ConvBench: A Multi-Turn Conversation Evaluation Benchmark with Hierarchical Capability for Large Vision-Language Models

Shuo Liu[1], Kaining Ying[1,2], Hao Zhang[1,3], Yue Yang[1,4], Yuqi Lin[1,2], Tianle Zhang[1,5], Chuanhao Li[1,6], Yu Qiao[1], Ping Luo[1,7], Wenqi Shao[1⋆], and Kaipeng Zhang[1⋆]

[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[2] Zhejiang University, Hangzhou, China
[3] Xi'an Jiao Tong University, Xi'an, China
[4] Shanghai Jiao Tong University, Shanghai, China
[5] The University of Electronic Science and Technology of China, Chengdu, China
[6] Beijing Institute of Technology, Beijing, China
[7] The Chinese University of HongKong, HongKong, China
liushuo@pjlab.org.cn, shaowenqi@pjlab.org.cn, zhangkaipeng@pjlab.org.cn

Abstract. This paper presents ConvBench, a novel multi-turn conversation evaluation benchmark tailored for Large Vision-Language Models (LVLMs). Unlike existing benchmarks that assess individual capabilities in single-turn dialogues, ConvBench adopts a three-level multimodal capability hierarchy, mimicking human cognitive processes by stacking up perception, reasoning, and creativity. Each level focuses on a distinct capability, mirroring the cognitive progression from basic perception to logical reasoning and ultimately to advanced creativity. ConvBench comprises 577 meticulously curated multi-turn conversations encompassing 215 tasks reflective of real-world demands. Automatic evaluations quantify response performance at each turn and overall conversation level. Leveraging the capability hierarchy, ConvBench enables precise attribution of conversation mistakes to specific levels. Experimental results reveal a performance gap between multi-modal models, including GPT4-V, and human performance in multi-turn conversations. Additionally, weak fine-grained perception in multi-modal models contributes to reasoning and creation failures. ConvBench serves as a catalyst for further research aimed at enhancing visual dialogues.

Keywords: Multi-Turn Conversation Evaluation · Progressive Evaluation · Large Vision-Language Model

1 Introduction

Large Vision-Language Models (LVLMs) [10, 29, 37, 42], have demonstrated remarkable success in various multimodal applications such as open-world visual

⋆ Equal Corresponding Authors

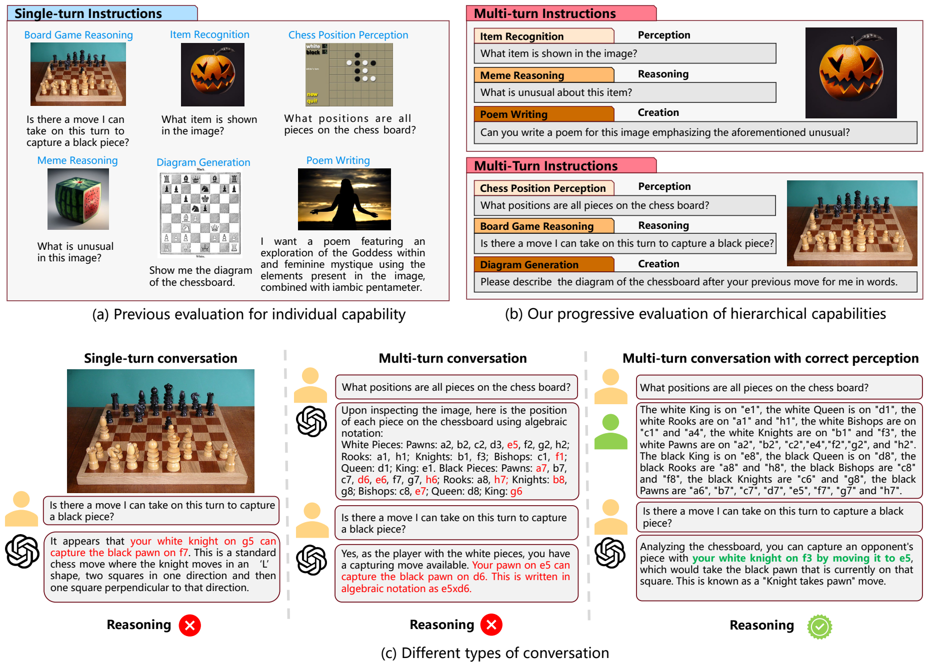


Fig. 1: The comparison between previous evaluation benchmarks (a) and our ConvBench (b). Previous benchmarks assess capabilities independently in a single-turn conversation, while our ConvBench evaluates multi-turn conversation by progressively assessing perception, reasoning, and creativity. (c) shows that multi-turn conversations with hierarchical capabilities can easily attribute mistakes to perception errors.

question answering [13,29], visual dialogue [43], and medical service [21,34]. Due to the great potential in advancing artificial general intelligence, a tremendous surge of research activity has been devoted to improving the performance of LVLMs by various techniques, including effective training strategies [10,17,24,29], high-quality image-text dataset [4,10,27], robust model architectures [13,27]. However, previous benchmarks such as VQAv2 [16] and COCO Captioning [6] are not enough to assess the performance of these LVLM models, which are designed to solve user's general-purpose requests. Therefore, it is urgent to build a challenging evaluation benchmark to measure the advancement of LVLMs.

Recent studies [12,31,40,48] have tackled this challenge by introducing a variety of evaluation benchmarks modeled on visual question-answering. Notably, LVLM-eHub [40], MME [12], SEED-Bench [20], and MMBench [31] have amassed numerous test samples to evaluate key multimodal capabilities such as perception and reasoning. Despite their simplicity, these benchmarks can reveal the downside of current LVLMs. In addition, VisIT-Bench [3] measures how LVLMs respond to various real-world requests by collecting a broad spectrum of tasks from humans. Mathvista [33] and MMMU [49] assess the reasoning and comprehension on expert-level domain knowledge.

However, the above benchmarks [12, 33, 40, 49] evaluate each capability dimension independently in a single-turn conversation as shown in Figure 1 (a), which suffers from two limitations. *i) Naive capability hierarchy.* Previous benchmarks treat different multimodal capabilities independently while ignoring the fact that multimodal capabilities are highly dependent on each other. Evaluating different capabilities independently would make it hard to conduct error attribution. For example, When the model gives the wrong response to a reasoning question, it is unclear whether it is attributed to the perception or reasoning error of LVLMs (see Figure 1 (c)). *ii) Disparity in human preference.* Multi-turn conversation and instruction-following ability are critical elements for human preference. Most previous benchmarks for LVLMs focus on single-turn tests with one instruction and response. However, multi-turn dialogue is the more likely way for general-purpose assistants to collaborate with humans to solve diverse tasks.

To tackle these challenges, we introduce ConvBench, a benchmark designed for multi-turn conversation evaluation that progressively examines the perception, reasoning, and creativity capabilities of LVLMs. Inspired by the fact that humans reason based on what they perceive and generate new ideas through a combination of perceptual and reasoning skills, we build a three-level hierarchy of multimodal capabilities ranging from perception to reasoning and finally to creativity. Such a capability hierarchy is instantiated for each test sample within a framework of multi-turn dialogue, as shown in Figure 1(b). Throughout the conversation, LVLMs are tasked with progressively addressing challenges across these three levels. Evaluating LVLMs using a multi-turn conversation framework with capability hierarchy enables error attribution as shown in Figure 1(c).

To establish the capability hierarchy, each instance in ConvBench is composed of an input image, three progressive instructions for the three-level capability hierarchy, three human-verified references, and an instruction-conditioned caption verified by humans. Specifically, the annotation starts by creating three progressive instructions based on an input image in a multi-turn manner. Initially, we curate the perception instruction for the first level, followed by the reasoning and creativity instructions, which are generated in response to the instructions and answers at preceding levels. We then annotate image captions tailored for these instructions following [3]. GPT-4V, with the help of instruction-conditioned captions, produces preliminary outputs by feeding it input images and instructions. A subsequent human verification step is employed to ensure the high quality of the reference responses.

Overall, ConvBench comprises 577 meticulously curated multi-turn QA samples, spanning 71, 65, and 79 distinct types of perception, reasoning, and creation tasks, as depicted in Figure 2, respectively. We assess 19 publicly available LVLMs, including the advanced GPT4-V [42], employing various assessment methods such as direct grading and pairwise comparison. The evaluation is conducted in an ablative manner, enabling error attribution, as illustrated in Figure 1(c). The evaluation results reveal several innovative findings: i) Our ConvBench poses a substantial challenge for current LVLMs, notably GPT4-V [42], which only achieves 39.51% overall score in pairwise evaluation. ii) Through extensive ablative evaluation, we conclude that weak perception capability undermines LVLMs' reasoning and creativity and limited reasoning capacity also hinders creativity. iii) LVLMs demonstrate weak performance in perception, particularly in fine-grained recognition, object detection, and tasks related to detailed descriptions.

The contributions of this work are summarized as follows: Firstly, we introduce ConvBench, a multi-turn conversation evaluation benchmark, to assess various LVLMs. ConvBench comprises 577 meticulously curated test samples, covering a wide range of multimodal tasks. Secondly, ConvBench is built upon a three-level hierarchy of multimodal capabilities from basic perception to intricate reasoning, and to advanced creativity. This capability hierarchy facilitates error attribution. Thirdly, our extensive evaluation of diverse LVLMs reveals that ConvBench represents a significant challenge for current LVLMs. For instance, even the advanced [42] only achieves 39.51%, 38.47%, 39.34%, and 37.61% overall conversation, perception, reasoning, and creativity score, respectively, in pairwise evaluation method.

2 Related Work

Large Vision-Language Models. Building upon the achievements of Large Language Models (LLMs) [9,35,39], Large Vision-Language Models (LVLMs) [2, 7,10,13,19,29,37,42,43,53] have recently showcased remarkable proficiency across a variety of tasks, demonstrating advanced perception, reasoning, and creative capabilities. A favored approach to enhancing LVLMs involves integrating visual knowledge into the semantic framework of LLMs, thereby leveraging the LLMs' strong performance in interpreting and responding to prompts. For instance, BLIP-2 [23] introduces the Q-Former to synchronize vision foundation models with LLMs without modifying the underlying models. MiniGPT4 [53] utilizes a straightforward fully connected layer, requiring only a minimal set of caption data. LLaVA [29] enhances the LLM with high-quality instructional data generated by GPT-4. QWen-VL [2] undergoes fine-tuning with high-resolution images, employing multi-task training strategies. Moreover, mPLUG-DocOwl [43] expands the capabilities of LVLMs to include document understanding by processing digital document data. Given that instructions for LVLMs challenge both the vision and language processing capabilities simultaneously, there is a pressing need for a benchmark designed to evaluate the two functions in a nuanced manner at the same time. ConvBench would facilitate a comprehensive assessment of LVLMs' abilities in tasks that demand intricate coordination between visual perception and linguistic interpretation.

Large Vision-Language Models Benchmarks. With the advancement of Vision-Language Models (LVLMs), existing standard evaluation benchmarks

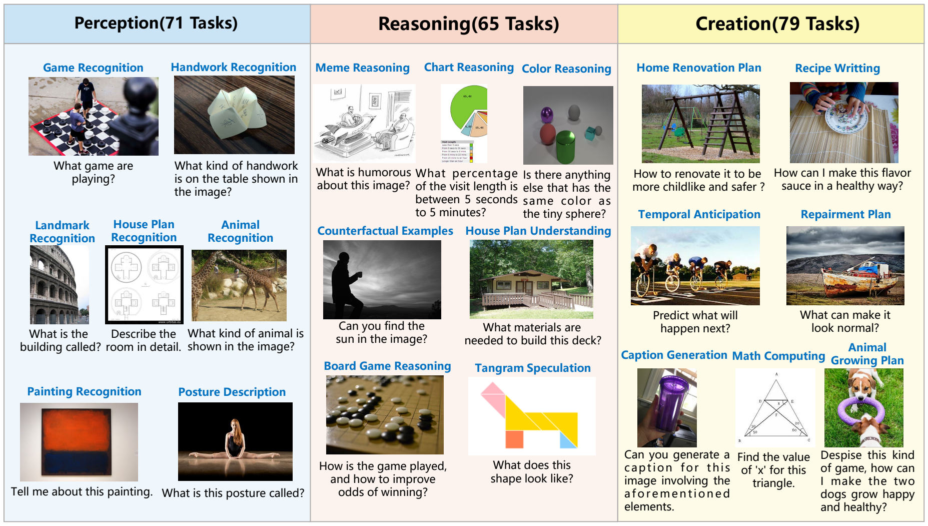


Fig. 2: Visualization of example tasks in ConvBench. It consists of 215 tasks constructed in perception, reasoning, and creation hierarchy.

like MSCOCO [26], GQA [18], VQA [1, 15], etc., are no longer sufficient to assess the comprehensive multimodal abilities of LVLMs. In response, a variety of benchmarks have been developed specifically for LVLM evaluation, including OwlEval [43], LAMM [45], LVLM-eHub [41], SEED [20], MMBench [31], and MM-Vet [48]. These benchmarks primarily focus on assessing basic perceptual abilities. In addition, VisIT-Bench [3] covers a broad spectrum of tasks, ranging from simple recognition to complex reasoning. Recent research has also introduced LVLM benchmarks requiring expert-level domain knowledge and intricate reasoning, such as MathVista [33], MMMU [49] and MMT-Bench [46]. However, these benchmarks tend to address perception, reasoning, and creation tasks in isolation, without establishing connections among these tasks. Furthermore, the current benchmarks predominantly focus on single-turn interactions, comprising one instruction and one response, with less emphasis on multi-turn interactions between users and chatbots. The ConvBench addresses these gaps by not only offering a progressive evaluation that moves from basic perception through logical reasoning to advanced creation but also by evaluating LVLMs' capabilities in multi-turn conversational contexts.

3 ConvBench

3.1 Overview of ConvBench

We introduce the ConvBench Benchmark, an innovative benchmark designed to evaluate multi-turn conversation capabilities and conduct detailed, progressive

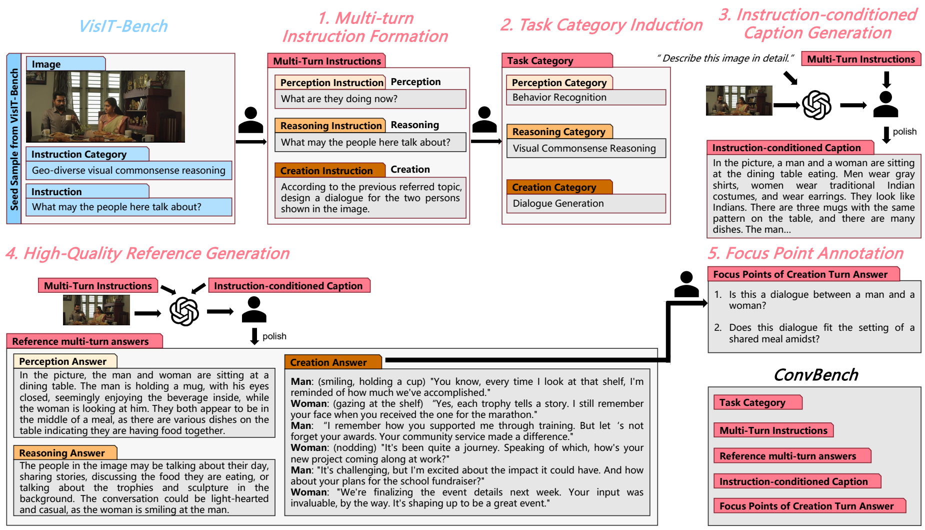


Fig. 3: The pipeline of data curation. We develop three multi-turn instructions for each image to assess the perception, reasoning, and creation capabilities. We also annotate the referenced answers to facilitate automatic evaluation and error analysis.

assessments of Large Vision-Language Models (LVLMs). This benchmark concentrates on three pivotal skills of LVLMs: perception, reasoning, and creation. These skills are meticulously arranged in a bottom-up way, facilitating error attribution.

The ConvBench includes 577 premium image-instruction pairs tailored for multi-turn dialogues. Each pair is structured around three sequential instructions, each targeting a distinct cognitive skill—beginning with perception, followed by reasoning, and culminating in creation. This structure underscores the cognitive evolution from basic perceptual comprehension to logical reasoning and finally to sophisticated creative expression. Encompassing 215 tasks, the benchmark is divided into 71 tasks focused on perception, 65 on reasoning, and 79 on creation. Drawn from real-world contexts, these demanding tasks are seen as benchmarks for the competencies of advanced LVLMs, as depicted in Figure 2. Beyond the instructions, we manually annotate each instruction's referenced answer, which facilitates error attribution and automatic evaluation in our evaluation framework, ConvEval.

