关工作

科学领域的大型多模态模型：将领域知识融入大型多模态模型，已成为现代科学发现的基石，推动研究从简单的文本处理转向全面的视觉理解。这一演进在化学和生物医学等领域尤为明显：ProGen 等模型利用专业知识解读复杂序列；在地球科学领域，EarthGPT [8] 则整合多模态卫星数据，用于环境分析。沿着这一发展趋势，材料科学领域的大型多模态模型应用日益增多，涌现出 MatterChat [9]，将晶体结构与文本相结合；以及 HoneyComb [10]，为研究建立智能体工作流。然而，与基于序列的领域不同，材料科学从根本上受制于不同物理尺度上的视觉特征。尽管近期研究已引入多模态能力，但主要挑战在于，如何使这些模型不仅能解读孤立图像，还能理解其中蕴含的层级物理定律。

科学基准：近年来，科学领域对大型多模态模型的评估工作催生了 ScienceQA [5]，等基准，该基准包含 ~21 k 道涵盖各类科学主题的多模态选择题。MathVista [6] 等专项基准则进一步通过数学可视化考验模型。在科学图表理解领域，MMSCI [11] 采用了涵盖 72 个领域的 *NatureCommunications* 图表，但其评估仍侧重于一般的图表解读，而非深入的领域专业知识。同样，MaCBench [12]为材料科学引入了多模态评估，但其内容在很大程度上局限于常识性解读。重要的是，这些通用基准未能体现材料的多尺度特性，也未评估模型跨越不同物理尺度、从原子尺度到厘米级块体材料进行推理的能力，因此学界仍缺乏用于评估尺度感知推理的严谨基准。

III. CSMBENCH

A. 数据收集

为确保基准具有权威性，并能代表跨尺度特征，我们制定了如下严格的数据收集与处理流程。

1）源数据收集：CSMBench数据集源自八本在该领域享有权威声誉的知名材料科学期刊：Nature、Nature Materials、Nature Communications、Science Advances、Advanced Materials、Materials Today 和 Materials Characterization。尤其值得一提的是，Materials Characterization 重点关注材料的结构行为，并广泛采用多种表征技术。数据集收录的文章截至2025年9月，确保数据集不受过时数据的影响，从而为评估当代

[1]该示意图展示了固体中的晶格缺陷类型如何受到成分复杂性的影响。该图改编自 [1] 发表于 2023 年 Nature Computational Science 的材料设计研究。

表 I：四个尺度（原子尺度、微观尺度、介观尺度和宏观尺度）下开放式问答的性能比较。报告的指标包括 BERTScore（F1）、语义文本相似度（STS）和 LLM评审得分（大语言模型）。所有指标均遵循数值越高越好的趋势（↑）。粗体条目表示最佳结果。

