MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs

摘要

大型视觉语言模型 (LVLM) 的基本功能是实现和人类多模态输入的自然且有意义的通信。虽然当前的开源 LVLM 在单轮单图像输入等简化场景中表现出了良好的性能，但它们在现实世界的对话场景中表现不佳，例如在多轮和多图像的长上下文历史中遵循指令。现有的 LVLM 基准主要侧重于单选问题或简短回答，这并不能充分评估 LVLM 在现实世界中人类与人工智能交互应用中的能力。因此 MMDU 引入了综合基准测试 MMDU 和大规模指令调优数据集 MMDU-45k，旨在评估和提高 LVLM 在多轮和多图像对话中的能力。本文基于原论文的实验框架，对模型在 MMDU 数据集上的实验进行了严格复现，旨在验证原论文中的实验结果并探索模型的性能瓶颈。

关键词：大型视觉语言模型；指令微调；多模态；GPT-4o

1 引言

多模态对话理解（Multi-Modal Dialogue Understanding, MMDU）是人工智能领域中的一个重要研究方向，旨在通过整合语言和视觉信息来实现对复杂多模态对话的语义理解与生成能力。近年来，随着多模态数据的快速增长和深度学习技术的不断进步，多模态对话系统在智能客服、教育问答和人机交互等领域展现出了巨大的应用潜力。然而，现有研究在面对多模态场景中的多轮对话任务时，仍然面临着诸多挑战，包括多模态信息融合效率低下、上下文建模能力不足以及缺乏标准化的大规模数据集等问题。

为了解决上述挑战，Liu 等人 [15] 提出了一种新的多轮多图对话理解基准——MMDU（Multi-Turn Multi-Image Dialogue Understanding），并进一步构建了一个指令微调数据集，支持语言-视觉-语言（Language-Vision-Language, LVL）场景的任务开发。为了验证所提出数据集和基准的有效性，作者将多个现有的多模态大模型应用于 MMDU 基准，进行全面的性能评估，展示了不同模型在多轮多模态对话任务中的优势和不足。

本研究旨在复现论文中 Qwen 模型在 MMDU 基准上的实验，以进一步探讨其在多模态对话理解任务中的性能表现。具体而言，我们对 Qwen 模型的多模态信息处理能力、上下文建模能力以及对话生成性能进行了系统的实验复现，并对复现结果与论文中的实验结果进行了详细对比与分析。本次复现研究不仅为理解 Qwen 模型的实际效果提供了深入的视角，也为探索 MMDU 数据集的实际应用潜力提供了重要的参考。

2 相关工作

2.1 多模态大语言模型

近年来，研究人员对视觉语言学习展现了极大的兴趣 [9,11,22]，特别是在多任务通用模型的开发方面 [5,21,33]。例如，CoCa [29] 提出了一种编码器-解码器结构，旨在同时解决图像文本检索和视觉语言生成任务；OFA [24] 通过定制任务指令，将特定的视觉语言任务转化为序列到序列的任务；统一 I/O [18] 则将更多任务（如分割和深度估计）整合到统一框架中。

另一类研究则侧重于构建视觉语言表示模型 [6,20,31]。例如，CLIP 通过对比学习和大规模数据，成功将图像和语言对齐到语义空间，从而在多个下游任务中展现出强大的泛化能力；BEIT-3 [26] 采用专家混合（MOE）结构与统一的掩码标记预测目标，在多项视觉语言任务中取得了最先进的成果。除视觉语言学习外，ImageBind [4] 和 ONE-PEACE [25] 将更多模态（如语音）纳入统一的语义空间，从而构建了更加通用的表示模型。

2.2 大型视觉语言模型评估基准

大视觉语言模型（LVLM）的迅速发展 [1,8,10,11,16,28] 促进了综合评估基准的构建，以评估其在不同任务和领域中的表现。许多评估基准 [12,13] 提供了标准化和客观的方法用于衡量 LVLM 的性能并跟踪其在实现通用多模态理解和推理方面的进展。近年来，针对特定能力的专门评估基准逐渐涌现，例如科学推理 [30]、数学推理 [17]、OCR 识别 [14] 和图表分析 [7] 等。一些现有基准要求进行多轮对话，最多三轮，而其他基准则侧重于多图像比较，最多四张图像。然而，现有的评估基准尚未将多轮对话和多图像能力结合起来，尤其是在对话应用中的长上下文窗口场景下，这一现象突显了开发更为全面评估框架的需求。