3.2 Data Curation Process

Data Collection. Our benchmark collection is structured into five distinct stages, as depicted in Figure 3. These stages are as follows:

i) Multi-turn Instruction Formation. Starting with VisIT-Bench as our foundational guide, we develop three-level instructions for each sample from

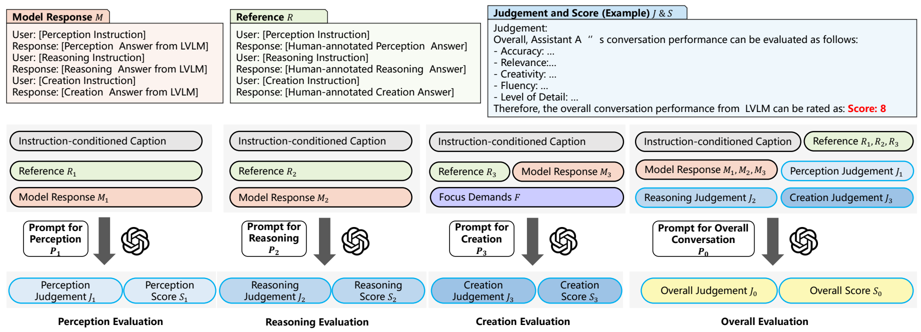


Fig. 4: ConvBench Evaluation Pipeline. ֍ indicates the ChatGPT-3.5.

VisIT-Bench. The three-level instructions align with perception, reasoning, and creation capability.

ii) Task Category Induction. We derive the task categories by inducing them from instructions. The bottom-up approach to task collection guarantees that the tasks under investigation are tailored to meet real-world requirements. ConvBench consists of 215 tasks, and we have included the details in the supplymentarl materials.

iii) Instruction-Conditioned Caption Annotation. We then proceed to annotate detailed captions for images tailored to the instructions. We first prompt GPT-4V with "Describe this image in detail." We then polish the responses according to the instructions to obtain the final instruction-conditioned caption. These captions serve a dual purpose: they provide a comprehensive description of the image relevant to executing the instructions and support the generation of raw reference answers in the next step. Additionally, these annotated captions are instrumental in evaluating responses from candidate models, ensuring a robust assessment framework.

iv) High-Quality Reference Generation. For each sample, we feed GPT-4V with the instruction-conditioned caption, the image, multi-turn instructions, and our well-designed prompt in a multi-turn conversation fashion to generate each instruction's response. We meticulously refine these responses as reference answers, enhancing their quality and relevance.

v) Focus Point Annotation. The creativity instruction is an open-ended question without a standard answer. Therefore, we annotate specific focus points related to each creation instruction. These annotations are used as criteria to assess whether the model produces instructive answers to the instruction, seeing Step 5 in Figure 3.

4 Multi-turn Conversation Evaluation

In this section, we introduce ConvEval, an evaluation pipeline designed for multi-turn conversation assessment. Building upon prior research [3, 52], which has shown the effectiveness of employing Large Language Models (LLMs) with tailored prompts for automatic evaluation, ConvEval utilizes a large language model (LLM) as the evaluator to assess model responses across various levels, ensuring cost-effectiveness and efficiency in evaluation procedures.

Specifically, ConvEval comprises two kinds of grading schemes for models' responses: direct and pairwise grading. The direct grading will give a score of 0-10 for the model's responses under specific prompts while the pairwise grading would output a model preference when comparing the response of the test model and reference. In the sequel, we introduce the evaluation pipeline in Sec. 4.1 and present direct and pairwise grading in Sec. 4.2 and Sec. 4.3, respectively.

4.1 Pipeline of ConvEval

ConEval comprises four key components: perception, reasoning, creation, and overall conversation evaluation modules, as shown in Figure 4. The multi-turn conversation evaluation is complicated as creativity may be affected by inaccurate perception or reasoning in previous levels. To enable error attribution, we recursively use ConvEval in three settings as follows. For clarity, we denote the instruction, model response, reference, and focal points as I_i, M_i, R_i, and F_i at each level, respectively, where i indicates the level index. Note that F_1 and F_2 are null focal points.

i) ConvEval with model's responses. In this setting, the model response at each level is obtained by $M_i = f(I_0^{i-1}, M_0^{i-1}, I_i), i = 1, 2, 3$ where f denotes the model inference function. The evaluation process can be expressed as $S_i, J_i = \text{ConvEval}(\{M_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \{F_i\}_{i=1}^3; P_i), i = 1, 2, 3$ where S_i, J_i, and P_i are the overall capability score, judgment from the LLM, and prompt specific to level i. Finally, we feed all instructions, responses, focal points, and judgments into LLM to obtain the overall conversation score as given by $S_0, J_0 = \text{ConvEval}(\{M_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \{J_i\}_{i=1}^3; P_0)$ where P_0 is the prompt to obtain S_0.

ii) ConvEval with perfect perception reference. To enable error attribution, we first use perfect perception reference as the response at the perception level. In this way, the influence of inaccurate perception on reasoning, creation, and overall conversation can be derived. In this setting, the model inference can be written as $\hat{M}_1 = R_1$ and $\hat{M}_i = f(I_0^{i-1}, \hat{M}_0^{i-1}, I_i), i = 2, 3$. The evaluation process without considering perception error can be expressed as $\hat{S}_i, \hat{J}_i = \text{ConvEval}(\{\hat{M}_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \{F_i\}_{i=1}^3; P_i), i = 2, 3$. Finally, the overall conversation score without considering perception score can be given by $\hat{S}_0, \hat{J}_0 = \text{ConvEval}(\{\hat{M}_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \{\hat{J}_i\}_{i=2}^3; P_0)$. By comparing S_i and $\hat{S}_i$ $(i = 0, 2, 3)$, we can see how perception error affects the performance of overall conversation, reasoning and creativity.

iii) ConvEval with perfect perception and reasoning reference. We further explore how reasoning errors affect creativity and overall conversation. To this end, we use perfect perception and reasoning reference as responses at perception and reasoning levels. In this setting, the model inference can be written as $\tilde{M}_i = R_i, i = 1, 2$ and $\tilde{M}_i = f(I_0^{i-1}, \tilde{M}_0^{i-1}, I_i), i = 3$. The evaluation process without considering perception and reasoning error can be expressed as $\tilde{S}_i, \tilde{J}_i = \text{ConvEval}(\{\tilde{M}_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \{F_i\}_{i=1}^3; P_i), i = 3$. Finally, the overall conversation score without considering perception and reasoning score can be given by $\tilde{S}_0, \tilde{J}_0 = \text{ConvEval}(\{\tilde{M}_i\}_{i=1}^3, \{R_i\}_{i=1}^3, \tilde{J}_3; P_0)$. By comparing $\hat{S}_i$ and $\tilde{S}_i$ $(i = 0, 3)$, we can check how reasoning errors affect the performance of overall conversation and creativity.

Note that the function in i) - iii) ConvEval$(\cdot)$ can be pairwise or direct grading schemes as described in the following.

4.2 Pairwise Grading

When ConvEval$(\cdot)$ employs pairwise grading scheme, a model preference is returned when comparing the response of the test model and reference. Following the evaluation pipeline in Sec. 4.1, we feed LLM with the following components for evaluation: system prompts, instructions, model responses, human-verified references, and focal points at all levels. To make sure that LLM can infer the correct answers, the instruction-condition caption tailored to the specific instructions is also added.

The response sets are anonymously presented to the LLM in a random order. The LLM is tasked with a pairwise comparison to decide which set of responses is superior. The percentages of cases where the LLM prefers the output from the model rather than the human-verified reference output are obtained as the final metric, *i.e.* win rate. The system prompts, detailed in Appendix, encourage the LLM to engage in a step-by-step thought process, making its reasoning explicit. In our forced-choice setup, ties are not permitted; thus, if the LLM deems the responses of equal quality, it is instructed to select one arbitrarily. This prompt, including the presentation of two full conversations within a single prompt, addresses the challenge of the LLM potentially struggling to accurately recall previous responses by the assistant, as noted in previous work [52].

4.3 Direct Grading

When ConvEval$(\cdot)$ employs the pairwise grading scheme, a score of 0-10 is returned when comparing the response of the test model and reference. Unlike pairwise comparison, the response sets are now identifiable; specifically, the sets from the tested LVLM and a human participant are labelled as Assistant A's Conversation and Reference Answer, respectively. In this context, the LLM judge is tasked with assigning scores directly to each turn answer and the overall conversation quality.

The prompt designed for these scenarios is detailed in Appendix. This prompt not only encourages the LLM to engage in a chain-of-thought process but also includes the presentation of two full conversations within a single prompt. This approach is aimed at enhancing the LLM's ability to accurately evaluate and score the responses by maintaining a clear context and facilitating a comprehensive assessment.

5 Experiment and Analysis

We undertake a thorough evaluation of 19 LVLMs using the ConvBench. Section 5.1 outlines the evaluation framework, detailing the LVLMs under study and the methodologies employed for assessment. Section 5.2 delves into progressive evaluation comparisons and analysis, offering insights into how these models perform over a range of tasks. Section 5.3 focuses on multi-turn conversation comparisons and analysis, examining the models' capabilities in engaging in dialogues that require sustained interaction.

5.1 Evaluation Settings

LVLMs. We evaluate 19 representative LVLMs, including GPT-4V [42], GeminiProVision [37], Reka Flash [38], Qwen-VL-Chat [2], LLaMA-Adapter-v2 [13], XComposer [51], XComposer2 [10], mPLUG-Owl2 [43], Monkey [24], Otter [19], MMAlaya [32], MiniGPT-4 [53], InternVL-Chat-V1-2 [7], InternVL-Chat-V1-5 [7], LLaVA (V1.5) models [29] like LLaVA-7B, LLaVA-13B, ShareGPT4V models [4] like ShareGPT4V-7B, ShareGPT4V-13B, BLIP2 models [23] like BLIP2-FLAN-T5-XL and BLIP2-FLAN-T5-XXL. Appendix provides the models' configuration.

Evaluation Methods. We supplied each Large Vision-Language Model (LVLM) with an image accompanied by a set of carefully curated progressive instructions to elicit corresponding responses. The chatting prompt specific to each model played a crucial role in the generation of multi-turn responses. Upon gathering all responses, we employed our proposed *ConvBenchEval* methodology to perform a quantitative analysis. This involved comparing the model-generated responses with high-quality, human-verified reference answers. We utilized 9 distinct evaluation scores as metrics to assess the multi-turn conversation capabilities and to provide detailed, progressive evaluation results.

5.2 Progressive Evaluation Comparisons and Analysis

The outcomes of the evaluation results are detailed in Table 1 and Table 2. In these tables, S_1, S_2, and S_3 denote the scores for perception, reasoning, and creation, respectively. Meanwhile, $\hat{S}_2$ and $\hat{S}_3$ correspond to the scores for reasoning and creation, respectively, but under conditions of perfect perception. $\tilde{S}_3$ is the score for creation, assuming perfect conditions for both perception and reasoning. We present our principal insights from the evaluation results as follows:

(1) **The Challenge of Progressive Evaluation:** This benchmark sets formidable challenges for modern models. GPT-4V, despite being a sophisticated model, shows only modest achievements in perception, reasoning, and creation. In the Pairwise Grading approach, it attains scores of 38.47, 39.34, and 37.61 for perception, reasoning, and creation, respectively. On the other hand, through the Direct Grading method, it achieves 7.30, 7.48, and 7.12, correspondingly. Claude [36] is infinitely close to GPT-4V in performance. However, according

Table 1: Comparison of Performance for LVLMs on ConvBench. Quantitative ConvBench Evaluation Results for 19 LVLMs with Pairwise Grading method. The results in the table are win-rate vs human. The colors blue and red indicate positive and negative differences, respectively. R_2 is defined as $(S_1+S_2+S_3)/3$, indicative of the mean performance over three turns. Meanwhile, R_1 is computed as $(R_2+S_0)/2$, representing the model's overall score.

Model	R_1	R_2	S_1	S_2	S_3	S_0	$\hat{S}_2(\hat{S}_2-S_2)$	$\hat{S}_3(\hat{S}_3-S_3)$	$\hat{S}_0(\hat{S}_0-S_0)$	$\tilde{S}_3(\tilde{S}_3-\hat{S}_3)$	$\tilde{S}_0(\tilde{S}_0-\hat{S}_0)$
GPT-4V	**39.51**	**38.47**	**38.47**	**39.34**	**37.61**	**40.55**	**47.31**(16.97)	37.78(0.61)	37.61(2.94)	38.99(1.21)	38.30(0.69)
Claude	36.60	37.49	**38.99**	39.17	34.32	35.70	45.93(6.76)	**38.99**(4.67)	**43.15**(7.45)	**39.16**(0.17)	**40.21**(2.94)
Reka Flash	25.60	24.67	25.13	27.56	21.32	26.52	32.93(5.37)	22.88(1.56)	25.82(0.70)	24.78(1.90)	26.00(0.18)
InternVL-Chat-V1-2	21.17	22.41	24.96	21.31	20.97	19.93	32.06(10.75)	28.25(7.28)	29.64(9.71)	33.62(5.37)	35.18(5.54)
ShareGPT4V-13B	17.56	17.45	17.85	18.72	15.77	17.68	32.58(13.86)	30.33(14.56)	28.94(11.26)	32.41(2.08)	31.54(2.60)
LLaVA-V1.5-13B	16.93	18.08	20.45	18.02	15.77	15.77	32.76(14.74)	25.65(9.88)	28.94(13.17)	32.06(6.41)	28.94(0.00)
ShareGPT4V-7B	16.32	16.87	16.81	19.24	14.56	15.77	32.76(13.52)	23.05(8.49)	25.13(9.36)	29.46(6.41)	30.33(5.20)
LLaVA-V1.5-7B	16.15	17.56	19.06	19.24	14.38	14.73	33.80(14.56)	23.22(8.84)	26.52(11.79)	30.68(7.46)	32.58(6.06)
XComposer2	15.83	16.41	17.16	19.06	13.00	15.25	30.50(11.44)	20.97(7.97)	22.36(7.11)	28.60(7.63)	29.81(7.45)
mPLUG-Owl2	14.93	15.83	17.50	17.16	12.82	14.04	27.90(10.74)	17.50(4.68)	20.80(6.76)	24.26(6.76)	24.44(3.64)
Qwen-VL-Chat	14.33	14.62	16.29	18.37	9.19	14.04	28.25(9.88)	16.12(6.93)	22.70(8.66)	25.30(9.18)	26.52(3.82)
MiniGPT-4	10.95	10.80	11.61	11.27	9.53	11.09	27.56(16.29)	18.20(8.67)	22.53(11.44)	22.88(4.68)	23.74(1.21)
LLaMA-Adapter-v2	9.04	9.59	8.84	10.92	9.01	8.49	27.38(16.46)	15.60(6.59)	19.41(10.92)	18.37(2.77)	19.24(0.17)
GeminiProVision	8.44	8.55	9.01	9.36	7.28	8.32	21.84(12.48)	12.31(5.03)	15.08(6.76)	23.92(11.61)	23.92(8.84)
MMAlaya	5.55	5.89	7.28	6.41	3.99	5.20	22.53(16.12)	9.88(5.99)	15.25(10.05)	14.21(4.33)	16.81(1.56)
Monkey	3.70	4.10	3.64	5.20	3.47	3.29	16.64(11.44)	7.28(3.81)	10.75(7.46)	13.86(6.58)	15.94(5.19)
Otter	2.78	2.60	3.12	3.12	1.56	2.95	14.21(11.09)	5.37(3.81)	9.01(6.06)	8.49(3.12)	13.00(3.99)
XComposer	1.21	1.73	1.73	1.91	1.56	0.69	12.13(10.22)	2.77(1.21)	8.49(7.80)	10.40(7.63)	12.48(3.99)
BLIP2-FLAN-T5-XXL	0.32	0.29	0.35	0.52	0.00	0.35	3.47(2.95)	1.91(1.91)	2.95(2.60)	5.72(3.81)	8.49(5.54)
BLIP2-FLAN-T5-XL	0.06	0.11	0.00	0.17	0.17	0.00	3.12(2.95)	0.17(0.00)	2.25(2.25)	3.97(3.80)	8.67(6.42)

to progressive evaluation, Claude [36] has greater potential for performance improvement compared to GPT-4V. The least successful models, exemplified by BLIP2, are utterly unsuccessful, underscoring a significant opportunity for advancements. Our analysis exposes a stark discrepancy between the performance of these models and that of humans, highlighting the benchmark's stringent and exacting criteria.