模型	原子尺度			微观尺度			介观尺度			宏观尺度			总体		
	Bert	STS	LLM	Bert	STS	LLM	Bert	STS	LLM	Bert	STS	LLM	Bert	STS	LLM
GPT-5.1	0.6101	0.7006	**7.8679**	0.6129	**0.6917**	**8.2672**	0.6613	**0.7011**	**8.0971**	0.5975	0.6613	**7.5217**	0.6205	0.6887	**7.9385**
Gemini-2.5-pro	**0.6268**	**0.7280**	7.0566	**0.6228**	0.6777	7.5854	**0.6232**	0.6949	7.4429	**0.6101**	**0.6599**	6.7935	**0.6207**	**0.6901**	7.2196
Doubao-1.6-vison	0.5837	0.6991	7.2075	0.5851	0.6794	7.6842	0.5829	0.6874	7.5593	0.5755	0.6530	7.0000	0.5818	0.6797	7.3628
Qwen2.5-VL-7B	**0.6273**	0.7169	5.2264	0.6322	0.7026	6.3806	0.6234	0.6975	5.8490	0.6040	0.6533	5.5217	0.6217	0.6926	5.7444
Qwen2.5-VL-32B	0.5827	0.6660	4.5849	0.5890	0.6595	5.1174	0.5870	0.6664	5.1032	0.5720	0.6274	4.7065	0.5827	0.6548	4.8780
Qwen2.5-VL-72B	0.6245	**0.7199**	**6.5849**	**0.6323**	**0.7055**	**6.9838**	**0.6291**	**0.7074**	6.5485	**0.6102**	**0.6613**	**5.8587**	**0.6240**	**0.6985**	**6.4940**
Qwen3-VL-8B	0.6215	0.7183	6.3774	0.6238	0.7028	6.9069	0.6208	0.7040	**6.6256**	0.6037	0.6559	5.8043	0.6175	0.6953	6.4286
InternVL3-8B	0.6108	0.6874	4.7358	0.6122	0.6867	5.3482	0.6082	0.6731	5.0277	0.5967	0.6413	4.8043	0.6070	0.6721	4.9790
InternVL3-38B	0.6125	0.7045	5.7170	0.6100	0.6846	6.2186	0.6082	0.6839	5.9353	0.6009	0.6460	5.6848	0.6079	0.6798	5.8889
InternVL3-78B	0.6142	0.7094	5.6792	0.6134	0.6824	6.3968	0.6084	0.6825	6.0139	0.6009	0.6446	5.7935	0.6092	0.6797	5.9709

大型多模态模型。我们从432篇经同行评审的材料科学文章中，精选了涵盖四种不同物理尺度的1,041幅图。文章。

遵循材料信息学原则，我们根据物理 d 将材料研究尺度划分为四个层级
维度。收集的数据涵盖以下四个尺度：

- **原子尺度：** 在这一 A尺度（10^{-10}m）下，材料研究聚焦于电子结构、缺陷和界面等原子层面的现象，这些现象是理解决定材料性质的原子键合与相互作用的基础。

- **微观尺度：** 在纳米尺度（10^{-9}m）下，研究关注扩散、界面演化和材料退化等现象，这些现象会影响晶粒或颗粒层面的强度和导电性等性质。

- **介观尺度：** 在微米尺度（10^{-6}m）下，研究考察相界、孔隙和裂纹等微观结构特征。该尺度连接原子尺度模型与宏观力学，并影响电化学性质和力学性质。

- **宏观尺度：** 在宏观尺度（厘米至米）下，材料被视为连续介质，重点关注工程应用中的变形、流体流动和热传递等体相行为。

2) 数据预处理： 为了获得每个图的详细说明，用作评估感知能力的真实标注，我们首先使用 MinerU 将原始 PDF 转换为机器可读格式。MinerU 会对图进行分割，将公式转换为 LaTeX，并将文本提取为 Markdown。随后，我们开发了一个正则表达式匹配函数，用于识别正文中的图引用（例如 Fig. n）。由于初始匹配通常会为一个图识别出多个候选段落，我们设计了清洗流程，排除引用多个图或包含外部表格数据的段落。经过这一步筛选后，我们将属于同一张图的零散描述合并为一个完整段落。最后，我们保留长度在 100–300 个词元范围内的描述，以剔除长尾数据。

3) 人工筛选：为进一步确保图像数据符合我们的跨尺度要求，我们会进行专家人工审查，筛除自动提取的、仅包含通用流程图或纯统计图表的图像数据，保留那些能够直接揭示材料形貌、相或成分的图像。完成筛选后，我们会通过人工核验每张图像中的比例尺和上下文元数据，进行尺度标注。经过这一流程，最终得到 1,041 张经过准确分类的图像，用于评估跨尺度推理能力：原子尺度 53 张、微观尺度 247 张、介观尺度 649 张、宏观尺度 92 张。数据分布不均衡源于样本本身的固有分布，而这可能受到表征难度和机制研究重点的影响。

B. 任务设计

我们在 CSMBench 中设计了两项设置各异的任务，以全面评估大型多模态模型解读材料科学图表的能力。

1）开放式任务：在此任务中，我们评估模型的图描述能力：模型以一张图及其说明文字为输入，并需生成一段 100 至 300 词的详细说明。模型需要聚焦于图像中可见的特征，并依据所呈现的证据得出科学结论。通过比较生成内容与从原始论文中提取的高保真真实标注段落之间的相似度，对生成内容进行评估。