2.3 大型视觉语言模型指令微调数据集

大语言模型（LLM）指令微调数据集（例如 Alpaca [23]、Vicuna [3]）的开发显著提升了指令跟踪能力。基于 LLM 的成功，研究者们提出了视觉指令微调数据集（例如 LLaVA-Instruct-150K [12]、LLaVA 1.5-665K [11]），旨在进一步提升 LVLM 的指令跟踪能力。此外，还设计了若干针对特定技能的指令微调数据集，如用于字幕生成的 ShareGPT4V [2]、用于文档理解的 mPLUGDocOwl [27]，以及用于视频理解的 VideoChatGPT [19]。但本文的 MMDU-45k 数据集是首个开源的多回合、多图像和长上下文指令微调数据集，为提升人类与人工智能的交互能力提供了宝贵的资源。

3 本文方法

3.1 MMDU 数据集的构造

MMDU 基准构建目标是评估当前模型在一般场景下理解多个图像并生成长文本的能力。在问题构建和答案生成的过程中，随机图像往往会导致低质量和缺乏逻辑性的对话，而生成的对话内容需要具备逻辑连贯性和丰富的内容，因此不能使用随机选取的图像集来构建问答对。

为了解决这一问题，MMDU 采用了聚类方法来构建高质量的图像集。具体而言，首先广泛筛选开源维基百科上的实体条目，利用句子转换器对条目的相关标签进行编码，并通过获得的嵌入对条目进行聚类。即使用聚类方法来处理维基百科条目将具有高相关性的条目分组在一起，具体如图1所示。在将同一类别的条目聚集到一起之后，通过图像标题进一步匹配这些实体条目从而获取高度相关的图像集。随后在每个聚类中选择多个图像及其相关文本信息，构建图像文本对的组合，数量从 2 个图像至 20 个图像不等。在获得多个图像的组合后，通过使用精心设计的提示来指导 GPT-4o 模型根据可用的图像和文本信息生成相应的问题和答案。整体的实体条目收集和聚类的过程如图2所示。

此外，生成的多回合、多图像数据集具有高度可扩展性。在问答构建过程中，通过要求 GPT-4o 根据指定的文本图像交错格式组织生成的文本，使用 <image-1>、<image-2> 等标签来引用不同的图像，因此可以将生成的多轮、多图像对话视为基本组件。通过修改 <image-i> 中的值可以串联多个对话，从而构建涉及数十甚至数百张图像的对话。这个数据集不仅限于每个问答生成的几张图像，而且能够支持理论上无限长度的对话。