(2) **Weak Perception Undermines LVLMs' Reasoning and Creation Performance:** Under conditions of perfect perception, we see significant improvements in reasoning and creation abilities, as indicated by the data in the $\hat{S}_2$ and $\hat{S}_3$ columns of Table 1 and Table 2. The figures in parentheses reflect the enhancement in reasoning and creation attributed to impeccable perception. Across 19 LVLMs, the average increase in reasoning and creation scores are 11.21 and 5.31, respectively, when using the Pairwise Grading method. Similarly, improvements of 1.25 both are observed in reasoning and creation, respectively, via the Direct Grading approach. These enhancements underscore the challenges models face in accurately interpreting images, which in turn leads to mistakes in reasoning and creative efforts. Such challenges highlight a crucial opportunity for enhancing the visual comprehension capabilities of LVLMs. Our benchmark clar-

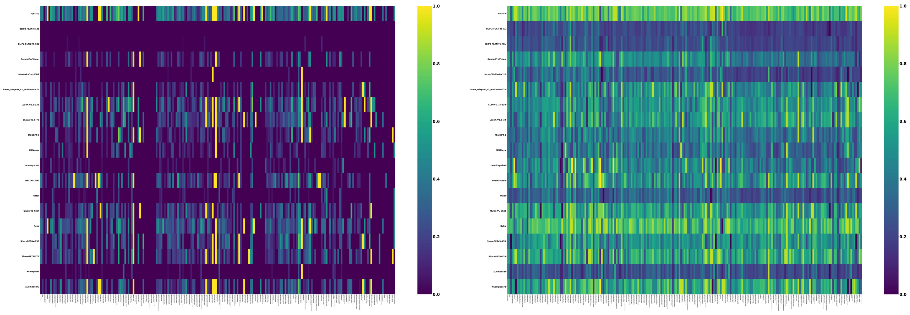
Table 2: Comparison of Performance for LVLMs on ConvBench. Quantitative ConvBench Evaluation Results for 19 LVLMs with Direct Grading method. The results in the table is the average scores of all the samples. The colors blue and red indicate positive and negative differences, respectively. R_2 is defined as $(S_1+S_2+S_3)/3$, indicative of the mean performance over three turns. Meanwhile, R_1 is computed as $(R_2+S_0)/2$, representing the model's overall score.

Model	R_1	R_2	S_1	S_2	S_3	S_0	$\hat{S}_2(\hat{S}_2-S_2)$	$\hat{S}_3(\hat{S}_3-S_3)$	$\hat{S}_0(\hat{S}_0-S_0)$	$\tilde{S}_3(\tilde{S}_3-\hat{S}_3)$	$\tilde{S}_0(\tilde{S}_0-\hat{S}_0)$
GPT-4V	**7.09**	**7.30**	**7.30**	**7.48**	**7.12**	**6.88**	**8.23**(0.75)	**8.00**(0.88)	**8.25**(1.37)	**7.34**(0.66)	**8.18**(0.07)
Claude	6.54	6.75	6.53	7.04	6.68	6.32	7.48(0.44)	7.06(0.38)	7.55(1.23)	7.18(0.12)	8.13(0.58)
Reka Flash	6.78	6.86	6.93	7.25	6.41	6.70	7.10(0.15)	6.41(0.00)	7.32(0.62)	4.95(1.46)	6.95(0.37)
ShareGPT4V-7B	5.83	5.99	6.02	6.14	5.80	5.67	7.19(1.05)	6.77(0.97)	7.31(1.64)	6.93(0.16)	8.19(0.88)
XComposer2	5.82	5.98	5.98	6.17	5.78	5.66	7.35(1.18)	7.04(1.26)	7.66(2.00)	7.00(0.04)	8.20(0.54)
Qwen-VL-Chat	5.54	5.65	5.96	5.78	5.22	5.43	7.04(1.26)	6.53(1.31)	7.26(1.83)	6.57(0.04)	8.00(0.74)
InternVL-Chat-V1-2	5.49	5.69	5.80	5.88	5.39	5.29	6.66(0.78)	6.12(0.73)	6.75(1.46)	6.31(0.19)	7.70(0.95)
LLaVA-V1.5-7B	5.16	5.29	4.95	5.59	5.34	5.03	7.28(1.69)	6.68(1.34)	7.28(2.25)	6.72(0.04)	7.97(0.69)
mPLUG-Owl2	5.04	5.17	4.98	5.38	5.14	4.91	6.77(1.39)	6.64(1.50)	7.22(2.31)	5.93(0.71)	7.62(0.40)
LLaVA-V1.5-13B	4.94	5.14	5.03	5.41	4.99	4.74	7.43(2.02)	7.13(2.14)	7.70(2.95)	6.14(0.99)	7.60(0.10)
ShareGPT4V-13B	4.85	5.03	5.16	5.06	4.86	4.67	7.42(2.36)	7.17(2.31)	7.65(2.98)	6.24(0.93)	7.65(0.00)
LLaMA-Adapter-v2	4.77	4.91	4.77	5.47	4.48	4.64	6.68(1.21)	5.49(1.01)	6.68(2.04)	5.19(0.30)	7.36(0.68)
Monkey	4.49	4.60	5.11	4.68	4.01	4.37	6.28(1.60)	5.66(1.65)	6.76(2.39)	5.39(0.27)	7.30(0.54)
GeminiProVision	4.42	4.60	5.18	4.95	3.66	4.24	6.16(1.21)	5.05(1.39)	6.28(2.04)	5.07(0.02)	7.05(0.77)
MiniGPT-4	3.85	4.04	3.99	4.40	3.73	3.66	6.66(2.26)	5.80(2.07)	6.75(3.09)	4.97(0.83)	7.01(0.26)
MMAlaya	3.60	3.75	4.07	3.91	3.28	3.44	5.64(1.73)	4.76(1.48)	5.91(2.47)	4.02(0.74)	6.47(0.56)
Otter	2.96	3.11	3.33	3.52	2.47	2.80	5.00(1.48)	4.11(1.64)	5.75(2.95)	3.25(0.86)	6.03(0.28)
XComposer	2.61	2.70	2.90	2.82	2.39	2.51	4.67(1.85)	3.84(1.45)	5.30(2.79)	3.90(0.06)	6.47(1.17)
BLIP2-FLAN-T5-XXL	2.37	2.45	2.81	2.59	1.95	2.28	3.18(0.59)	2.35(0.40)	4.03(1.75)	2.59(0.24)	5.75(1.72)
BLIP2-FLAN-T5-XL	2.14	2.21	2.55	2.33	1.74	2.07	2.74(0.41)	2.17(0.43)	3.81(1.74)	2.24(0.07)	5.44(1.63)

ifies the origins of these errors in reasoning and creativity, determining whether they stem from visual perception issues or language reasoning shortcomings. With the aid of human-verified visual comprehension, the authentic strengths of the language module in reasoning and creation will be more precisely evaluated.

(3) **Limited Reasoning Impacts LVLM's Creation Abilities:** Under ideal conditions for perception and reasoning, shifts in creation capabilities are documented in the $\tilde{S}_3$ column of Table 1 and Table 2. The numbers in brackets indicate adjustments in creation scores due to human-verified perception and reasoning accuracy. Among the 19 LVLMs evaluated, an average increase of 6.96 in creation scores is noted with the Pairwise Grading method. This indicates that reasoning inaccuracies can adversely affect LVLMs' performance in creative tasks. On the other hand, a modest average decrease of 0.39 in creation scores is noted with the Direct Grading approach. This minor decline may result from the approach's challenges in recognizing subtle distinctions between specific pairs, which may render the outcomes less stable.

(4) **LVLM's Performance across Various Categories:** As depicted in Figure 5, LVLMs demonstrate weak performance in areas of perception, par-



(a) Visualization of model performance on different tasks for 19 LVLMs with Pairwise Grading. (b) Visualization of model performance on different tasks for 19 LVLMs with Direct Grading.

Fig. 5: The Visualizations of model performance. The horizontal and vertical axes represent task and model names respectively. The performance heatmaps across various tasks for 19 LVLMs, under the two grading strategies. They have high agreement in comparing the performance of different models. GPT-4V and Reka have excellent performance. BLIP2, Otter and Xcomposer have poor performance. However, Figure(a) is darker and has a lower score than Figure(b), which may be due to the Direct Grading method struggles to identify subtle differences between specific pairs resulting in higher scores.

ticularly in fine-grained recognition, object detection, and providing detailed descriptions. Fine-grained recognition encompasses identification in specialized fields such as art, science, medicine, and technology. Object detection pertains to tasks similar to "Chess Position Perception". Detailed descriptions involve elaborate explanations akin to "image description", "chart description", and "scene description", among others. In terms of reasoning, even with perfect perception, LVLMs encounter challenges in abstract reasoning tasks, including mathematical reasoning, game strategy analysis, and solving anagrams. Regarding creation, there remains a significant disparity between LVLMs and human-verified responses, especially in areas like custom service planning, computer programming, and science application questions. In disciplines such as art design and the humanities and social sciences, LVLMs are still unable to offer detailed answers, explanations, or suggestions. Similarly, in fields like science, health & medicine, and technology & engineering, the models continue to struggle with delivering accurate results.

(5) **Comparison Between Pairwise Grading and Direct Grading:** As indicated by the data in Table 1 and Table 2, the two grading strategies exhibit a high level of agreement. GPT-4V ranks first and Reka second under both evaluation methodologies. Other models like ShareGPT4V, LLaVA-V1.5, mPLUG-Owl2, Xcomposer2, and Qwen-VL-Chat also achieve high rankings. In contrast, BLIP2, InterVL-Chat, XComposer, Otter, and MMAlaya are positioned lower in the rankings. Surprisingly, GeminiProVision, an open-source LVLM, does not perform as well, potentially due to the benchmark's challenging instructions and

its weak multi-turn ability leading to adversely affecting its ability to answer questions. As illustrated in Figure 5, the performance heatmaps across various tasks for 19 LVLMs, under the two grading strategies, are comparable. This similarity supports the high level of agreement between the two grading methods, each with its distinct advantages and disadvantages. The Direct Grading provides clear scores but struggles to identify subtle differences between specific pairs, potentially leading to inconsistent results. Conversely, the Pairwise Grading excels at detecting fine distinctions but does not offer numerical scores and may be affected by biases, such as position bias.

5.3 Multi-Turn Conversation Comparisons and Analysis

The results of the multi-turn conversation evaluation are meticulously outlined in Table 1 and Table 2. Within these tables, S_0 represents the scores for multi-turn conversation performance. Concurrently, $\hat{S}_0$ signifies the scores for multi-turn conversation performance under the assumption of flawless perception. Moreover, $\tilde{S}_0$ reflects the score for multi-turn conversation performance, premised on ideal conditions for both perception and reasoning. We delineate our key findings from the multi-turn conversation evaluation as follows:

(1) **The Challenge of Multi-Turn Conversation:** This benchmark presents substantial challenges in multi-turn conversation to contemporary models. GPT-4V, for example, records a modest accuracy of only 40.55 and 6.88 in the Pairwise Grading and Direct Grading approaches, respectively. This highlights the existing disparity between the multi-turn conversation capabilities of models and humans, underscoring the need for further improvement.

(2) **Weak Perception and Reasoning Impact Multi-Turn Conversation Performance:** When comparing the S_0 column against the $\hat{S}_0$ and $\tilde{S}_0$ columns in Table 1 and Table 2, the numbers in parentheses illustrate the improvement in overall multi-turn conversation performance resulting from flawless perception and reasoning. Among 19 LVLMs, an average increase of 12.29 in multi-turn conversation performance is noted under conditions of perfect perception and reasoning with the Pairwise Grading method. Additionally, an improvement of 2.75 in multi-turn conversation performance is observed using the Direct Grading approach. These enhancements indicate that the performance of individual turns influences the overall quality of the conversation. Our benchmark establishes a framework for exploring multi-modal, multi-turn conversations.

(3) **Multi-Turn Conversation Score vs. Average Score of Three Turns:** As indicated in Table 1 and Table 2, the average differences between the multi-turn conversation scores (S_0) and the corresponding average scores of the three turns (R_2) for the 19 LVLMs are 0.44 and 0.29, respectively. This suggests that the multi-turn conversation scores are generally lower than the corresponding average scores of the three turns. This discrepancy implies that the LLM judge evaluates more than just the performance of individual responses. The LLM judge should also consider the instruction-following ability in a multi-turn conversation.

6 Conclusion

We introduce ConvBench, a benchmark focusing on the three critical abilities of LVLMs: perception, reasoning, and creation. These capabilities are thoughtfully sequenced to facilitate a comprehensive exploration of LVLMs' extensive potential. In parallel, we establish an evaluation pipeline designed to conduct progressive and multi-turn conversation assessments for LVLMs. Our research uncovers a discrepancy between model performances and human capabilities in multi-turn conversations, highlighting that inadequate perception can result in failures in reasoning and creative tasks. We aim for ConvBench to clearly identify and illuminate the shortcomings of multimodal AI systems.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision **123**, 4 – 31 (2015), https://api.semanticscholar.org/CorpusID:3180429 5
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023) 4, 10, 20
3. Bitton, Y., Bansal, H., Hessel, J., Shao, R., Zhu, W., Awadalla, A., Gardner, J., Taori, R., Schimdt, L.: Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. ArXiv **abs/2308.06595** (2023), https://api.semanticscholar.org/CorpusID:260887670 2, 3, 5, 7
4. Chen, L., Li, J., wen Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. ArXiv **abs/2311.12793** (2023), https://api.semanticscholar.org/CorpusID:265308687 2, 10
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023) 20
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 2
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. ArXiv **abs/2312.14238** (2023), https://api.semanticscholar.org/CorpusID:266521410 4, 10
8. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 20
9. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Valter, D., Narang, S., Mishra, G., Yu, A.W., Zhao, V., Huang, Y., Dai, A.M., Yu, H., Petrov, S., hsin Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models.

ArXiv **abs/2210.11416** (2022), https://api.semanticscholar.org/CorpusID:253018554 4
10. Dong, X.W., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. ArXiv **abs/2401.16420** (2024), https://api.semanticscholar.org/CorpusID:267311889 1, 2, 4, 10
11. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024) 20
12. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023) 2, 3
13. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.J.: Llama-adapter v2: Parameter-efficient visual instruction model. ArXiv **abs/2304.15010** (2023), https://api.semanticscholar.org/CorpusID:258418343 2, 4, 10
14. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023) 20
15. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. International Journal of Computer Vision **127**, 398 – 414 (2016), https://api.semanticscholar.org/CorpusID:8081284 5
16. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021) 2
18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6693–6702 (2019), https://api.semanticscholar.org/CorpusID:152282269 5
19. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. ArXiv **abs/2305.03726** (2023), https://api.semanticscholar.org/CorpusID:258547300 4, 10, 20
20. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. ArXiv **abs/2307.16125** (2023), https://api.semanticscholar.org/CorpusID:260334888 2, 5
21. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023) 2
22. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023) 20
23. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International

Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:256390509 4, 10
24. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. ArXiv **abs/2311.06607** (2023), https://api.semanticscholar.org/CorpusID:265150038 2, 10
25. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. arXiv preprint arXiv:2311.06607 (2023) 20
26. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014), https://api.semanticscholar.org/CorpusID:14113767 5
27. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 2
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 20
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023), https://api.semanticscholar.org/CorpusID:258179774 1, 2, 4, 10
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 20
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? ArXiv **abs/2307.06281** (2023), https://api.semanticscholar.org/CorpusID:259837088 2, 5
32. Ltd., D.: mmalaya. https://github.com/DataCanvasIO/MMAlaya (2024) 10, 20
33. Lu, P., Bansal, H., Xia, T., Liu, J., yue Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts (2023), https://api.semanticscholar.org/CorpusID:264491155 3, 5
34. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E.P., Rajpurkar, P., Leskovec, J.: Med-flamingo: a multimodal medical few-shot learner (2023) 2
35. OpenAI: Introducing chatgpt. (2022), https://openai.com/blog/chatgpt 4
36. Team, C.: Meet claude; claude is a family of foundational ai models that can be used in a variety of applications. (2024), https://www.anthropic.com/claude/ 10, 11, 20
37. Team, G.: Gemini: A family of highly capable multimodal models (2023) 1, 4, 10, 20
38. Team, R.: Reka flash: An efficient and capable multimodal language model (2024), https://reka.ai/reka-flash-an-efficient-and-capable-multimodal-language-model/ 10, 20
39. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404 4
40. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023) 2, 3

41. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y.J., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. ArXiv **abs/2306.09265** (2023), https://api.semanticscholar.org/CorpusID:259165040 5
42. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9**(1), 1 (2023) 1, 3, 4, 10, 20
43. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality. ArXiv **abs/2304.14178** (2023), https://api.semanticscholar.org/CorpusID:258352455 2, 4, 5, 10
44. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023) 20
45. fei Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Sheng, L., Bai, L., Huang, X., Wang, Z., Ouyang, W., Shao, J.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. ArXiv **abs/2306.06687** (2023), https://api.semanticscholar.org/CorpusID:259138958 5
46. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., Lei, J., Lu, Q., Chen, R., Xu, P., Zhang, R., Zhang, H., Gao, P., Wang, Y., Qiao, Y., Luo, P., Zhang, K., Shao, W.: Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi (2024) 5
47. You, H., Sun, R., Wang, Z., Chen, L., Wang, G., Ayyubi, H.A., Chang, K.W., Chang, S.F.: Idealgpt: Iteratively decomposing vision and language reasoning via large language models. ArXiv **abs/2305.14985** (2023), https://api.semanticscholar.org/CorpusID:258865952 22
48. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. ArXiv **abs/2308.02490** (2023), https://api.semanticscholar.org/CorpusID:260611572 2, 5
49. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multidiscipline multimodal understanding and reasoning benchmark for expert agi. ArXiv **abs/2311.16502** (2023), https://api.semanticscholar.org/CorpusID:265466525 3, 5
50. Zhang, P., Dong, X., Wang, B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Zhang, W., Yan, H., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 (2023) 20
51. Zhang, P., Wang, X., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. ArXiv **abs/2309.15112** (2023), https://api.semanticscholar.org/CorpusID:262824937 10
52. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J., Stoica, I.: Judging llm-as-a-

judge with mt-bench and chatbot arena. ArXiv **abs/2306.05685** (2023), https://api.semanticscholar.org/CorpusID:259129398 7, 9

53. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. ArXiv **abs/2304.10592** (2023), https://api.semanticscholar.org/CorpusID:258291930 4, 10, 20

A LVLMs Configuration

Table 1 summarizes the LVLMs information used in this paper, including the corresponding parameter sizes, visual encoders, and LLMs.