这项描述任务专门用于评估跨尺度感知能力。与通用的图表到文本任务不同，解读材料科学图表需要对局部和整体特征都有细致入微的理解。例如，模型必须能够正确识别原子尺度上的原子排列、微观尺度上的界面演化、介观尺度上的微观结构形貌，以及宏观尺度上的整体变形模式。通过要求模型分析这些视觉线索，该任务能够有效检验模型是否具备对材料科学深刻且基于物理的理解。

2）多项选择任务：第二种任务形式为多项选择，旨在评估图注匹配任务：模型需要从四个选项中，为给定图像选出正确的图注，其中包含

表 II：不同模型在四个尺度上的选择题性能比较。我们报告准确率分数。最佳结果以粗体突出显示。

模型	原子尺度。	微观尺度。	中观尺度。	宏观尺度。	总体
GPT-5.1	0.8491	0.9271	0.8829	0.9565	0.9039
Gemini-2.5-pro	0.9434	0.9433	0.9137	0.9348	0.9338
Doubao-1.6-vison	**0.9434**	**0.9636**	**0.9492**	**0.9674**	**0.9559**
Qwen2.5-VL-7B	0.6604	0.7895	0.7458	0.7065	0.7256
Qwen2.5-VL-32B	0.7170	0.7814	0.7935	0.7935	0.7714
Qwen2.5-VL-72B	0.7170	0.8300	0.8136	**0.8587**	0.8048
Qwen3-VL-8B	**0.8626**	**0.9057**	**0.8988**	0.8536	**0.8802**
InternVL3-8B	0.6981	0.7166	0.7057	0.6304	0.6877
InternVL3-38B	0.7736	0.8866	0.8582	0.8478	0.8416
InternVL3-78B	0.7736	0.8421	0.8320	0.8152	0.8157

细粒度的干扰项。为严格测试模型的精确识别能力，我们利用 GPT-4.1 通过有针对性的扰动生成三类具有挑战性的负向干扰项，具体方式如下：

- **表征方法：**我们提示模型将原始技术替换为相似但不正确的方法（例如，将透射电子显微镜替换为扫描电子显微镜，或将X射线衍射替换为X射线光电子能谱）。这用于检验模型能否识别不同实验仪器各自特有的视觉特征。
- **材料成分：**我们要求模型将特定材料名称或化学式替换为合理的替代项（例如，将 ZrB_2 替换为 TiB_2，或将 Al_2O_3 替换为Fe_2O_3）。这项测试评估模型将化学知识文本与视觉上下文联系起来的能力。
- **数值：**我们要求模型扰动温度、浓度、尺寸和时间等关键参数（例如，将 1050°C 改为 950°C，或将 2 h 改为 4 h）。准确识别数值对于精确的科学推理至关重要。

通过要求模型区分这些细微差异，我们评估其高精度识别表征方法、成分和定量数据的能力，而这些都是正确解读材料研究的基础。

IV. 实验

A. 实验设置

1）基准测试候选模型：为在 CSMBench 上建立全面的性能基线，我们评估了多种大型多模态模型（LMMs）。所选模型涵盖了前沿专有模型和具有代表性的开源模型。具体而言，在专有模型方面，我们纳入了 GPT-5.1 [13], Grok-4 [14],和 Gemini-2.5-Pro [15]等最先进的候选模型。对于开源大语言模型，我们重点关注以强大多模态能力著称的模型，具体包括 Qwen-2.5-VL-{7, 32, 72}B [16], Qwen-3-VL-8B [17]和 InternVL3-{8, 38, 78}B [18]。

2）评估指标：对于多项选择问答任务，我们以准确率作为主要指标。对于开放式问答任务，我们报告三项指标。对于**BERTScore** [19], ，我们采用 F1分数，利用上下文嵌入评估语义相似度