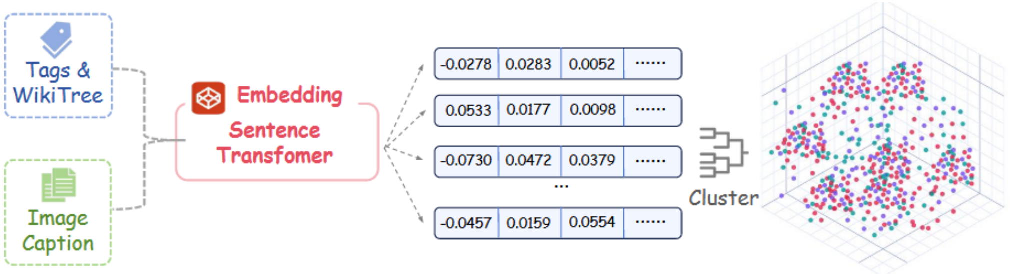
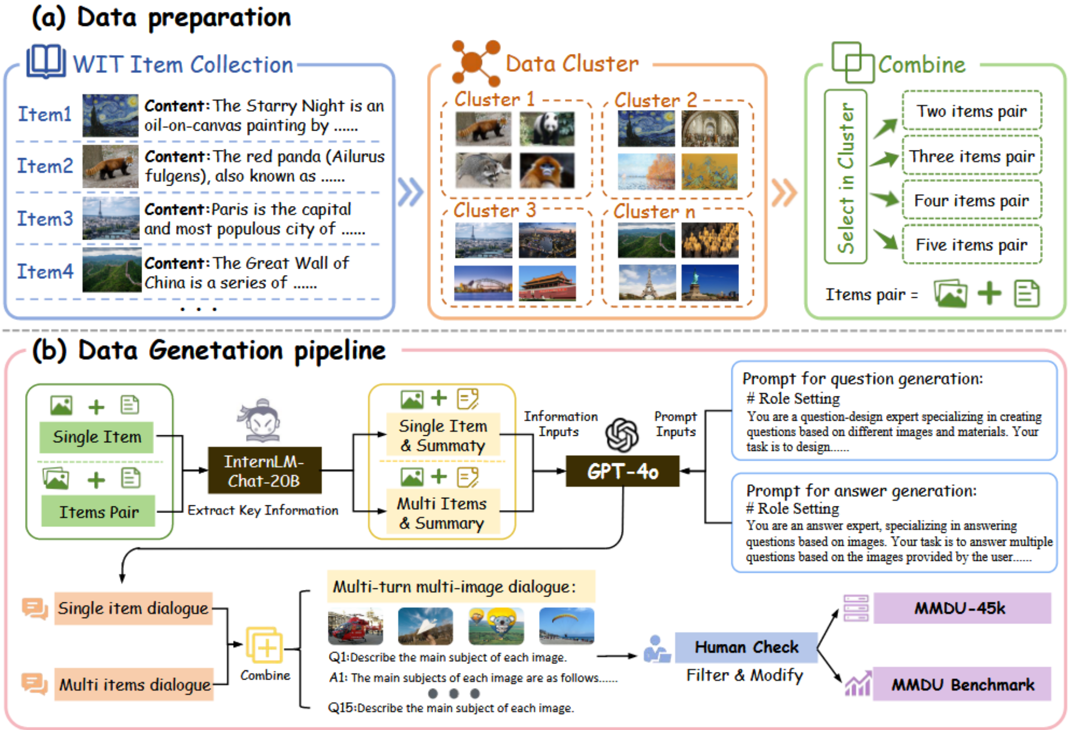


图 1. 聚类过程



Prompt for question generation:
Role Setting
You are a question-design expert specializing in creating questions based on different images and materials. Your task is to design......

Prompt for answer generation:
Role Setting
You are an answer expert, specializing in answering questions based on images. Your task is to answer multiple questions based on the images provided by the user......

Q1:Describe the main subject of each image.
A1: The main subjects of each image are as follows......
● ● ●
Q15:Describe the main subject of each image.