Table 1: Model architecture of 19 LVLMs evaluated on MMT-Bench.

Models	Parameters	Vision Encoder	LLM
GPT-4V [42]	-	-	-
Claude [36]	-	-	-
GeminiProVision [37]	-	-	-
Reka Flash [38]	-	-	-
ShareGPT4V-7B [5]	7.2B	CLIP ViT-L/14	Vicuna-v1.5-7B
ShareGPT4V-13B [5]	13.2B	CLIP ViT-L/14	Vicuna-v1.5-13B
LLaVA-v1.5-7B [28, 30]	7.2B	CLIP ViT-L/14	Vicuna-v1.5-7B
LLaVA-v1.5-13B [28, 30]	13.4B	CLIP ViT-L/14	Vicuna-v1.5-13B
XComposer [50]	8B	EVA-CLIP-G	InternLM-7B
XComposer2 [11]	7B	CLIP ViT-L/14	InternLM2-7B
mPLUG-Owl2 [44]	8.2B	CLIP ViT-L/14	LLaMA2-7B
QWenVL [2]	9.6B	CLIP ViT-G/16	QWen-7B
LLaMA-Adapter-v2 [14]	7B	CLIP-ViT-L/14	LLaMA-7B
BLIP2-Flan-T5-XL [22]	12.1B	EVA-CLIP ViT-G/14	Flan-T5-XL
BLIP2-Flan-T5-XXL [22]	12.1B	EVA-CLIP ViT-G/14	Flan-T5-XXL
InternVL-Chat-V1.2 [8]	40B	InternViT-6B	Nous-Hermes-2-Yi-34B
Monkey [25]	9.8B	CLIP-ViT-BigHuge	Qwen-7B
MiniGPT-4 [53]	8.0B	EVA-G	Vicuna-7B
MMAlaya [32]	7.8B	BLIP2-opt-2.7b	Alaya-7B-Chat
Otter [19]	1.3B	CLIP ViT-L/14	LLaMA-7B

B Task Category

	Task Category
Perception	"Car Recognition","OCR","Role Identification","Celebrity Recognition","OCR Math","Material Recognition","Sign Description","Furniture Recognition","Location Recognition","Board Chess Position Detection","Painting Recognition","Device Recognition","Movie Recognition","Structure Recognition","Tangram Description","House Plan Recognition","Human Description","Flower Recognition","Outfit Recogntion","Graph Description","Statue Recognition","Product Recognition","Image Description","Item Recognition","Handwork Recognition","Profession Identification","Food Recognition","Behavior Recognition","Color Recognition","Scene Description","Expression Recognition","Chemical Identification","Object Counting","Medicine Recognition","Shape Recognition","Plant Identification","Item Recogntion","Animal Recognition","Medical Recognition","Photo Recognition","Landmark Recognition","Length Estimate","Chart Description","Posture Description","Gestures Recognition","Aircraft Recognition","Traffic Sign identification","Event Recognition","Injury Description","Hazard Identification","Emotion Conditioned Output","Exercise Recognition","Geometry Problems Description","Astronomy Identification","Weather Recogntion","Game Recognition","Meter Reading","Recipe Description","Food Chain Description","Food Description","Position Description","Historical Relic Identification","Art Work Description","Sculpture Description","Logo Recognition","Make-up Description","Spatial Relationship Understanding","Weather Recognition","Attire Recognition","Differently Abled Recognition","Capacity Estimate"
Reasoning	"In-context Visual Scenace Understanding","Meme Reasoning","Math Reasoning","Structure Understanding","Traffic Sign Reasoning","Furniture Understanding","Location Understanding","Board Game Reasoning","Art Knowledge Reasoning","Device Reasoning","Dressing Reasoning","Tangram Speculation","Anagrams Reasoning","House Plan Understanding","Question Generation","Visual Commonsense Reasoning","Flower Understanding","Pop Culture Reasoning","Exercise Reasoning","Celebrity Understanding","Product Instruction","Figurative Speech Explanation","Paper Folding Reasoning","History Knowledge Reasoning","Physical Knowledge Reasoning","Cultural Knowledge Reasoning","Climate and Weather Understanding","Food Reasoning","Human Emotion Reasoning","Chemical Knowledge Reasoning","Biology Knowledge Reasoning","Medical Reasoning","Count Reasoning","Rhetoric Reasoning","Plant Identification Reasoning", "Chart Reasoning","Counterfactual Examples","In-context Visual Scene Understanding","Flavor Reasoning","Capacity Estimate Reasoning","Color Reasoning","Gestures Understanding","Location Relative Position","Contextual Knowledge of Events","Word Translation","Rational Action Identification","Geography Reasoning","Hazard Reasoning","Physical Knowledge Reasoning ","Position Reasoning","Graph Reasoning","Sign Understanding","Material Reasoning","Make-up Reasoning","Object Counting Reasoning","Sport Level Reasoning","Astronomy Reasoning","OCR Math Reasoning","Age Reasoning","Damage Evaluation Reasoning","Abstract Reasoning","Music Reasoning","Environment Reasoning","Role Identification Reasoning"
Creation	"Slogan Generation","Caption Generation","Math Computing","Building Materials Plan","Home Renovation Plan","Blog Writing","Algorithm Design","Artistic Appreciation","Device Principle Explanation","Movie Synopsis Writing","Travel Plan Writing","Tangram Segmentation","Computer Programming","Physical Computing","Advertisement Writing","Legalization Discussion","Chemistry Discussion","Roleplay","Plant Growing Plan","Exercise Plan","Place Recommendation","Dialogue Generation","Computer Knowledge","Essay Writing","How Visual Content Arouses Emotions","Diagram Generation","Poem Writing","Painting Drawing Teaching","Prompt Generation for Image Generation","News Report Writing","Temporal Anticipation","Multilingual Multicultural Understanding","Appliance Evaluation Report","Career Plan Generation","Treatment Plan","Catchy Titles Generation","Prompt Generation for Image Edition","Device Instructions Teaching","Calorie Estimate","Recipe Writting","Computer Program Description","Navigation","Science Fiction Scene Writing","Medical Suggestion","Humanity Discussion","Math Funtion Graphing","Story Writing","Metaphor Writing","Constrained Prompting","Photography Plan","Customized Captioner","Makeup Design","Chemical Computing","Powerpoint Production","Self-driving Design","Exercise Promotional Article Writing","Animal Growing Plan","Event Infulence and Meaning Discussion","Lyric Writing","Movie Review Writing","Packing Plan","Rubik's Cube Solution","Repairment Plan","Insurance Report Generation","Chemical Knowledge Computing","Mathematical Proof","Activity Recommendation","Biology Discussion","Geography Discussion","Meme Writing","Clothing Recommendation","Hairstyle Design","Planing","Nail Art Design","Damage Evaluation","Food Chain Computing","Architectural Plan","Metaphor Writting","Legal Rights Protection"

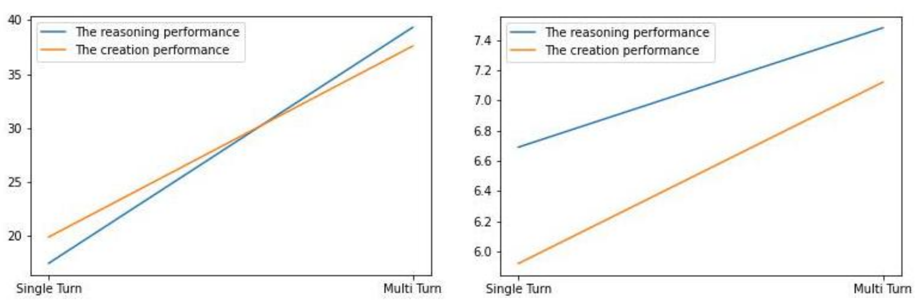
Fig. 1: List of exisiting task categories in ConvBench.

C Additional Experimental Results

C.1 Chain of Thoughts

ConvBench can also be employed to examine studies on chain-of-thought reasoning. As illustrated in Figure 2, the single-turn approach, which involves directly requesting reasoning or creation instructions, yields inferior results compared to the multi-turn method. Figure 3 is an example in ConvBench which demonstrates how a multi-turn approach can bolster reasoning capabilities. Several studies have introduced techniques to enhance this performance. For instance, IdealGPT [47] generates sub-questions derived from the main question and responds with the corresponding sub-answers. It then analyzes the aggregate information from these sub-answers to deduce the most likely answer to the main question. Its experimental results also indicate that the chain-of-thought framework's performance surpasses that of Large Vocabulary Language Models (LVLMs) in a zero-shot context.

However, it is indeed worthwhile to investigate whether the decomposed sub-questions can contribute to solving the main question. ConvBench provides annotations to test the effectiveness of these decomposed sub-questions in answering the main question.



(a) Performance comparison between single-turn and multi-turn for GPT-4V with Pairwise Grading. (b) Performance comparison between single-turn and multi-turn for GPT-4V with Direct Grading.

Fig. 2: The performance comparison between single-turn and multi-turn manner for GPT-4V, using both Pairwise Grading method and Direct Grading method.

C.2 Ablation Studies

Are Focus Demands neccessary?

The creativity instruction presents an open-ended question that lacks a standard answer. We identify and annotate specific focal points related to each creative prompt to guide the evaluation process. To examine the effectiveness of these focal points, we conduct an experiment involving 50 creative instances.

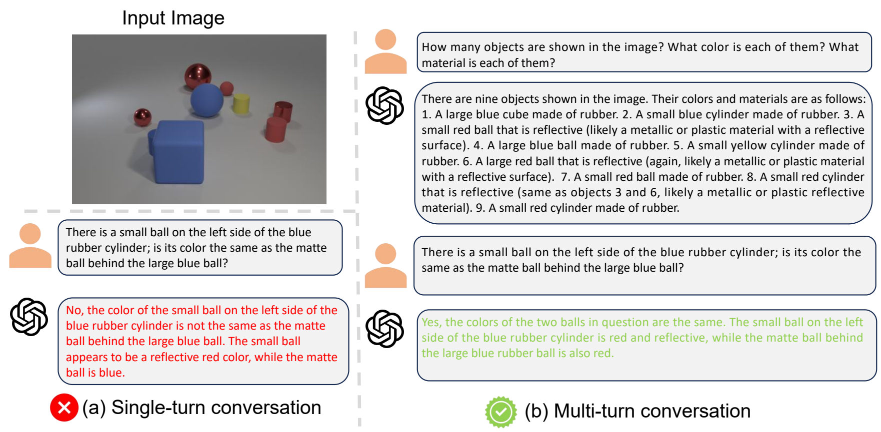


Fig. 3: This example which demonstrates that a multi-turn approach can bolster reasoning capabilities.

The experiments are divided into two groups: one that does not incorporate focal points in creation evaluations and the other that does. All other variables between the groups are kept constant. We manually assessed the evaluation outcomes from both groups to determine which group of results appeared more logical. A comparative win-rate, based on the quality of evaluation outcomes for identical responses, is calculated for both groups. The findings indicated success rates of 56% for the group using focal points and 44% for the group without. Figure 5 provides an illustrative example of the focal points' effectiveness.

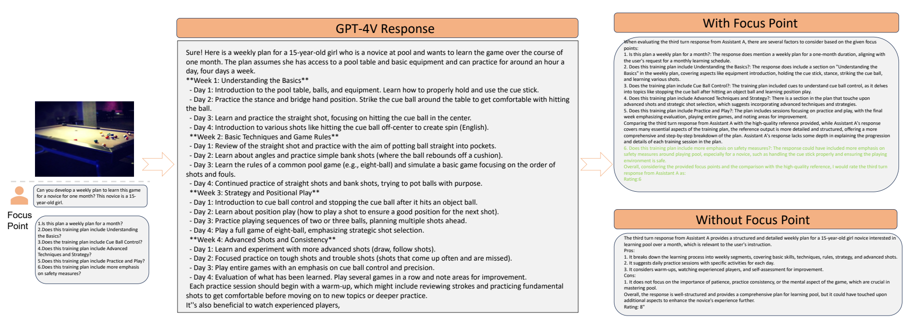


Fig. 4: This example assesses the effectiveness of focus points. Provided with an image and instructions, GPT-4V generates responses. The evaluation outcome incorporating focus points is deemed more appropriate compared to the result without focus points.

Are the three turns' judgements helpful for the overall conversation evaluation? We feed the perception, reasoning and creation judgements into the final overall conversation evaluation. To examine their necessities, we conduct

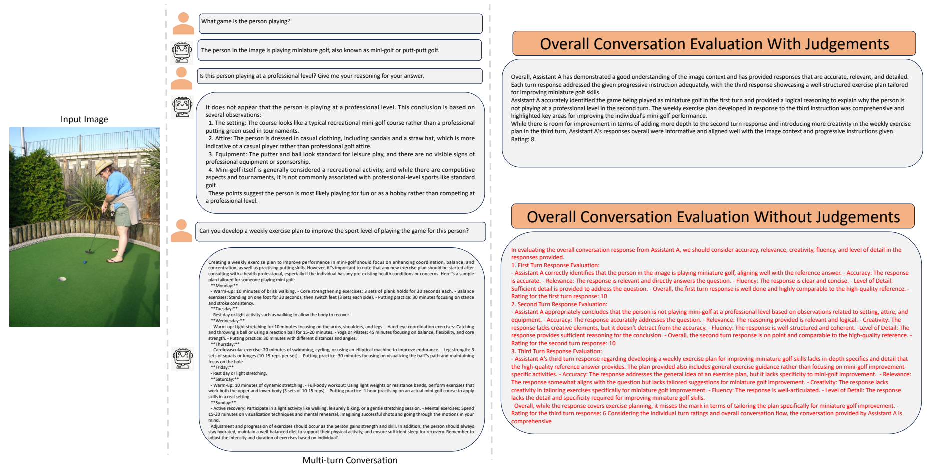


Fig. 5: This example evaluates the influence of judgments made across three turns on the entire conversational process. In this case, evaluating the overall conversation without incorporating these judgments fails to yield a definitive overall rating for the overall conversation.

an experiment involving 50 instances. The experiments are divided into two groups: one that does not involve three turns' judgements in overall conversation evaluations and the other that does. All other variables between the groups are kept constant. We manually assessed the evaluation outcomes from both groups to determine which group of results appeared more logical. A comparative win-rate, based on the quality of evaluation outcomes for identical responses, is calculated for both groups. The findings indicated success rates of 60% for the group using the three turns' judgements and 40% for the group without. The method with these judgements provides ultimate rating judgements for all examples, however, 8% fail to provide ultimate rating judgement if without these judgements. Figure 5 provides an illustrative example according to the necessities of three turns' judgements.