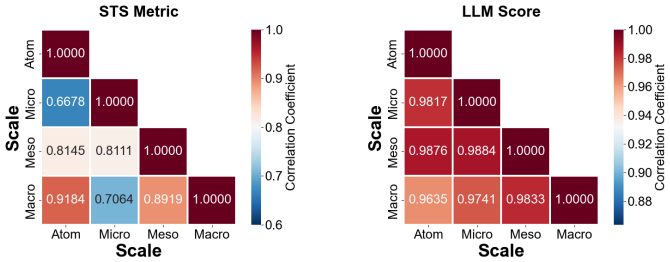


图 2：使用 STS 指标和大语言模型评分，分别表示原子、微观、中观和宏观尺度之间的皮尔逊相关系数。

。**语义文本相似度（STS）**用于从整体解读层面进行高层次的概念评估。我们使用句子转换器模型 [20],将模型输出和真实标注投影到统一的 384 维向量空间，并计算二者的余弦相似度，以此计算 STS。对于**LLM评审得分**，我们使用 GPT-4o-2024-11-20 [13] 作为评审模型，根据先前工作 [21], 中开发的实现方法和提示，对照模型预测对答案的正确性进行语义验证，得分范围为 1 到 10。将 LLM 得分与 STS 结合使用，有助于排除家族偏差，即 GPT-4o 可能偏向 GPT 系列模型。只有在各项得分一致时，结果才被视为可靠。

B. 主要结果

观察 1. 专有模型在科学推理和判别准确率方面均表现更佳。在所有评估的候选模型中，专有模型均展现出系统性优势。如表 I 所示，GPT-5.1 在开放式描述任务中的 LLM 评审得分最高，达到 8.0749，表明其能够生成基于物理且科学准确的描述。在表 II 所示的多项选择匹配任务中，Doubao-1.6-vision 表现最佳，整体准确率接近完美，达到 95.59%；Gemini-2.5-pro 紧随其后，达到 93.38%。专有模型与开源模型之间仍存在显著的性能差距。在表 I 所示的开放式任务中，表现最强的开源代表模型 Qwen2.5-VL-72B 的整体大语言模型评分为 6.4940，比 GPT-5.1 低约 1.82 分。在细粒度匹配任务中，这一差距同样明显：即使是能力最强的开源模型，整体准确率也往往比领先的专有模型低近 8%。这些结果表明，尽管开源模型不断进步，专有系统在材料科学所需的复杂多模态推理方面仍具备更强的能力。

观察 2。大型多模态模型呈现出依赖尺度的性能波动。对四个物理尺度的结果进行分析发现，模型在微观尺度和介观尺度上的整体表现优于原子尺度和宏观尺度。GPT-5.1 在微观尺度上取得了 8.2672 的最高 LLM评审得分，但在宏观尺度上显著下降至 7.5217。