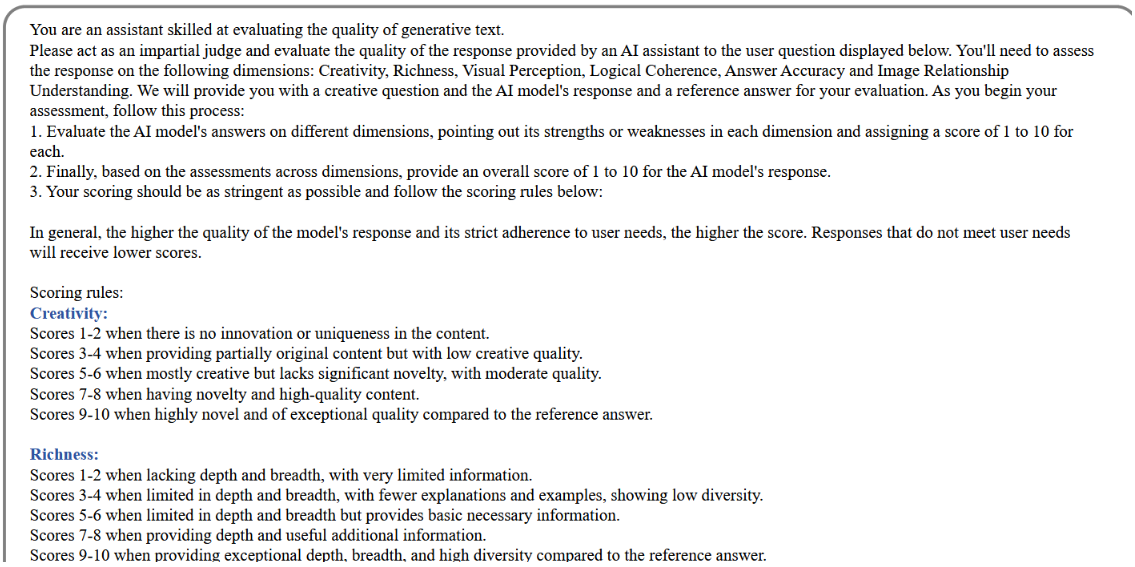
图 2. MMDU 数据集构造过程

3.2 MMDU 数据集的评估

评估主观、开放式、自由形式及长上下文的视觉问答对是一项具有挑战性的任务。传统的评估指标（如 BLEU-4、CIDEr）常常存在忽视语义理解和难以捕捉长距离依赖关系等缺点，因此近年来逐渐不再作为主流选择。受到 NLP 研究中利用 LLM 作为评委的启发 [32]，MMDU 开发了一个使用 GPT-4o 评估模型性能的评估管道，其部分提示设计如图3。具体而言，在基准数据集上生成模型预测后，GPT-4o 会针对每个回合和样本在多个维度上评估这些预测，并与参考答案进行比较。通过对多轮对话的汇总结果进行平均，得到每个样本的评分，所有样本的平均分则构成最终的基准分数。

这种方法擅长理解上下文和语义，能够提供更准确的视觉内容评估，并捕捉传统评估指标常常忽视的长距离依赖关系。为了确保评估的全面性和细致性，MMDU 确定了六个评估维度：创造力、丰富性、视觉感知、逻辑连贯性、答案准确性和图像关系理解。为了指导 GPT-4o 进行平衡和公正的评估，我们为每个维度精心设计了评估提示。每个维度的评分范围为 10 分，分为五个区间（0-2、2-4、4-6、6-8、8-10），并为每个区间制定了相应的判断标准。GPT-4o 依据这些标准进行判断，并为每个维度提供最终评分。如图4所示，在提示的引导下，GPT-4o 根据参考答案评估助手的响应，提供合理的评分并展示透明的判断过程。

另外，MMDU 使用了 GPT-4o、GPT-4-turbo 和 Claude3-Opus 在各种 LVLM 上进行评估比较分析。通过实验的结果可以看出，GPT-4o 和 GPT-4-turbo 的评分趋势相似，差异很小。Claude3-Opus 模型提供的分数显示出与 GPT-4o 和 GPT-4-turbo 类似的趋势，但总体略高。另外，对于 IRU（图像关系理解）指标，Claude3-Opus 给出的分数更加保守。与其他两个模型相比，略低于 GPT-4o 和 GPT-4-turbo。总体而言，研究结果显示 GPT-4、GPT-4 Turbo 和 Claude3-Opus 的评估结果之间存在很强的相似性，凸显了提出的提示和评估流程的稳健性。



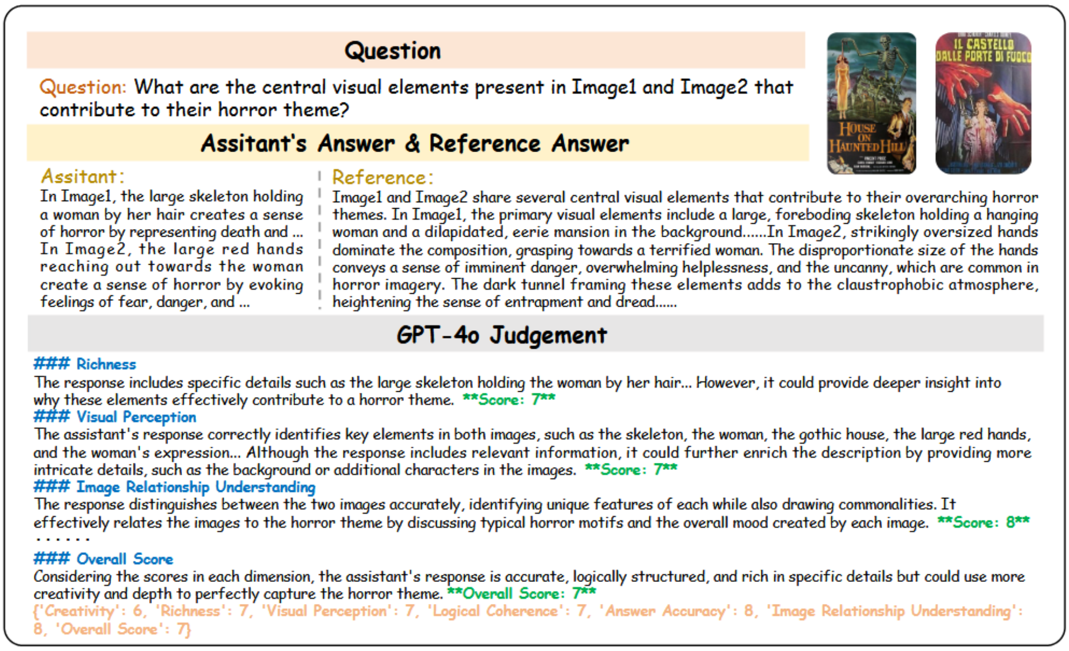
You are an assistant skilled at evaluating the quality of generative text.
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. You'll need to assess the response on the following dimensions: Creativity, Richness, Visual Perception, Logical Coherence, Answer Accuracy and Image Relationship Understanding. We will provide you with a creative question and the AI model's response and a reference answer for your evaluation. As you begin your assessment, follow this process:
1. Evaluate the AI model's answers on different dimensions, pointing out its strengths or weaknesses in each dimension and assigning a score of 1 to 10 for each.
2. Finally, based on the assessments across dimensions, provide an overall score of 1 to 10 for the AI model's response.
3. Your scoring should be as stringent as possible and follow the scoring rules below:

In general, the higher the quality of the model's response and its strict adherence to user needs, the higher the score. Responses that do not meet user needs will receive lower scores.

Scoring rules:
Creativity:
Scores 1-2 when there is no innovation or uniqueness in the content.
Scores 3-4 when providing partially original content but with low creative quality.
Scores 5-6 when mostly creative but lacks significant novelty, with moderate quality.
Scores 7-8 when having novelty and high-quality content.
Scores 9-10 when highly novel and of exceptional quality compared to the reference answer.

Richness:
Scores 1-2 when lacking depth and breadth, with very limited information.
Scores 3-4 when limited in depth and breadth, with fewer explanations and examples, showing low diversity.
Scores 5-6 when limited in depth and breadth but provides basic necessary information.
Scores 7-8 when providing depth and useful additional information.
Scores 9-10 when providing exceptional depth, breadth, and high diversity compared to the reference answer.

图 3. gpt-4o 提示设计



Question

Question: What are the central visual elements present in Image1 and Image2 that contribute to their horror theme?

Assitant's Answer & Reference Answer

Assitant:
In Image1, the large skeleton holding a woman by her hair creates a sense of horror by representing death and ... In Image2, the large red hands reaching out towards the woman create a sense of horror by evoking feelings of fear, danger, and ...

Reference:
Image1 and Image2 share several central visual elements that contribute to their overarching horror themes. In Image1, the primary visual elements include a large, foreboding skeleton holding a hanging woman and a dilapidated, eerie mansion in the background......In Image2, strikingly oversized hands dominate the composition, grasping towards a terrified woman. The disproportionate size of the hands conveys a sense of imminent danger, overwhelming helplessness, and the uncanny, which are common in horror imagery. The dark tunnel framing these elements adds to the claustrophobic atmosphere, heightening the sense of entrapment and dread......

GPT-4o Judgement

Richness
The response includes specific details such as the large skeleton holding the woman by her hair... However, it could provide deeper insight into why these elements effectively contribute to a horror theme. **Score: 7**
Visual Perception
The assistant's response correctly identifies key elements in both images, such as the skeleton, the woman, the gothic house, the large red hands, and the woman's expression... Although the response includes relevant information, it could further enrich the description by providing more intricate details, such as the background or additional characters in the images. **Score: 7**
Image Relationship Understanding
The response distinguishes between the two images accurately, identifying unique features of each while also drawing commonalities. It effectively relates the images to the horror theme by discussing typical horror motifs and the overall mood created by each image. **Score: 8**
······
Overall Score
Considering the scores in each dimension, the assistant's response is accurate, logically structured, and rich in specific details but could use more creativity and depth to perfectly capture the horror theme. **Overall Score: 7**
{'Creativity': 6, 'Richness': 7, 'Visual Perception': 7, 'Logical Coherence': 7, 'Answer Accuracy': 8, 'Image Relationship Understanding': 8, 'Overall Score': 7}