D Case Study

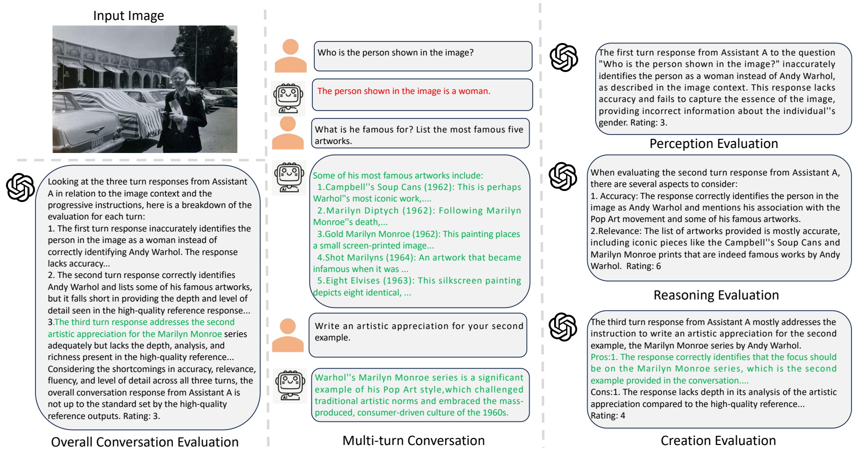


Fig. 6: This example demonstrates that the evaluation process encompasses not just the assessment of individual turns, but also the overall conversation. Specifically, it examines whether the LVLM precisely chooses the second example from the previous responses when addressing the instructions in the third turn.

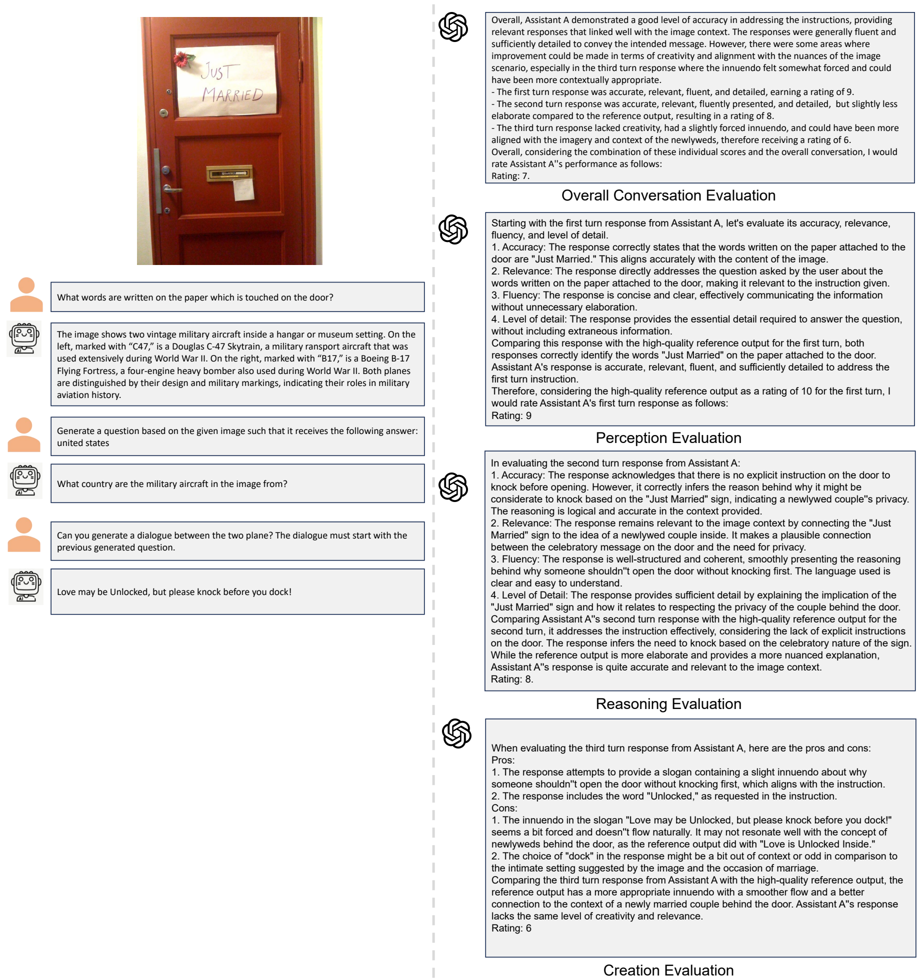


Fig. 7: This example precisely interprets the text on the door without impairing further reasoning. In assessing creative performance, it is determined that the response is deficient in creativity. The performances across the three turns are considered to evaluate the multi-turn conversation's overall performance.

What items are positioned against the step to two retial dorrways?

There are two large wedge-shaped items with the text ''STOP N'' DROP'' on them. The one in the foreground is red, and the other is green. Both are placed flat on the ground against a step leading to two separate retail doorways. This placement is typically used to prevent people from slipping as they exit, especially when surfaces may be wet or icy. In terms of object classification using COCO categories: these could be classified under ''Miscellaneous Objects'', more specifically as a type of ''Safety Sign''. However, they don''t neatly fit into any predefined COCO category like furniture or sports equipment due to their unique function and design.

Would a disabled wheelchair bound individual find it easy to go into this store?

It appears that the store owners are taking consideration for wheelchair users by installing a wide ramp in front of the entrance.
However,without seeing inside it's hard to assess if the path through the door is clear and there aren't any steps or other obstructions. Even though this particular obstacle with 'STOP N DROP' wedge is helpful for customers entering when surfaces may be wet or icy, they must also find a way through and around them before being able to enter at all.
It seems like they have not completely overcome an obvious barier preventing free access - suggesting unobserved difficulties might exist once attempting to progress past these wedges into store.

If the total weight of the wheelchair and the disabled person is 700N, if you want to push the disabled person and the wheelchair up a 0.02-meter-high threshold along a 1-meter portable wedge ramp in 2 seconds, and the thrust used is 150N, what is the power of the push? ? What is the mechanical efficiency of the portable wedge ramp in this process? What is the friction force exerted on the wheelchair by the portable wedge ramp? (Friction between internal parts of the wheelchair is ignored)

Let's break this down step by step:\n\n1. Power of the Push: Power is calculated as force multiplied by velocity. However, since the ramp is inclined and work needs to be done against gravity, we use a component of force (cos(theta)*F), where theta is 90 - the angle with vertical which is cos(0) =1.
Mechanical Efficiency (\u03B7):
\u03B7 = useful output / input energy
useful output = W * h = m*g*h * y,\nwhere W was weight on ramp, g=9.81m/s^2, m=700N,(h+y)=0.02-0.01m from threshold to ground.\ninput energy would be Work Done up Ramp or 150N*1m.\
So \u03B7=(700g*0)-150/100\n\nFrictional Force exerted:\nFirst calculate normal force N which is equal in magnitude but opposite in direction to gravitational force: Gf=m*g*sin(theta), so sin(\u03B8) must be found using trigonometry --> tan(\u03B8)=opposite/adjacent <= height/ramp-length\nAfter calculating \u03B8 using arctan(h/lengt). Then \nFG=N*cos(\u03B8)\nwhich gives you frictional forces

Overall, the conversation response from Assistant A has several shortcomings in accuracy, relevance, detail, and completeness throughout the progressive urns.
1. The first turn response accurately identifies the items as ramps, but lacks the specific details mentioned in the high-quality reference, thus falling short of providing a comprehensive answer. The response is mostly relevant and fluent, but it could benefit from more specific details matching the reference better.
2. The second turn response lacks detail in explaining how the presence of the ramps facilitates ease of access for wheelchair-bound individuals, failing to elaborate on the importance of the ramps in accessibility.
3. The third turn response is particularly problematic as it incorrectly calculates the thrust power, mechanical power, and friction force, deviating significantly from the accurate calculations and explanations provided in the high-quality reference.
Considering the overall conversation, Assistant A''s responses demonstrate a lack of precision, detail, and accuracy required to effectively address the progressive instructions given in the context of the image. The responses fall short of the high-quality references on multiple levels in terms of accuracy, relevance, and detail.
Rating: 3.

Overall Conversation Evaluation

The first turn response from Assistant A identifies the items as "Two green and red wooden ramps" positioned against the step to two retail doorways. This response accurately conveys the presence of the ramps but lacks the specific details mentioned in the high-quality reference, which mentions the ramps as portable wedge ramps with the brand "STOPGAP.CA" printed on them. Hence, this response could be more detailed and specific to match the high-quality reference better.
In terms of relevance, the response directly addresses the question posed regarding the items against the doorways, so it is relevant to the instruction given.
Fluency-wise, the response is clear and concise, presenting the information in a straightforward manner without any major grammatical errors or confusion.
Considering the level of detail provided in the response, it falls short of the level expected in comparison to the high-quality reference. The reference includes additional details about the purpose of the ramps in bridging the vertical distance and providing easier access for individuals with mobility devices, which could have been beneficial to include in the response for a more comprehensive answer.
Comparing the first turn response from Assistant A with the first turn high-quality reference that is rated as 10, the response could be improved by adding more specific details about the ramps, such as their portability, wedge shape, and the "STOPGAP.CA" text on them.
Therefore, I would rate the first turn response from Assistant A as follows:
Rating: 6

Perception Evaluation

When evaluating the second turn response from Assistant A:
- The response lacks detail in explaining how the presence of the ramps contributes to making it easier for a disabled wheelchair-bound individual to enter the store.
- It could have elaborated on how the ramps bridge the vertical distance between the sidewalk and the doorways, providing a smooth transition for wheelchair users.
- The answer is quite brief and lacks a thorough explanation of why a disabled wheelchair-bound individual would find it easy to enter the store due to the ramps present.
Comparing this with the high-quality reference, which provides a detailed explanation of how the ramps help overcome the step at the entrance and make the store more accessible for individuals with mobility challenges, Assistant A''s response falls short in terms of detail, relevance, and completeness.
Rating: 4

Reasoning Evaluation

Analyzing the third turn response from Assistant A step by step reveals several issues:
1. In terms of the thrust power calculation, the response states the mechanical efficiency of the portable wedge ramp instead of calculating the thrust power correctly like the reference output.
2. The response does not calculate the mechanical power of the wheelchair as instructed, providing an incorrect value instead of the proper calculation.
3. The stated friction force exerted on the wheelchair by the portable wedge ramp is significantly inaccurate, deviating from the correct value provided in the reference output.
Considering the focus points, Assistant A''s response falls short in all three aspects. The calculations are incorrect, leading to inaccurate results compared to the reference output. The response lacks the necessary detail, precision, and accuracy required to address the third turn instruction effectively.
Rating: 2

Creation Evaluation

Fig. 8: The accurate outcomes from the third turn include a push power of 75W, a mechanical power efficiency for the wheelchair of 9.33%, and a frictional force exerted by the wheelchair on the inclined surface of 136N. This example highlights the ongoing challenges faced by LVLMs in scientific domains.

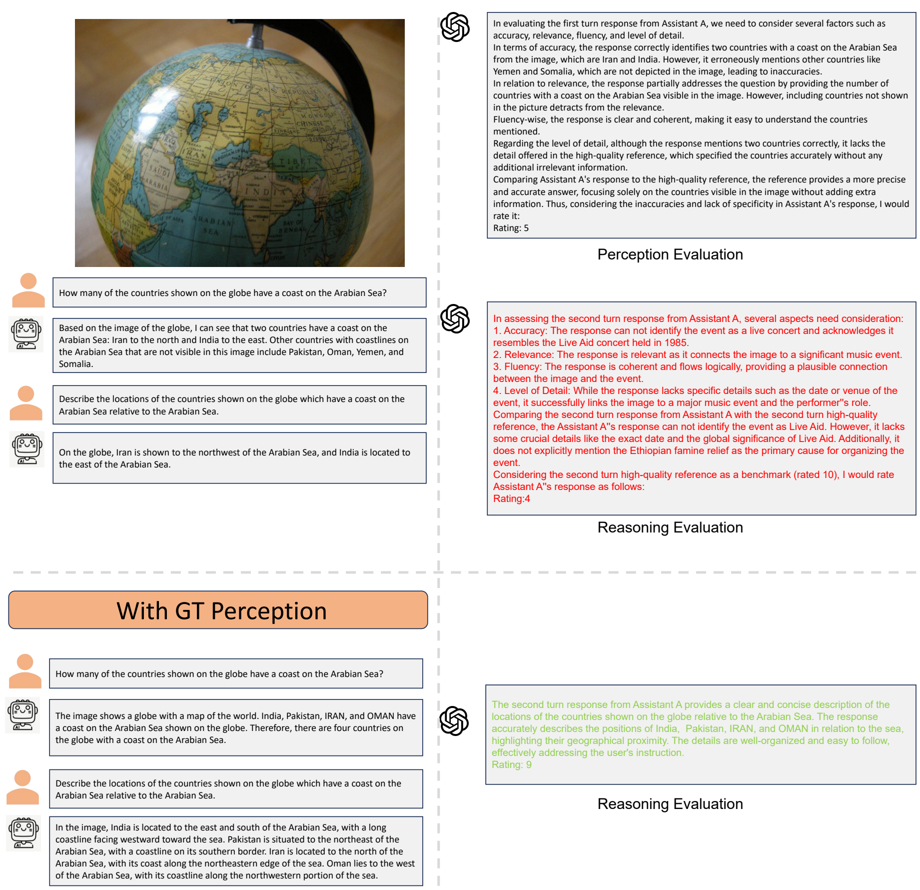


Fig. 9: This example fails to offer a clear understanding of the countries surrounding the Arabian Sea, which impacts subsequent reasoning. However, when a clear perception is provided, the reasoning process is accurate. This indicates that inadequate perception can adversely affect reasoning.

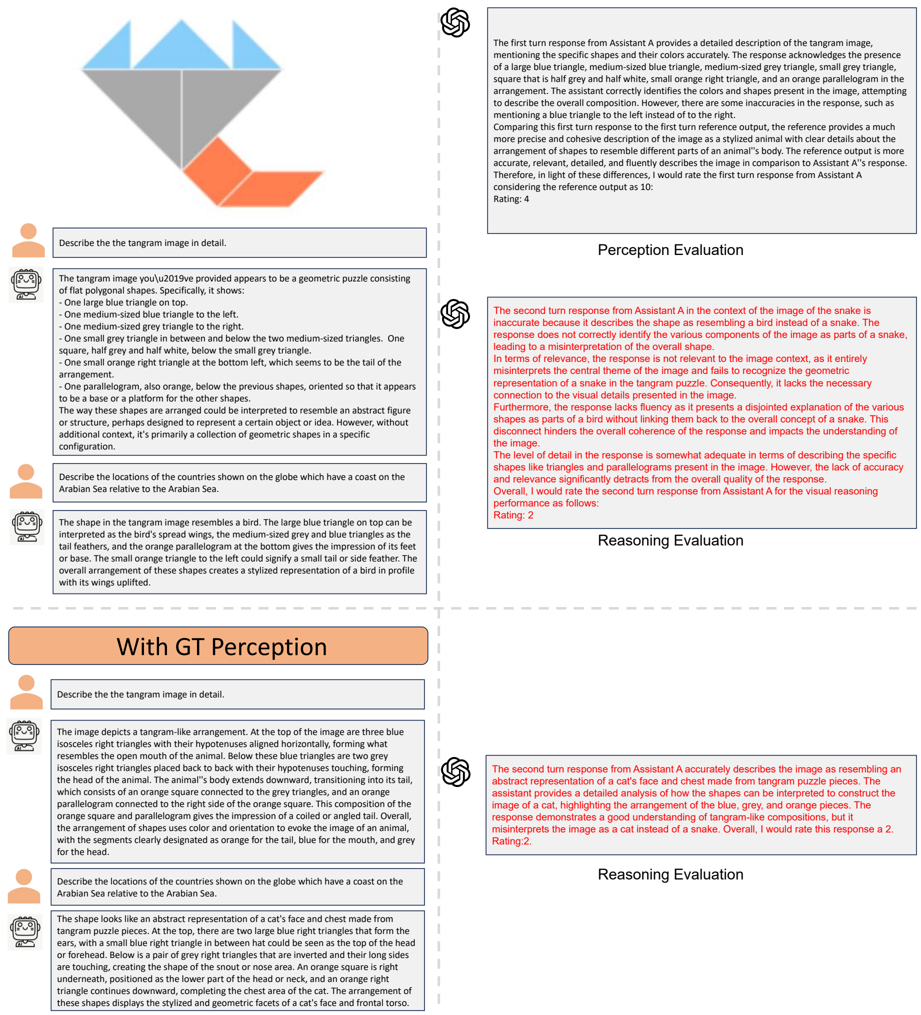


Fig. 10: In this example, reasoning errors stem from the inability to accurately perceive each tangram and understand their positional relationships to one another. However, even with the provision of detailed image descriptions, the reasoning outcomes still contain errors.