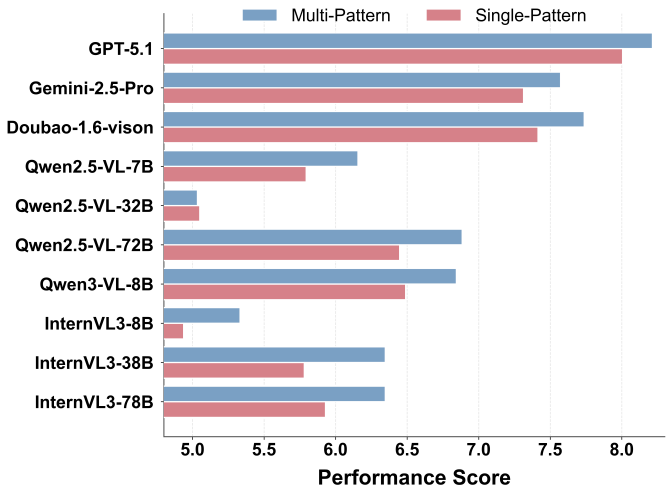


图3：纯形态图像与图表化混合多模式图像上的性能比较。

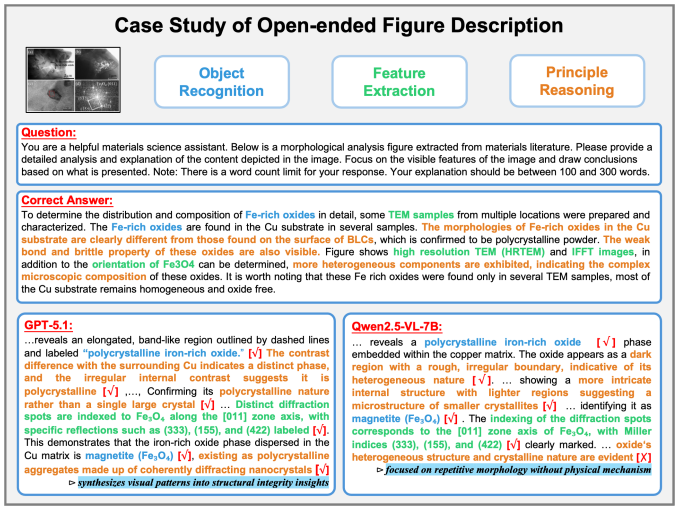


图4：GPT-5.1 和 Qwen2.5-VL-7B 在物体识别、特征提取和原理推理各维度上的比较案例研究。

开源模型也呈现出相同趋势，例如 Qwen3-VL-8B 的得分从微观尺度的 6.9069 下降至宏观尺度的 5.8043。这种性能差异主要归因于各尺度视觉数据的性质。微观尺度和介观尺度主要由实验表征中的标准化图像构成，包括扫描电子显微镜、计算机断层扫描和X射线光电子能谱图像。这些图像具有相对一致的视觉模式，模型能够更容易地进行解读。相比之下，原子尺度和宏观尺度的数据集包含更高比例的多样化示意图和异质视觉表征，增加了解读任务的复杂性。此外，图 2 所示的相关性分析进一步证实了这种尺度依赖性。尽管 LLM评审指标的得分表现出很高的跨尺度一致性，超过 0.97，这可能归因于模型内部逻辑的稳定性，但 STS 指标显示出显著更高的方差，并下降至 0.6678。这表明，语义和术语的精确性对具体视觉语境以及材料科学中依赖尺度的语言具有高度敏感性。

观察3。推理能力优先于材料科学解读中的缩放定律。我们可以从图 II 中观察到，在 Qwen2.5-VL 系列中，匹配准确率从 7B 模型的 72.56% 提升至 72B 版本的 80.48%。然而，在复杂任务中，这种缩放趋势并不一致。如图 I 所示，32B 模型在开放式图像描述任务中的得分为 4.8780，未能超过 7B 模型的 5.7444。同样，InternVL3-38B 在多项选择题图注匹配任务中的准确率达到 84.16%，超过了更大的 78B 版本的 81.57%。这些发现表明，单纯增加模型参数并不能确保模型在科学语境中实现更强的语义整合。此外，Qwen3-VL-8B 的表现进一步凸显了架构演进的价值。

该模型得分为 88.02%，优于规模大得多的 Qwen2.5-VL-72B（得分为 80.48%）。通过引入内部思维过程，Qwen3-VL-8B 能够系统地分析材料科学图像的层级特征，从而更精细地对齐视觉证据与物理原理。

观察4。开放式描述任务与多项选择题图注匹配任务之间的差异。直接比较这两项任务可以发现，有些模型擅长选出正确选项，却无法阐述图中蕴含的领域信息。最值得注意的是，Doubao-1.6-vision 在图注匹配任务中取得了最高的总体准确率 95.59%，展现出出色的细粒度图注辨别能力。然而，在开放式图像描述任务中，它的表现不及 GPT-5.1，后者的大语言模型评审评分更高，达到 7.9385。同样，在开源模型中，InternVL3-38B 的辨别能力较强，匹配准确率为 84.16%，但其科学叙述表现明显不及 Qwen3-VL-8B，后者的大语言模型评分更高，达到 6.4286。这些结果表明，许多模型依赖视觉启发式来解答选择题，却缺乏进行高质量科学描述所必需的跨尺度结构映射和物理性质推导能力。