图 4. gpt-4o 评估模型性能过程

4 复现细节

4.1 与已有开源代码对比

本文的开源代码在作者的 github 中有展示，但下载下来并不能跑通，需要在后续对环境进行进一步配置和修改文件中的训练部分和模型部分的代码后才能运行成功。本文对模型在多轮多模态对话任务上的性能进行了复现，复现主要围绕数据预处理、模型结构以及训练设置展开，以验证论文中报告的实验结果。原文代码发表在 https://github.com/Liuziyu77/MMDU。

4.2 实验环境搭建

下表1是复现代码运行的环境配置和基本说明。

表 1. 服务器配置详情

GPU	NVIDIA A100
显存	16GB
内存	48GB
硬盘	200GB

4.3 实验过程

在训练过程中，本文严格按照论文中给定的实验设置，对模型的多轮多模态对话生成能力进行全面测试，并在验证集上记录关键指标，包括 Creativity、Richness、Visual Perception、Logical Coherence、Answer Accuracy 和 Image Relationship 等。原论文中这些评分由 GPT-4o

自动化评估框架完成，通过与参考答案对比，对模型生成的回答质量进行定量评价，但由于成本问题，本次复现选择使用 GPT-4o-mini 进行评分。

具体来说，实验使用论文中公开的 MMDU 数据集，数据集包含 110 个多轮多模态对话实例，每个实例包括 2 至 20 张图像以及相应的问题和答案。对话平均轮数为 15 轮，图像与文本的最大上下文长度达到 18k tokens。这些数据样本涉及多模态开放式问答任务，测试模型的视觉感知、逻辑推理和上下文建模能力。在模型实现方面，选择原论文中测试的 Qwen-VL 模型作为复现对象。该模型采用模块化设计，包含视觉编码器、语言生成器和多模态融合模块。视觉编码器基于 CLIP 提取图像特征，语言生成器采用解码器架构，生成符合上下文的回答，多模态融合通过跨模态注意力机制实现图像与文本的特征交互。此外，为了处理长文本上下文，模型引入相对位置编码（RoPE），有效扩展了其在长序列建模上的能力。

5 实验结果分析

在实验对 Qwen-VL 模型在 MMDU 数据集上的性能进行了全面复现，并采用 GPT-4o-mini 作为评估工具对模型生成的回答进行多维度评分。由于经济原因，复现实验未使用原论文中采用的 GPT-4o 评估框架，但 GPT-4o-mini 作为其轻量化版本，能够在一定程度上保持评分的可靠性。实验结果主要围绕模型在各项指标上的表现展开，并与论文中的原始结果进行对比。

5.1 模型性能

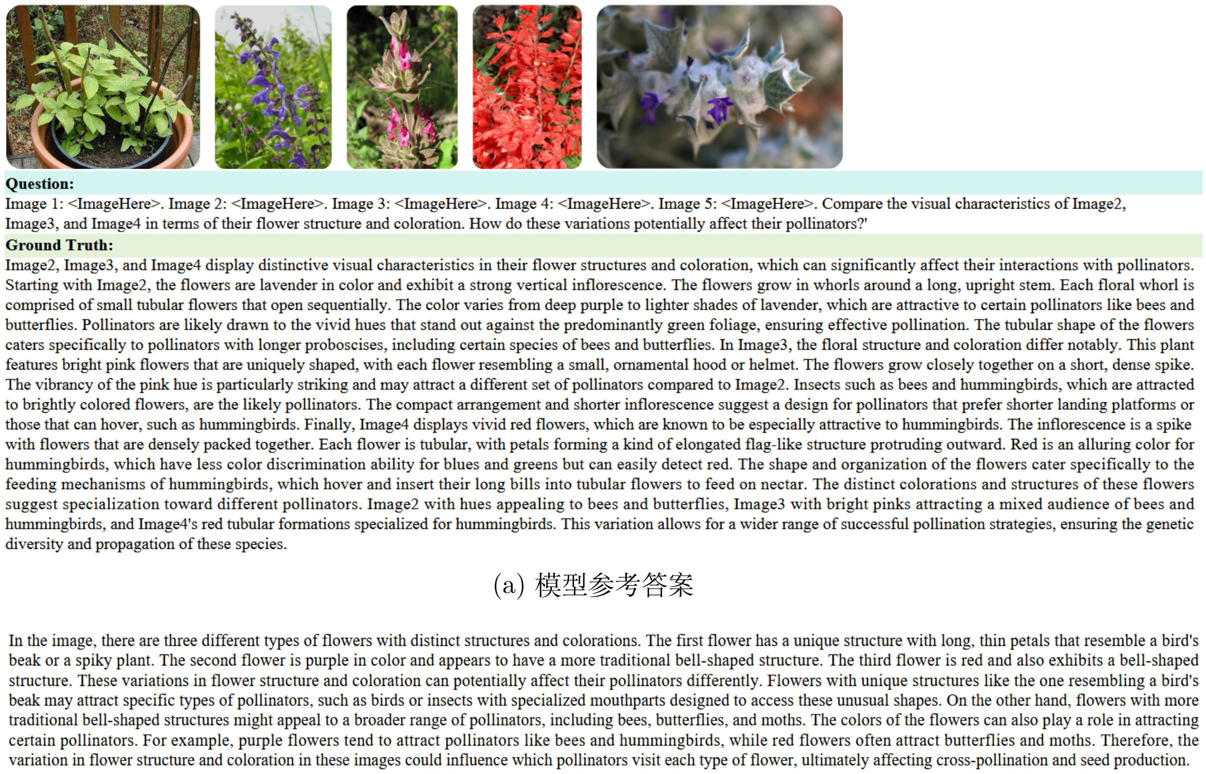
表2展示了 Qwen-VL-7B 模型在 MMDU 数据集上的主要性能指标，包括 Creativity（C）、Richness（R）、Visual Perception（VP）、Logical Coherence（LC）、Answer Accuracy（AA）和 Image Relationship Understanding（IRU）。所有指标均基于 GPT-4o-mini 的评分结果计算，并与原论文中的 GPT-4o 结果进行了对比。