E Prompt Templates

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Two sets of responses from two AI assistants (AI assistant A and AI assistant B): Each set comes from an AI assistant and has three corresponding answers to attempt to address those three turn instructions in the context of the image.

Your job is to judge whether the first turn response from Assistant A or the first turn response from Assitant B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You should choose the assistant the follows the user's first instruction and answers the user's first question better.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency etc.
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.
- You think step-by-step, but ultimately respond with "Response A" or "Response B".

user (human authored)

I will describe an image to you, and provide three progressive instructions. Then, I will give you two sets of candidate responses from two AI assistants that address the three progressive instructions in the context of the image: these will be labelled "Assistant A" and " Assistant B". Your job is to first reason step-by-step about which response for the first turn instruction is better in terms of accuracy, relevance, fluency, etc. After reasoning step-by-step and comparing the pros/cons of corresponding responses for the first turn instruction, in the end, respond with "Overall, Response X is better." where X is either A or B.

assistant (human authored)

Sure, please provide the image context, the three progressive instructions, and the two sets of candidate responses. Then, I will think step-by-step and provide my ultimate judgement as to which response for the first turn instruction is better.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: This is an image of a straight road in a desert-like landscape. The road is marked with double yellow lines, indicating a no-passing zone. The asphalt appears slightly worn but in generally good condition. On either side of the road, there is dry, sandy soil with sporadic, drought-resistant vegetation like small shrubs and grasses. ..

<|The Start of Assistant A' s Conversation with User|>
The first turn question from user:
What is the landscape and climate of this location like shown in the image?

The first turn response from Assistant A:
The landscape in the image appears to be typical of a desert or arid region. You can infer the climate might be hot and dry based on the dry, sandy soil and presence of drought-resistant vegetation such as small shrubs and grasses that are adapted to such environments...

The second turn question from user:
If you are driving through this area, what are vital supplies you will need? Why are these particularly important in this location?

The second turn response from Assistant A:
When driving through a desert or arid landscape as show in the image, it is essential to carry the following supplies:
1. Water: This is the most critical supply given the high temperatures and dry conditions, which can lead to dehydration. Enough water should be carried to last the duration of the journey, plus some extra in case of emergencies...

The third question from user:
According to previous responses, can you give me a three-day and two-night camping plan for three men and two women in this location from March 10th to March 12th, 2024.The budget is now set at $2,000.

The third turn response from Assistant A:
Creating a camping plan for a group of five in a desert location with a budget of $2,000 over three days and two nights would involve careful planning to ensure that the group has a safe and enjoyable experience. Here\u2019s a suggested itinerary: Pre-Trip Preparations:\nGear: Rent high-quality camping gear suitable for desert conditions, including tents, sleeping bags, portable stove, cooking equipment, and a cooler...

<|The End of Assistant A' s Conversation with User|>

<|The Start of Assistant B' s Conversation with User|>
The first turn question from user:
What is the landscape and climate of this location like shown in the image?

The first turn response from Assistant B:
The image depicts a desert-like landscape with a straight road cutting through it. The terrain is relatively flat with sparse vegetation, consisting primarily of small shrubs and bushes typical of arid environments...

The second turn question from user:
If you are driving through this area, what are vital supplies you will need? Why are these particularly important in this location?

The second turn response from Assistant B:
When driving through such a desert area, some vital supplies you will need include:
1. Water: This is the most crucial supply as deserts are extremely dry and can have very high temperatures, leading to rapid dehydration...

###The third turn question from user:
According to previous responses, can you give me a three-day and two-night camping plan for three men and two women in this location from March 10th to March 12th, 2024.The budget is now set at $2,000.

The third turn response from Assistant B:
With a budget of $2,000 for a three-day, two-night desert camping trip for five people, here's a suggested plan:
Day 1: Arrival (March 10, 2024) Morning:- Depart early morning to maximize daylight hours.- Pick up a rental 4WD vehicle if necessary (approx. $150/day). - On the way, buy supplies (food, water, fuel, etc.) - Allocate around $300 for groceries including water (approx. 1 gallon per person/day) and non-perishables. Afternoon:- Arrive at the designated camping site by early afternoon.- Set up camp with rented/bought tents and sleeping gear...

<|The End of Assistant B' s Conversation with User|>

Think step-by-step, compare the first turn responses from the two assistants, and finish your response with "Overall, Response X is better." where X is either A or B.'

assistant (machine generated; used to derive the prediction)

Based on the detailed analysis provided for the first turn responses from Assistant A and Assistant B, it is clear that Assistant A"s response is better. Assistant A not only accurately describes the landscape and climate of the location shown in the image but also provides a thorough explanation of the key elements depicted in the desert-like landscape. The response from Assistant A delves into the presence of desert vegetation, the condition of the road, the surrounding mountains, and the clear sky, painting a vivid picture of a hot and dry climate in a remote, isolated environment.
Therefore, for the first turn instruction, **Overall, Response A is better.

Fig. 11: The prompt used for evaluating perception turn in a pairwise grading method, accompanied by a sample completion from ChatGPT-3.5, is provided. The response conversations are anonymously presented to ChatGPT-3.5, where Assistant A is identified as a human, and Assistant B is recognized as GPT-4V.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Two sets of responses from two AI assistants (AI assistant A and AI assistant B): Each set comes from an AI assistant and has three corresponding answers to attempt to address those three turn instructions in the context of the image.

Your job is to judge whether the second turn response from Assistant A or the second turn response from Assitant B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You should choose the assistant the follows the user's second instruction and answers the user's second question better.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.
- You think step-by-step, but ultimately respond with "Response A" or "Response B.

user (human authored)

I will describe an image to you, and provide three progressive instructions. Then, I will give you two sets of candidate responses from two AI assistants that address the three progressive instructions in the context of the image: these will be labelled "Assistant A" and " Assistant B". Your job is to first reason step-by-step about which response for the second turn instruction is better in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of corresponding responses for the second turn instruction, in the end, respond with "Overall, Response X is better." where X is either A or B.

assistant (human authored)

Sure, please provide the image context, the three progressive instructions, and the two sets of candidate responses. Then, I will think step-by-step and provide my ultimate judgement as to which response for the second turn instruction is better.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: There is portrait of a Native American named Geronimo. He is wearing a blue coat and a red shirt. He is also wearing a yellow tie. There is a old man in the painting. He has short ear-length hair. He is a native American. His skin is winked. He wears a red shirt inside and a blue coat outside. He is wearing a yellow scarf. "GERONIMO APACHE" and "COPYRIGHT 1898 F.A. RINEHART OMAHA" written on the painting.

<|The Start of Assistant A’ s Conversation with User|>
The first turn question from user:
Who is this man shown in the image?

The first turn response from Assistant A:
The man shown in this image represents Geronimo, who was a prominent leader and medicine man from the Apache tribe.

The second turn question from user:
Who did this fearless man stand up to? How do we remember him today?

The second turn response from Assistant A:
Geronimo, a fearless Native American leader,_ stood up against both Mexican and United States forces that encroached on the lands and rights of his Apache people. He fought to protect his people's way of life, culture, and land. Today, we remember him as a symbol of resistance and a defender of indigenous culture. He is remembered today as a symbol of Native American resistance and a defender of indigenous rights...

The third question from user:
Write an essay for an assessment of this historical figure.

The third turn response from Assistant A:
Title: The Dialectical Evaluation of Geronimo: Resistance and Reflection in Historical Context
Introduction:
In the annals of history, few figures have embodied th spirit of resistance and the fight for cultural survival as Geronimo, the renowned Apache leader...
Thesis Statement:
Through the dialectic process, one may gain a profound understanding of Geronimo''s actions, motives, and legacy within the larger historical narrative...
Affirmation:
Geronimo''s legacy is often celebrated for his unyielding determination to resist the encroachment of foreign forces on Apache lands. He led a life committed to the...
<|The End of Assistant A’ s Conversation with User|>

<|The Start of Assistant B’ s Conversation with User|>
The first turn question from user:
Who is this man shown in the image?

The first turn response from Assistant B:
Sorry, I can't provide the identity of real people in images.

The second turn question from user:
Who did this fearless man stand up to? How do we remember him today?

The second turn response from Assistant B:
While I can''t identify the person in the image, I can provide generic historical information that might apply to an individual resembling the description of a "fearless man" from Native American history. There were numerous Native American leaders who stood up to various adversaries, including European-American settlers and the United States government during the 19th and early 20th centuries...

###The third turn question from user:
Write an essay for an assessment of this historical figure.

The third turn response from Assistant B:
With a budget of $2,000 for a three-day, two-night desert camping trip for five people, here's a suggested plan:
Title: A Legacy of Resistance: Assessing the Impact of a Native American Leader\n\nIntroduction:
The American frontier of the 19th century was a battleground of cultures, ideologies, and survival. Among the most notable figures of this era was a legendary Apache leader\u2014a man whose name became synonymous with Native American resistance against U.S...

<|The End of Assistant B’ s Conversation with User|>

Think step-by-step, compare the second turn responses from the two assistants, and finish your response with "Overall, Response X is better." where X is either A or B.'

assistant (machine generated; used to derive the prediction)

Assistant A"s response for the second turn question provides a detailed and specific overview of who Geronimo stood up against and how he is remembered today...
On the other hand, Assistant B"s response for the second turn question takes a more generic approach, not confirming the identity of the individual in the image and instead providing historical information about Native American leaders in general...
Comparing the second turn responses, Assistant A"s response is more accurate, specific, and directly relevant to the image context and user"s question. Therefore, in terms of accuracy, specificity, and relevance to the given image description, Assistant A"s response for the second turn instruction is better.

Overall, Response A is better.

Fig. 12: The prompt used for evaluating reasoning turn in a pairwise grading method, accompanied by a sample completion from ChatGPT-3.5, is provided. The response conversations are anonymously presented to ChatGPT-3.5, where Assistant A is identified as a human, and Assistant B is recognized as GPT-4V.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Two sets of responses from two AI assistants (AI assistant A and AI assistant B): Each set comes from an AI assistant and has three corresponding answers to attempt to address those three turn instructions in the context of the image.
4. Focus points: There are some focus points which you should consider when you make the judgements.

Your job is to judge whether the third turn response from Assistant A or the third turn response from Assitant B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You should choose the assistant the follows the user's third instruction and answers the user's third question better.
- You are capable of judging response quality, accounting for important factors like correctness, relevance, fluency, specificity, etc.
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.
- You think step-by-step, but ultimately respond with "Response A" or "Response B- You think step-by-step, but ultimately respond with "Response A" or "Response B.

user (human authored)

I will describe an image to you, and provide three progressive instructions. Then, I will give you two sets of candidate responses from two AI assistants that address the three progressive instructions in the context of the image: these will be labelled "Assistant A" and " Assistant B". Your job is to first reason step-by-step about which response for the third turn instruction is better in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of corresponding responses for the third turn instruction, in the end, respond with "Overall, Response X is better." where X is either A or B.

assistant (human authored)

Sure, please provide the image context, the three progressive instructions, and the two sets of candidate responses. Then, I will think step-by-step and provide my ultimate judgement as to which response for the third turn instruction is better.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: In this photo, the sun is setting which casts an orange huge against the black tree-line. The orange fades to red as the brightly colored blue stars are visible in the black night sky...

<|The Start of Assistant A' s Conversation with User|>
The first turn question from user:
What are the bright dots? What is the red hue?

The first turn response from Assistant A:
The bright dots in the image are stars scattered across the night sky. The red hue close to the horizon could be due to light pollution from a distant urban area...

The second turn question from user:
Why does it look like that, what causes the bright dots, what causes the red hue?

The second turn response from Assistant A:
The bright dots in the image are caused by the light from distant stars reaching the camera sensor or film...

The third question from user:
Write a romantic poem for explaining the previous causes.

The third turn response from Assistant A:
Beneath the celestial tapestry so wide and deep,
Stars whispersecrets that the night does keep.
Each a tiny heartbeat in the cosmos' breast,
Light-years filled with love's soft cadence, unrest...
<|The End of Assistant A' s Conversation with User|>

<|The Start of Assistant B' s Conversation with User|>
The first turn question from user:
What are the bright dots? What is the red hue?

The first turn response from Assistant B:
The bright dots are stars. The red hue is the afterglow of the setting sun. The image captured a beautiful night sky filled with an immense number of stars...

The second turn question from user:
Why does it look like that, what causes the bright dots, what causes the red hue?

The second turn response from Assistant B:
The image looks like that due to the scattering of sunlight by the Earth's atmosphere during sunset, which causes the orange and red hues...

###The third turn question from user:
Write a romantic poem for explaining the previous causes.

The third turn response from Assistant B:
In the tender clutch of evening''s soft descent,
Where the sun kisses the horizon, and day is spent,
The sky, a canvas vast, begins its nightly hue,
A ballet of light where orange fades to red anew...
<|The End of Assistant B' s Conversation with User|>

There are some concerns which you should focus when make your judgements for the response:
1. Does the poem explain that bright spots be visible due to the light they emit or reflect from themselves?
2. Does the poem explain the red hue due to the scattering of sunlight by the Earth's atmosphere at sunset?
3. Is this poem romantic?'

Think step-by-step, compare the third turn responses from the two assistants, and finish your response with "Overall, Response X is better." where X is either A or B.'

assistant (machine generated; used to derive the prediction)

When comparing the third turn responses from Assistant A and Assistant B, let's consider the following points:
1. **Accuracy and Relevance**:
- Assistant A's response beautifully integrates the concepts of stars, their distant nature...
Assistant B's response also includes romantic elements and metaphorical language to describe the stars...
2. **Fluency and Poetic Quality**:
- Assistant A's poem flows well, utilizing poetic language effectively...
- Assistant B's poem also maintains a poetic tone..
3. **Tie Back to Previous Explanations**:
- Both responses successfully tie the romantic poem back to the scientific explanations provided earlier about stars, light pollution, and the red hue, offering a creative and artistic interpretation of those concepts.\n\nBased on the above considerations,
Overall, Response A is better. Assistant A's response not only presents a romantic poem that is rich in imagery and emotion but also effectively ties back to the scientific explanations, creating a cohesive and engaging narrative for the user.

Fig. 13: The prompt used for evaluating creation turn in a pairwise grading method, accompanied by a sample completion from ChatGPT-3.5, is provided. The response conversations are anonymously presented to ChatGPT-3.5, where Assistant A is identified as GPT-4V, and Assistant B is recognized as a human.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging the multi-turn conversation instruction-following ability of an AI assistant. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instruction.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Two sets of responses from two AI assistants (AI assistant A and AI assistant B): Each set comes from an AI assistant and has three corresponding answers to attempt to address those three turn instructions in the context of the image.
4. Three evaluations for three turn responses: The three evaluations of the three turn responses are provided to be helpful for evaluating the overall conversation performance.

Your job is to judge whether the overall conversation from Assistant A or the overall conversation from Assitant B better. A and B are randomly ordered.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- You should choose the assistant the follows the user's instructions and answers the user's questions better.
- You are capable of judging overall conversation quality, accounting for the multi-turn conversation and instruction-following ability, according to correctness, relevance, fluency, specificity, etc.
- Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants. Be as objective as possible.
- You think step-by-step, but ultimately respond with "Response A" or "Response B

user (human authored)

I will describe an image to you, and provide three progressive instructions. Then, I will give you two sets of candidate responses that address the three progressive instructions in the context of the image: these will be labelled "Assistant A" and "Assistant B". Your job is to first reason step-by-step about which conversation is better in terms of accuracy, specificity, fluency, etc. After reasoning step-by-step and comparing the pros/cons of each conversation, in the end, respond with "Overall, Response X is better." where X is either A or B.

assistant (human authored)

Sure, please provide the image context, the three progressive instructions, and the two sets of candidate responses. Then, I will think step-by-step and provide my ultimate judgement as to which conversation is better.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: There are three rows of kettlebells, with six kettlebells in each row. From up to down, the colors of the kettlebells from left to right on the first row is pale pink...

<|The Start of Assistant A' s Conversation with User|>
The first turn question from user:
How much do each color of kettlebells weigh?