观察5.异质视觉模式可增强解读表现。材料科学中的视觉数据主要包括用于呈现物理结构的形貌表征，以及用于展示与性质相关的信息或统计分析的辅助统计图。为评估视觉多样性的影响，我们比较了模型在两种不同数据类别上的表现：单一形貌模式，以及融合两种图表类型的多模式。我们根据大语言模型评分指标计算了

各类别的得分，如图3所示。大多数模型，包括 GPT-5.1、Gemini-2.5-Pro 和 Qwen2.5-VL-72B，在混合视觉配置上的得分更高。这一提升表明，辅助统计图提供了关键的上下文锚点，通过补充定量数据，帮助消歧材料的复杂结构。

C. 案例研究

为考察定性推理能力，我们从三个层面比较 GPT-5.1 和 Qwen-3-VL-8B 的能力，具体包括：以识别各相为例的物体识别、解析技术特征等特征提取，以及推断物理机制等原理推理。分析结果如图4所示。两种模型在物体识别和特征提取方面均表现出色，能够正确识别形态为 Fe_3O_4 的磁铁矿析出物，并提取米勒指数等晶体学特征。这些结果表明，两者在低层次视觉感知方面表现相当。然而，在原理推理方面仍存在显著差距。真实标注指出这些氧化物具有脆性且结构键合较弱，而开源模型仍专注于重复描述形貌。相较之下，专有模型表现出色，能够将视觉数据综合为关于结构完整性的更高层次阐释。这些结果证实，尽管专有模型与开源模型在感知能力上的差距正在缩小，但推理能力方面的差距依然存在，因为专有模型更擅长将视觉模式转化为抽象物理原理。

V. 结论

在本研究中，我们提出了 CSMBench，旨在评估模型在材料研究中跨越四种物理尺度的层次化理解能力。我们从截至2025年9月的同行评审文献中整理了1041幅高质量图表，以对科学感知能力进行细粒度评估。我们发现，由于不同物理尺度具有 distinct visual characteristics，模型性能在各物理尺度之间存在显著差异。我们进一步发现了一个关键缺口：较高的识别准确率并不能转化为深入的物理推理能力。我们认为，CSMBench 是一个具有挑战性的前沿基准，有助于推动面向科学发现的基于物理的模型发展。

VI. 致谢

本工作得到新一代人工智能国家科技重大专项（2025ZD0124104）的支持，并与上海人工智能实验室开展合作。

参考文献

[1] Dierk Raabe、Jaber Rezaei Mianroodi 和 J¨org Neugebauer，《利用物理信息人工智能加速成分复杂材料的设计》，*Nature Computational Science*，第 3 卷，第 3 期，第 198–209 页，2023 年。