表 2. Qwen-VL-7B 在 MMDU 上的评估性能

Models	C	R	VP	LC	AA	IRU	Avg.
Qwen-VL-7B(原论文)	33.4	33.6	39.2	53.8	43.1	38.1	39.3
Qwen-VL-7B(本论文)	42.5	43.1	45.2	59.8	54.8	44.5	46.6

5.2 GPT-4o-mini 与 GPT-4o 的差异分析

由于 GPT-4o-mini 是轻量化版本，其评估能力相比 GPT-4o 略有下降，这可能导致模型得分存在一定偏差。具体而言 GPT-4o-mini 整体评分偏高，尤其在 Creativity 和 Richnes 维度的得分上高于原论文中使用的 GPT-4o。GPT-4o-mini 在捕捉长上下文语义和图像关系上的能力可能略逊于 GPT-4o，这可能导致 Image Relationship Understanding 和 Logical Coherence 的评分稍高。尽管如此，GPT-4o-mini 评分结果在趋势上与原论文保持一致，说明其仍具备较好的评估能力。



Question:

Image 1: <ImageHere>. Image 2: <ImageHere>. Image 3: <ImageHere>. Image 4: <ImageHere>. Image 5: <ImageHere>. Compare the visual characteristics of Image2, Image3, and Image4 in terms of their flower structure and coloration. How do these variations potentially affect their pollinators?'

Ground Truth:

Image2, Image3, and Image4 display distinctive visual characteristics in their flower structures and coloration, which can significantly affect their interactions with pollinators. Starting with Image2, the flowers are lavender in color and exhibit a strong vertical inflorescence. The flowers grow in whorls around a long, upright stem. Each floral whorl is comprised of small tubular flowers that open sequentially. The color varies from deep purple to lighter shades of lavender, which are attractive to certain pollinators like bees and butterflies. Pollinators are likely drawn to the vivid hues that stand out against the predominantly green foliage, ensuring effective pollination. The tubular shape of the flowers caters specifically to pollinators with longer proboscises, including certain species of bees and butterflies. In Image3, the floral structure and coloration differ notably. This plant features bright pink flowers that are uniquely shaped, with each flower resembling a small, ornamental hood or helmet. The flowers grow closely together on a short, dense spike. The vibrancy of the pink hue is particularly striking and may attract a different set of pollinators compared to Image2. Insects such as bees and hummingbirds, which are attracted to brightly colored flowers, are the likely pollinators. The compact arrangement and shorter inflorescence suggest a design for pollinators that prefer shorter landing platforms or those that can hover, such as hummingbirds. Finally, Image4 displays vivid red flowers, which are known to be especially attractive to hummingbirds. The inflorescence is a spike with flowers that are densely packed together. Each flower is tubular, with petals forming a kind of elongated flag-like structure protruding outward. Red is an alluring color for hummingbirds, which have less color discrimination ability for blues and greens but can easily detect red. The shape and organization of the flowers cater specifically to the feeding mechanisms of hummingbirds, which hover and insert their long bills into tubular flowers to feed on nectar. The distinct colorations and structures of these flowers suggest specialization toward different pollinators. Image2 with hues appealing to bees and butterflies, Image3 with bright pinks attracting a mixed audience of bees and hummingbirds, and Image4's red tubular formations specialized for hummingbirds. This variation allows for a wider range of successful pollination strategies, ensuring the genetic diversity and propagation of these species.