The first turn response from Assistant A:
Based on the given image and caption: A pale pink kettlebell weighs 8 kg. A light blue kettlebell weighs 10 kg. An orange kettlebell weighs 16 kg...

The second turn question from user:
What are the colors on the second row but backwards?

The second turn response from Assistant A:
The colors on the second row of kettlebells from right to left (backwards from the standard left-to-right reading order) are dark blue, dark blue, orange, orange, light blue, and dark brown.

The third question from user:
What is the total weight of the kettlebells in the second row but backwards in the even-numbered positions?

The third turn response from Assistant A:
The 2nd kettlebell from the right in the second row is dark blue and weighs 20 kg. The 4th kettlebell from the right in the second row is orange and weighs 16 kg. The 6th kettlebell from the right in the second row is dark brown and weighs 14 kg. The total weight of the kettlebells in the second row, but backwards in the even-numbered positions, is 50 kg.

<|The End of Assistant A' s Conversation with User|>

<|The Start of Assistant B' s Conversation with User|>
The first turn question from user:
How much do each color of kettlebells weigh?

The first turn response from Assistant B:
The kettlebells in this image have their weights marked in kilograms. Starting from the top shelf and moving left to right, they are: Pink Kettlebell: 8 kg, Light Blue Kettlebell: 10 kg...

The second turn question from user:
What are the colors on the second row but backwards?

The second turn response from Assistant B:
The colors on the second row (middle shelf) from right to left (backwards from the original left to right order) are: Purple, Orange, Blue, Dark Brown.

###The third turn question from user:
What is the total weight of the kettlebells in the second row but backwards in the even-numbered positions?

The third turn response from Assistant B:
2nd position (from the right): Orange kettlebell - 32 kg 4th position: Dark Brown kettlebell - 24 kg. Adding these weights together: 32 kg (Orange) + 24 kg (Dark Brown) = 56 kg. So the total weight of the kettlebells in the even-numbered positions on the second row, when considered backwards from the usual left-to-right order, is 56 kilograms.
<|The End of Assistant B' s Conversation with User|>

The first turn evaluation: Assistant A accurately provides the weights of each color of kettlebells as requested by the user,...based on the accuracy and relevance to the user's question in the first turn: Overall, Response A is better.1. Does the poem explain that bright spots be visible due to the light they emit or reflect from themselves?

Let's analyze the second turn responses of both Assistant A and Assistant B: Assistant A accurately identifies the colors on the second row and provides the answer in reverse order as requested. Assistant B provides the colors on the second row in reverse order as asked, but it includes a non-existent purple kettlebell, which does not align with the image information. Overall, Response A is better.

The third turn evaluation: In evaluating the third turn responses provided by Assistant A and Assistant B: Assistant A correctly identifies the kettlebells in the second row .... Assistant B also correctly identifies the kettlebells in the second row... Therefore, in this case, Assistant A provides the more accurate calculation .. Overall, Response A is better.
Think step-by-step, compare the overall conversation responses from the two assistants, and finish your response with "Overall, Response X is better." where X is either A or B.

assistant (machine generated; used to derive the prediction)

Let"s analyze the conversations:
Assistant A:
In the first turn, Assistant A correctly provides the weights of each color of kettlebell as asked by the user. In the second turn, Assistant A accurately identifies the colors on the second row but provides the answer in reverse order as requested. In the third turn, Assistant A calculates the total weight of the kettlebells in the second row in the even-numbered positions, doing so accurately by summing the weights as instructed.
Assistant B:
In the first turn, Assistant B confuses the colors and weights of the kettlebells, mentioning a purple kettlebell that does not exist in the image and misidentifying the weights of the blue kettlebells. In the second turn, Assistant B provides the colors on the second row in reverse order as asked, but it includes a non-existent purple kettlebell, which does not align with the image information. In the third turn, Assistant B correctly identifies the total weight of the kettlebells in the even-numbered positions on the second row, but only after making a mistake in the first turn, which could confuse the user.
Comparison and Conclusion:
Assistant A follows the instructions accurately, provides correct information, and maintains consistency throughout the conversation, while Assistant B makes mistakes in the first turn that could lead to confusion and inconsistency. Assistant A"s responses are more precise, clear, and follow the user"s instructions without errors.
Overall, Response A is better.'

Fig. 14: The prompt used for evaluating overall conversation in a pairwise grading method, accompanied by a sample completion from ChatGPT-3.5, is provided. The response conversations are anonymously presented to ChatGPT-3.5, where Assistant A is identified as a human, and Assistant B is recognized as GPT-4V.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instructions.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Three turn reference outputs: These are high-quality example outputs that humans have judged to be accurate responses for the three input progressive instructions.
4. Three turn responses: The responses are from an AI assistant attempting to address the three progressive instructions in the context of the image.

Your job is to rate the first turn response from the AI assistant for the visual perception performance on a scale of 1 to 10. Rate the first turn response from the AI assistant by regarding the rating of the first turn reference output as 10.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgement can be accurate.
- Regard the ratings of the high-quality references as 10. Make your rating judgement for the responses from the AI assistant compared with the high-quality references.
- You are capable of judging responses quality. The first turn instruction is visual perception perspective. Correctness, relevance, fluency and the level of detail of responses are the most important factors which should be accounted for the first turn response.
- You think step-by-step and be as objective as possible, after providing your explanation, you must rate the first turn response for the visual perception performance on a scale of 1 to 10 by strictly following this format:"Rating:{rating}", for example: "Rating:{5}".

user (human authored)

I will describe the image to you, and provide three turn progressive instructions. Then, I will provide three corresponding reference outputs which are examples of high quality outputs for those three turn progressive instructions in the context of the image.
Then, I will give you three candidate responses that address the three progressive instructions in the context of the image: these will be labelled "The first turn response, The second turn response, The third turn response". Your job is to first reason step-by-step about the procs/cons of the first turn candidate response in terms of accuracy, relevance, fluency, the level of detail of responses etc. After reasoning step-by-step, comparing between the first turn candidate response and the first turn reference output and making the judgement by regarding the rating of the first turn reference output as 10, in the end, respond with "Rating:X." where X is a scale of 1 to 10.

assistant (human authored)

Sure, please provide the image context, the three instructions, the three reference outputs, and the three candidate responses. Then, I will think step-by-step and provide my ultimate rating judgement for the first turn candidate response.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: A woman is playing a game of miniature golf. She is holding a mini golf putt. The image portrays an outdoor setting, most likely a mini-golf course or putting green, given the context...

<The Start of Reference Answer|>
The first turn question from user:
What game is the person playing?

The first turn high quality reference:
A woman is playing a game of miniature golf.

The second turn question from user:
Is this person playing at a professional level? Give me your reasoning for your answer.

The second turn high quality reference:
It does not appear that the person is playing mini-golf at a professional level. Here are some reasons for this determination:1. Attire: The individual is dressed in casual clothing\u2014a blue T-shirt...

The third question from user:
Can you develop a weekly exercise plan to improve the sport level of playing the game for this person?

The third turn high quality reference:
Improving at miniature golf largely involves enhancing hand-eye coordination, precision, and understanding the physics of how the ball moves on different surfaces. Here's a weekly exercise plan that focuses on these skills, which are crucial for improving performance in miniature golf:
Week 1: Fundamentals and Putting Technique
Monday - Understanding Basics: 30 minutes of putting on a flat surface. Focus on stance, grip, and swing mechanics.
Tuesday - Control Practice:Practice putting different distances (3 ft, 6 ft, 9 ft). 20 putts from each distance, noting consistency and accuracy. Regular practice on different types of putting surfaces and inclines to develop a consistent putting stroke and better judgment of distances and slopes.
Wednesday - Rest Day.
Thursday - Directional Control: Place obstacles on the putting surface and practice putting around them. 30 minutes of practice aiming for precision rather than power.
<|The End of Reference Answer|>

<|The Start of Assistant A' s Conversation with User|>
The first turn question from user:
What game is the person playing?

The first turn response from Assistant A:
The person in the image is playing miniature golf, also known as mini-golf or putt-putt golf.

The second turn question from user:
Is this person playing at a professional level? Give me your reasoning for your answer.

The second turn response from Assistant A:
1. The setting: The course looks like a typical recreational mini-golf course rather than a professional putting green used in tournaments. 2. Attire: The person is dressed in casual clothing, including sandals and a straw hat, which is more indicative of a casual player rather than professional golf attire...

The third question from user:
Can you develop a weekly exercise plan to improve the sport level of playing the game for this person?

The third turn response from Assistant A:
Creating a weekly exercise plan to improve performance in mini-golf should focus on enhancing coordination, balance, and concentration, as well as practising putting skills. However, it''s important to note that any new exercise plan should be started after consulting with a health professional, especially if the individual has any pre-existing health conditions or concerns.
Here''s a sample plan tailored for someone playing mini-golf:
Monday:
- Warm-up: 10 minutes of brisk walking.
- Core strengthening exercises: 3 sets of plank holds for 30 seconds each.
- Balance exercises: Standing on one foot for 30 seconds, then switch feet (3 sets each side).
- Putting practice: 30 minutes focusing on stance and stroke consistency.
Tuesday:
- Rest day or light activity such as walking to allow the body to recover...
<|The End of Assistant A' s Conversation with User|>

Think step-by-step, rate the first turn response from Assistant A, and finish your response with "Rating:X." where X is all a scale of 1 to 10.

assistant (machine generated; used to derive the prediction)

The first turn response from Assistant A accurately identifies the game being played in the image as miniature golf, which aligns perfectly with the image context and the first turn reference output. The response is relevant, clear, and fluent in its delivery. It explicitly states the specific game being played and uses alternative terms (mini-golf, putt-putt golf) that demonstrate a comprehensive understanding of the activity depicted in the image. Furthermore, the response is adequately detailed by providing specific terminology related to the game, showing attention to contextual accuracy.
Compared to the high-quality reference output for the first turn instruction, which is succinct and directly answers the question, Assistant A''s response mirrors the precision and correctness expected, hence deserving a high rating. Rating: 9

Fig. 15: The prompt used for evaluating perception in a direct grading method, accompanied by a sample completion from ChatGPT-3.5, is provided.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instructions.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Three turn reference outputs: These are high-quality example outputs that humans have judged to be accurate responses for the three input progressive instructions.
4. Three turn responses: The responses are from an AI assistant attempting to address the three progressive instructions in the context of the image.

Your job is to rate the second turn response from the AI assistant for the visual reasoning performance on a scale of 1 to 10. Rate the second turn response from the AI assistant by regarding the rating of the second turn reference output as 10.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgement can be accurate.
- Regard the ratings of the high-quality references as 10. Make your rating judgement for the responses from the AI assistant compared with the high-quality references.
- You are capable of judging responses quality. The second turn instruction is visual reasoning perspective. Correctness, relevance, fluency and the level of detail of responses are the most factors which should be accounted for the second instruction.
- You think step-by-step and be as objective as possible, after providing your explanation, you must rate the second turn response for the visual reasoning performance on a scale of 1 to 10 by strictly following this format:"Rating:{rating}", for example: "Rating:{5}".

user (human authored)

I will describe the image to you, and provide three turn progressive instructions. Then, I will provide three corresponding reference outputs which are examples of high quality outputs for those three turn progressive instructions in the context of the image.
Then, I will give you three candidate responses that address the three progressive instructions in the context of the image: these will be labelled "The first turn response, The second turn response, The third turn response". Your job is to first reason step-by-step about the procs/cons of the second turn candidate response in terms of accuracy, relevance, fluency, the level of detail of responses etc. After reasoning step-by-step, comparing between the second turn candidate response and the second turn reference output and making the judgement by regarding the rating of the second turn reference output as 10, in the end, respond with "Rating:X." where X is a scale of 1 to 10.

assistant (human authored)

Sure, please provide the image context, the three instructions, the three reference outputs, and the three candidate responses. Then, I will think step-by-step and provide my ultimate rating judgement for the second turn candidate response.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: The image portrays two individuals, both dressed in what appears to be period attire. Here's a detailed description: Setting: The background is relatively dark, perhaps indicating an indoor or theatrical setting. The floor appears to be wooden...

<The Start of Reference Answer|>
The first turn question from user:
Can you describe the people's attire shown in the image?

The first turn high quality reference:
A woman is playing a game of miniature golf.

The second turn question from user:
What religion do these people belong to? What time period and location are they from?

The second turn high quality reference:
Based on their attire, which includes modest and functional clothing with white bonnets, large white collars, brown skirts, white aprons, and black shoes, the individuals in the image seem to be representing Puritans...

The third question from user:
Outline a powerpoint what the slides could be like for a history teacher to introduce the relevant history based on the previous responses and image.

The third turn high quality reference:
I can outline what the slides could look like for a histor teacher''s presentation about Puritans in 17th-century Massachusetts based on the attire and context provided:
Slide 1: The Puritans of 17th-Century Massachusetts
- Image: (Insert image from above)
- Subtitle: "An Exploration of Colonial Life and Beliefs"
<|The End of Reference Answer|>

<|The Start of Assistant A' s Conversation with User|>
The first turn question from user:
Can you describe the people's attire shown in the image?

The first turn response from Assistant A:
The individuals in the image are wearing attire that resembles clothing from a historical period, perhaps suggestive of peasant or working-class dress from several centuries ago. Their costumes include the following elements:...

The second turn question from user:
What religion do these people belong to? What time period and location are they from?

The second turn response from Assistant A:
Based on the image alone, it's not possible to accurately determine the religion, exact time period, or the location of the people depicted...

The third question from user:
Outline a powerpoint what the slides could be like for a history teacher to introduce the relevant history based on the previous responses and image.

The third turn response from Assistant A:
Title: Life in the Past: Exploring Historical Clothing and Society
Slide 1: Introduction
- Title: "Understanding the Past through Fashion"
- Image of the people with historical attire
- Brief introductory text: "Discover how clothing can reveal information about historical periods, social status, and daily life."
<|The End of Assistant A' s Conversation with User|>

Think step-by-step, rate the second turn response from Assistant A, and finish your response with "Rating:X." where X is all a scale of 1 to 10.

assistant (machine generated; used to derive the prediction)

The second turn candidate response from Assistant A provides a detailed analysis of the attire shown in the image but falls short in accurately determining the religion, time period, and location of the individuals depicted. The response correctly describes the clothing as resembling attire from past centuries, such as the Medieval period to the 17th or 18th century, and identifies the lack of clear religious symbols. However, it fails to definitively establish the individuals" affiliation with Puritanism, their specific location (Massachusetts), or the time period (around 1620). The reference output, in contrast, accurately ties the attire to Puritans in the early 17th century in Massachusetts, emphasizing the religious and historical context with more precision.

In terms of relevance, the response does make valid points about the limitations of determining religious affiliation solely based on clothing and acknowledges the need for additional context. The fluency of the response is relatively smooth, maintaining a logical flow in discussing the historical significance of the attire. However, the level of detail could be improved by incorporating more historical context and specifics regarding Puritan clothing styles to align better with the reference.

Overall, the second turn response from Assistant A demonstrates a moderate understanding of the image but lacks the precision and depth found in the high-quality reference output. Therefore, I would rate the second turn response as follows:

Rating: 6

Fig. 16: The prompt used for evaluating reasoning in a direct grading method, accompanied by a sample completion from ChatGPT-3.5, is provided.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging whether or not a response adequately addresses an instruction in the context of an image. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instructions.
2. Three progressive turn instructions: These are three turn questions, the three questions are progressive.
3. Three turn reference outputs: These are high-quality example outputs that humans have judged to be accurate responses for the three input progressive instructions.
4. Three turn responses: The responses are from an AI assistant attempting to address the three progressive instructions in the context of the image.
5. Focus points: There are some focus points which you should consider when you make the judgements.