[2] Zicheng Zhang、Junying Wang、Yijin Guo、Farong Wen、Zijian Chen、Hanqing Wang、Wenzhe Li、Lu Sun、Yingjie Zhou、Jianbo Zhang 等，《Aibench：面向 45°定律下的可信评估》，*Displays*，第 103255 页，2025年。[3] Zijian Chen、Yuan Tian、Yuze Sun、Wei Sun、Zicheng Zhang、Weisi Lin、Guangtao Zhai 和 Wenjun Zhang，《大型多模态模型的刚好可觉差》，*arXiv* 预印本 *arXiv:2507.00490*，2025 年。[4] Kevin Maik Jablonka、Philippe Schwaller、Andres Ortega-Guerrero 和 Berend Smit，《利用大型语言模型进行预测化学研究》，*Nature Machine Intelligence*，第 6 卷，第 169 页，2024 年，第 2 期，第 161 页。[5] Pan Lu、Swaroop Mishra、Tanglin Xia、Liang Qiu、Kai-Wei Chang、Song-Chun Zhu、Oyvind Tafjord、Peter Clark 和 Ashwin Kalyan，《学会解释：通过思维链进行多模态推理，以回答科学问题》，神经信息处理系统进展，第 35 卷，第 2507–2521 页，2022 年。[6] Pan Lu、Hritik Bansal、Tony Xia、Jiacheng Liu、Chunyuan Li、Hannaneh Hajishirzi、Hao Cheng、Kai-Wei Chang、Michel Galley 和 Jianfeng Gao，《MathVista：在视觉上下文中评估基础模型的数学推理能力》，载于第十二届国际学习表征会议。[7] Ali Madani、Bryan McCann、Nikhil Naik、Nitish Shirish Keskar、Nam-rata Anand、Raphael R Eguchi、Po-Ssu Huang 和 Richard Socher，《ProGen：用于蛋白质生成的语言建模》，*arXiv* 预印本 *arXiv:2004.03497*，2020 年。[8] Wei Zhang、Miaoxin Cai、Tong Zhang、Yin Zhuang 和 Xuerui Mao，《EarthGPT：面向遥感领域多传感器图像理解的通用多模态大语言模型》，*IEEE* 地球科学与遥感汇刊，第 62 卷，第 1–20 页，2024 年。[9] Yingheng Tang、Wenbin Xu、Jie Cao、Weilu Gao、Steve Farrell、Benjamin Erichson、Michael W Mahoney、Andy Nonaka 和 Zhi Yao，《MatterChat：面向材料科学的多模态大语言模型》，*arXiv* 预印本 *arXiv:2502.13107*，2025年。[10] Huan Zhang、Yu Song、Ziyu Hou、Santiago Miret 和 Bang Liu，《HoneyComb：用于材料科学的灵活基于大语言模型的智能体系统》，载于计算语言学协会论文集：*EMNLP 2024* 研究成果，2024 年，第 3369–3382 页。[11] Zekun Li、Xianjun Yang、Kyuri Choi、Wanrong Zhu、Ryan Hsieh、HyeonJung Kim、Jin Hyuk Lim、Sungyoung Ji、Byungju Lee、Xifeng Yan 等，《MMSCI：面向博士级科学理解的多模态多学科数据集》，载于用于加速材料设计的人工智能——维也纳 *2024*，2024 年。[12] Nawaf Alampara、Indrajeet Mandal、Pranav Khetarpal、Hargun Singh Grover、Mara Schilling-Wilhelmi、NM Anoop Krishnan 和 Kevin Maik Jablonka，《MaCBench：多模态化学与材料科学基准》，载于第 *38* 届神经信息处理系统会议论文集（*NeurIPS 2024*），2024 年。[13] OpenAI，《GPT-5：生成式预训练变换器 5》，2025 年，访问日期：2025年12月29日。[14] xAI团队，《Grok 4 快速模型卡片》，*xAI*技术报告，2025 年 9 月。[15] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等，《Gemini 2.5：凭借先进推理、多模态性、长上下文和新一代智能体能力，推动前沿发展》，*arXiv* 预印本 *arXiv:2507.06261*，2025 年。[16] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等，《Qwen2.5-VL 技术报告》，*arXiv* 预印本 *arXiv:2502.13923*，2025 年。[17] Qwen 团队，《Qwen3 技术报告》，2025 年。[18] 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先进训练与测试时方案》，*arXiv* 预印本 *arXiv:2504.10479*，2025 年。[19] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi，《BERTScore：使用 BERT 评估文本生成》，载于国际学习表征会议。[20] Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou，《MiniLM：用于任务无关预训练变换器压缩的深度自注意力蒸馏》，神经信息处理系统进展，第 33 卷，第 5776–5788 页，2020 年。[21] Yuhao Zhou、Yiheng Wang、Xuming He、Ao Shen、Ruoyao Xiao、Zhiwei Li、Qiantai Feng、Zijie Guo、Yuejin Yang、Hao Wu 等，《科学家的首次考试：通过感知、理解和推理探测多模态大语言模型的认知能力》，*arXiv* 预印本 *arXiv:2506.10521*，2025 年。