(a) 模型参考答案

In the image, there are three different types of flowers with distinct structures and colorations. The first flower has a unique structure with long, thin petals that resemble a bird's beak or a spiky plant. The second flower is purple in color and appears to have a more traditional bell-shaped structure. The third flower is red and also exhibits a bell-shaped structure. These variations in flower structure and coloration can potentially affect their pollinators differently. Flowers with unique structures like the one resembling a bird's beak may attract specific types of pollinators, such as birds or insects with specialized mouthparts designed to access these unusual shapes. On the other hand, flowers with more traditional bell-shaped structures might appeal to a broader range of pollinators, including bees, butterflies, and moths. The colors of the flowers can also play a role in attracting certain pollinators. For example, purple flowers tend to attract pollinators like bees and hummingbirds, while red flowers often attract butterflies and moths. Therefore, the variation in flower structure and coloration in these images could influence which pollinators visit each type of flower, ultimately affecting cross-pollination and seed production.

(b) 模型生成的回答

图 5. 模型生成的回答与参考答案对比

为了更直观地展示模型的生成能力，本文选取了一个多轮对话任务样例，并将模型生成的回答与参考答案进行对比，具体如图5所示。在比较花朵结构和颜色的任务中，模型对视觉描述在某些情况下是有限且不正确的，但它识别了一些关于花颜色如何影响传粉等因素。

5.3 结果讨论

通过复现实验可以发现尽管使用了轻量化的 GPT-4o-mini 作为评估工具，实验结果仍在主要指标上与原论文保持一致。这也表明 MMDU 数据集在评估多模态对话任务中的能力具有稳健性并且 Qwen-VL 模型在多轮对话任务中展现出较强的生成能力，但仍需进一步提升对长上下文和复杂语义关系的建模能力。

6 总结与展望

本研究复现了论文中关于 Qwen-VL 模型在 MMDU 数据集上的实验，重点分析了模型在多轮多模态对话任务中的性能表现。通过严格的实验设计和多维度评估成功验证了原论文的主要结论，并进一步体现了 MMDU 数据集作为多模态对话理解任务基准的价值。

实验结果表明，MMDU 数据集在多模态对话任务中展现出了高度的评估能力，其复杂的多轮对话场景和任务多样性能够有效测试多模态大模型在真实应用中的表现。复现结果进一步验证了 Qwen-VL 模型作为一种开源 LVLM，在处理多模态任务时具备较强的生成能力，尤其在回答准确性、上下文逻辑连贯性和视觉感知能力上表现优异。然而，实验结果也揭示了当前开源模型在处理长上下文依赖和复杂图像关系建模方面仍存在显著瓶颈。此外，GPT-4omini 对部分主观性指标（如创造力和丰富性）的评分产生了一定影响，但总体趋势与原论文保持一致，验证了实验的可靠性。

基于复现实验的发现，后续可以考虑进一步优化模型的长上下文建模能力，例如通过扩大上下文窗口或引入更高效的位置编码方法；加强模型对多图像语义关系的捕捉能力以提升复杂视觉任务的表现；以及扩展数据集的多语言覆盖范围和领域任务类型来为模型能力评估提供更全面的参考等。

参考文献

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[5] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

[6] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[7] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[8] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[9] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[10] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[13] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2025.

[14] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[15] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiao wen Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and Jiaqi Wang. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *ArXiv*, abs/2406.11833, 2024.

[16] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[17] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[18] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[19] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[21] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[22] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023.

[23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[24] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

[25] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[26] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022.

[27] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[28] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[29] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[30] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[31] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[32] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[33] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022.