Your job is to rate the third turn response from the AI assistant for the composition on a scale of 1 to 10. Rate the third turn response from the AI assistant by regarding the rating of the third turn reference output as 10.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgement can be accurate.
- Regard the ratings of the high-quality references as 10. Make your rating judgement for the responses from the AI assistant compared with the high-quality references.
- You are capable of judging responses quality. Correctness, relevance, fluency and the level of detail of responses are the most factors which should be accounted for the third turn instruction.
- You think step-by-step and be as objective as possible, after providing your explanation, you must rate the third turn response for the composition on a scale of 1 to 10 by strictly following this format:"Rating:{rating}", for example: "Rating:{5}".

user (human authored)

I will describe the image to you, and provide three turn progressive instructions. Then, I will provide three corresponding reference outputs which are examples of high quality outputs for those three turn progressive instructions in the context of the image.
Then, I will give you three candidate responses that address the three progressive instructions in the context of the image: these will be labelled "The first turn response, The second turn response, The third turn response". Your job is to first reason step-by-step about the procs/cons of the third turn candidate response in terms of accuracy, relevance, creativity, fluency, the level of detail of responses etc. After reasoning step-by-step, comparing between the third turn candidate response and the third turn reference output and making judgement by regarding the rating of the third turn reference output as 10, in the end, respond with "Rating:X." where X is a scale of 1 to 10.

assistant (human authored)

Sure, please provide the image context, the three instructions, the three reference outputs, and the three candidate responses, the focus points. Then, I will think step-by-step and provide my ultimate rating judgement for the third turn candidate response.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: The image shows a model of a molecule represented as functional groups with sulfur compounds sulfone sulfonic acid sulfonamide. There are three parts in the molecule image. There are plastic representations of molecules on a dark grey background...
<The Start of Reference Answer|>
The first turn question from user:
What is the compound shown in the image?

The first turn high quality reference:
The image provided shows models of molecules typically used for educational purposes to demonstrate molecular structure. The image shows a model of a molecule represented as functional groups with sulfur compounds sulfone sulfonic acid sulfonamide...

The second turn question from user:
What is the connection between this compound and flatulence?

The second turn high quality reference:
The molecules depicted in the image are indeed models of sulfur-containing compounds, which can be related to flatulence due to the production of sulfur-bearing gases during digestion. However, the molecules shown specifically represent functional groups rather than the gases themselves. Here is a clarification of the connection:

The third question from user:
What other side effect does the compound which produces flatulence have. Pretend yourself to be a medical student.

The third turn high quality reference:
Sulfonamides, also known as sulfa drugs, are a class of antimicrobial agents that are widely used in the treatment of bacterial infections. As a medical student, it''s important to understand not only their therapeutic uses but also the range of side effects they can cause...
<|The End of Reference Answer|>

<|The Start of Assistant A’ s Conversation with User|>
The first turn question from user:
What is the compound shown in the image?

The first turn response from Assistant A:
The image shows molecular models of water (H\u2082O) and ammonia (NH\u2083). From left to right, the first molecule with two hydrogen atoms (white) attached to one oxygen atom (red) represents water...

The second turn question from user:
What is the connection between this compound and flatulence?

The second turn response from Assistant A:
The compound ammonia (NH\u2083), as represented by two of the molecular models in the image (with nitrogen in blue and hydrogen in white), is relevant to flatulence because it is one of the compounds that can be produced during the digestion process. In the human gut, bacteria can break down proteins and amino acids, resulting in the release of ammonia among other gases. ..

The third question from user:
What other side effect does the compound which produces flatulence have. Pretend yourself to be a medical student.

The third turn response from Assistant A:
As a medical student, you would understand that ammonia (NH\u2083), which is produced in the gut during protein digestion, not only contributes to flatulence but also to other physiological processes. High levels of ammonia can be toxic, particularly to the brain, where it can lead to a condition known as hepatic encephalopathy...
<|The End of Assistant A’ s Conversation with User|>

There are some concerns which you should focus when make your judgements for the response
1. Whether allergic reactions are mentioned in the answer?
2. Whether renal effects are mentioned in the answer?
3. Whether central nervous system effects are mentioned in the answer?:
Think step-by-step, rate the third turn response from Assistant A, and finish your response with "Rating:X." where X is all a scale of 1 to 10.

assistant (machine generated; used to derive the prediction)

In analyzing the third turn response from Assistant A, let''s break down the key points:
1. **Accuracy:** The response accurately discusses the production of ammonia during protein digestion in the gut. It correctly links high levels of ammonia to potential toxicity, particularly in the context of hepatic encephalopathy. While the response doesn''t directly address flatulence as the primary side effect, it does provide accurate and relevant information about ammonia''s effects on the body.
2. **Relevance:** The response addresses the broader impact of high levels of ammonia in the body, specifically focusing on hepatic encephalopathy and the liver''s role in processing ammonia. While it doesn''t directly
3. **Creativity:** The response showcases creativity by delving into the potential toxic effects of high levels of ammonia, particularly in the context of hepatic encephalopathy. It demonstrates a deeper understanding of the
4. **Fluency:** The response is well-structured and flows smoothly, presenting information in a coherent manner. The use of technical terms like hepatic encephalopathy and ammonia metabolism ..
5. **Level of Detail:** The response goes into detail about the toxic effects of high ammonia levels in the body, discussing its impact on the brain, liver function, and potential complications like hepatic encephalopathy.
Comparing the third turn response from Assistant A with the high-quality reference output, Assistant A provides accurate and detailed information about the effects of high levels of ammonia in the body. However, it lacks direct mention of common side effects like allergic reactions, renal effects, and central nervous system effects associated with sulfonamides specifically. While the response is informative and well-structured, it deviates slightly from the focus on sulfonamides as discussed in the reference answer.
Considering these points, I would rate the third turn response from Assistant A as follows:
Rating: 7

Fig. 17: The prompt used for evaluating creation in a direct grading method, accompanied by a sample completion from ChatGPT-3.5, is provided.

system prompt (human authored)

You are ImageTaskEvaluationGPT, an expert language model at judging the multi-turn conversation instruction-following ability of an AI assistant. More specifically, you will be given the following:

1. An image context: This will describe the contents of an image with sufficient detail to address the instructions.
2. Three progressive turn instruction: These are three turn questions, the three questions are progressive.
3. Three reference outputs: These are high-quality example outputs that humans have judged to be accurate responses for these input progressive instructions.
4. Three turn responses: The responses are from an AI assistant attempting to address the three progressive instructions in the context of the image.
5. Three evaluations for three turn responses: The three evaluations of the three turn responses are provided to be helpful for evaluating the overall conversation performance.

Your job is to rate the overall conversation on a scale of 1 to 10. Rate the overall conversation by regarding the ratings of the reference outputs as 10.

Some things to remember:

- Even though you are just a language model, the image description will be sufficiently detailed so that your judgements can be accurate.
- Regard the ratings of the high-quality references as 10. Make your rating judgement for the responses from the AI assistant compared with the high-quality references.
- Take the rating of each turn into account when making your judgements for the overall conversation.
- You are capable of judging overall conversation quality, accounting for the multi-turn conversation and instruction-following ability. Correctness, relevance, fluency and the level of detail of responses are the most factors which should be considered.
- As for the open-ended instructions, creativity, helpfulness, specificity and level of detail of responses should also be considered.
- You think step-by-step and be as objective as possible, after providing your explanation, you must rate the overall conversation on a scale of 1 to 10 by strictly following this format:"Rating: {rating}", for example: "Rating: 3.".

user (human authored)

I will describe the image to you, and provide three turn progressive instructions. Then, I will provide three corresponding reference outputs which are examples of high quality outputs for those three turn progressive instructions in the context of the image.
Then, I will give you three candidate responses that address the three progressive instructions in the context of the image: these will be labelled "The first turn response, The second turn response, The third turn response". Your job is to first reason step-by-step about the procs/cons of the overall conversation in terms of accuracy, relevance, creativity, fluency, the level of detail of responses etc. After reasoning step-by-step, comparing between the each turn candidate response and the each turn reference output and making judgement by regarding the rating of the each turn reference output as 10, in the end, respond with "Rating:X" where X is a scale of 1 to 10.

assistant (human authored)

Sure, please provide the image context, the three instructions, the reference outputs, the candidate responses and the three evaluations for the three turn responses. Then, I will think step-by-step and provide my ultimate rating judgement for the overall conversation.

user (from human authored template)

OK. Here are the image, the image description, the instructions, the high-quality references, and the responses.
Image context: There is are two shelves with a slightly reflective surface, which are each in a separate area of the cabinet. On the left of the image, there are three columns of round white medicine bottles...

<The Start of Reference Answer|>
The first turn question from user:
Describe the image in detail focusing on the words.

The first turn high quality reference:
There is are two shelves with a slightly reflective surface, which are each in a separate area of the cabinet. On the left of the image, there are three columns of round white medicinal bottles. Each bottle has the label \u2018stress killers 200pcs\u2019 written on the front in black text...

The second turn question from user:
What do these two medical products do?

The second turn high quality reference:
Based on the labels, the products are suggested to serve the following purposes: the bottles labeled "stress killers" imply that they are intended to eliminate or reduce stress, and the cans labeled "inner calmness" suggest that they are meant to promote or enhance a sense of calm and tranquility within an individual...

The third question from user:
Come up with a short, catchy ad title for the two products displayed in the cabinet, including the efficacy of these two medicines.

The third turn high quality reference:
Conquer Stress, Embrace Serenity: Unleash the power of 'Stress Killers' and 'Inner Calmness' for a peaceful you!_ Your Ultimate Wellness Duo!\"...
<|The End of Reference Answer|>

<|The Start of Assistant A’ s Conversation with User|>
The first turn question from user:
Describe the image in detail focusing on the words.

The first turn response from Assistant A:
The image displays two shelves containing products with labels that suggest they are related to stress management and well-being. On the left side of the image, there are white bottles with labels that read "stress killers." ...

The second turn question from user:
What do these two medical products do?

The second turn response from Assistant A:
Based on the image, the products are labeled in a way that suggests they may be designed to alleviate stress. The white bottles labeled "stress killers" imply that they might be used to help reduce or eliminate stress, while the silver cans labeled "inner calmness" suggest they could be intended to promote a feeling of calm or well-being...

The third question from user:
Come up with a short, catchy ad title for the two products displayed in the cabinet, including the efficacy of these two medicines.

The third turn response from Assistant A:
Conquer Stress and Embrace Serenity: Find Your Inner Calmness and Slay Stress with Every Dose!...
<|The End of Assistant A’ s Conversation with User|>

The first turn evaluation: The first turn response from Assistant A provides a thorough and detailed description of the image, focusing on the words featured on the products. The assistant accurately captures the details of the image, mentioning the labels "stress killers" and "inner calmness," along with the quantity and weight specifications on the labels...
The second turn evaluation: Firstly, let''s evaluate the second turn response from Assistant A in terms of accuracy, relevance, fluency, and level of detail compared to the high-quality second turn reference output...
The third turn evaluation: Upon examining the third turn response from Assistant A, here is a step-by-step analysis:
1. **Relevance to the instruction:** - The response does come up with a catchy ad title for the two products, which alignswith the instruction given. It emphasizes stress relief and serenity, as requested. It is relevant in addressing the task at hand...
Think step-by-step, rate the overall conversation from Assistant A, and finish your response with "Rating:X." where X is all a scale of 1 to 10.

assistant (machine generated; used to derive the prediction)

Overall, Assistant A demonstrated a strong understanding of the image content and responded well to the three progressive turn instructions. The responses were accurate, relevant, and provided a good level of detail throughout the conversation.
In the first turn, Assistant A accurately described the image, focusing on the words present on the products, which was similar to the high-quality reference output. The response was detailed and effectively conveyed the information in the image.
In the second turn, Assistant A correctly interpreted the potential functions of the medical products based on their labels. The response was relevant, cautious in its interpretation, and maintained fluency in explaining the implications of the labels.
For the third turn, Assistant A came up with a catchy ad title that highlighted the efficacy of the two medicines in stress relief and serenity. While the response was creative, it lacked specificity in mentioning the product names, which could have enhanced the impact of the advertisement.
Overall, Assistant A's responses were consistently accurate and relevant, with good fluency and detail. The conversation flowed logically and maintained a focus on the image content and instructions provided.
Rating: 8

Fig. 18: The prompt used for evaluating overall conversation in a direct grading method, accompanied by a sample completion from ChatGPT-3.5, is provided.

system prompt (human authored)

You are FinalAnswerExtractionGPT, an expert language model at extracting multiple choice answers from written out from longer explanations. You will be given several sentences describing a thought process which should eventually contain a final answer, either A or B. Your job is to extract the final answer that the explanation arrives at.

Some things to remember:

- Keep your answers short: only output "Final Answer: X" where X is A or B
- If the explanation doesn't mention a final choice explicitly, you can output "Unknown"

user (human authored)

Step-by-step reasoning:

- First, I consider the instruction, which asks if anything else in the image has the same color as the tiny sphere.
- Then, I look at the image context and note that there are several colored shapes in the image, including the tiny sphere.
- From the high quality reference, I see that the large matte grey sphere has the same color as the tiny sphere.
- Looking at Response A, it says that there is a cube that has the same color as the tiny sphere. But the high quality reference only mentions the large matte grey sphere having the same color, not the cube. Therefore, Response A is not accurate.
- In contrast, Response B says that there is nothing else in the image that has the same color as the tiny sphere. The high quality reference contradicts this response, stating that the large matte grey sphere shares the same color with the tiny sphere.
- Therefore, Response B is not accurate or correct.
- Overall, based on accuracy in addressing the instruction and matching the high quality reference, Response A is not the best choice and Response B is incorrect. Therefore, the better response is Response B: "No, there is nothing else in the image that has the same color as the tiny sphere.

Please extract the final answer from the above text.

assistant (human authored)

Final Answer: Response B

user (human authored)

Let's compare the two responses step-by-step:

Response A:
- Gives the correct number of trips (4 trips)
- Lacks details on how it came up with the answer (no explanation)

Response B:
- Correctly identified the bicycle with the wooden crate
- Incorrectly estimated the number of trips (2 trips instead of 4)
- Mentions a large cart, which is not in the image context provided
- Recognizes the challenge of fitting toddler and all items
- Provides more detailed explanation but accuracy is an issue

Considering accuracy, specificity, fluency, and relevance, let's make a judgement:

Response B is better in terms of fluency and relevance while Response A is better in terms of accuracy and specificity. Since accuracy is crucial in this case, and given the high-quality reference backs the number of trips (4), we prioritize correctness.

Overall, I choose A.
Please extract the final answer from the above text.

assistant (human authored)

Final Answer: Response A

user (from human authored)

Upon analyzing the image description and the instruction, it is clear that the instruction is asking for the implied meaning of a comment received by the creator of the burnt pizza in the image. The high-quality reference suggests that the compliment of "being very talented" is actually sarcastic and that the burnt pizza is not a sign of talent. Looking at Response A, it seems like the response is describing the pizza and the plate without really addressing the implication of the comment. Response B, on the other hand, does directly address the instruction and implies that the praise is sincere even though the pizza is burnt. Therefore, overall, Response B is better.
Please extract the final answer from the above text.

assistant (human authored)

Final Answer: Response B

Fig. 19: The prompt used for extracting answers from ill-formatted ChatGPT-3.5 responses in a pairwise grading method.

system prompt (human authored)

You are FinalAnswerExtractionGPT, an expert language model at extracting multiple choice answers
from written out from longer explanations. You will be given several sentences describing a thought process which should eventually contain a final answer,
1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10. Your job is to extract the final answer that the explanation arrives at.

Some things to remember:

- Keep your answers short: only output "Final Rating: X" where X is 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
- If the explanation doesn't mention a final choice explicitly, you can output "Unknown"

user (human authored)

Step-by-step reasoning:

- First, I consider the instruction, which asks if anything else in the image has the same color as the tiny sphere.
- Then, I look at the image and image context and note that there are several colored shapes in the image, including the tiny sphere.
- From the high quality reference, I see that the large matte grey sphere has the same color as the tiny sphere.
- Looking at the Response, it says that there is nothing else in the image that has the same color as the tiny sphere. The high quality reference contradicts this response, stating that the large matte grey sphere shares the same color with the tiny sphere.
- Therefore, Response B is not accurate or correct.
- Overall, based on accuracy in addressing the instruction and matching the high quality reference, the Response is incorrect. Therefore, the response is rated at 4

assistant (human authored)

Final Rating: 4

user (human authored)

Let's evaluate the the responses step-by-step:

- Correctly identified the bicycle with the wooden crate
- Incorrectly estimated the number of trips (2 trips instead of 4)
- Mentions a large cart, which is not in the image context provided
- Recognizes the challenge of fitting toddler and all items
- Provides more detailed explanation but accuracy is an issue

Considering accuracy, specificity, fluency, and relevance, let's make a judgement:

Compared with the reference, the response is good in terms of fluency and relevance while is bad in terms of accuracy and specificity. Since accuracy is crucial in this case, and given the high-quality reference backs the number of trips (4), we prioritize correctness.

Overall, I rate the response at 2

assistant (human authored)

Final Rating: 2

user (from human authored)

Upon analyzing the image description and the instruction, it is clear that the instruction is asking for the implied meaning of a comment received by the creator of the burnt pizza in the image. The high-quality reference suggests that the compliment of "being very talented" is actually sarcastic and that the burnt pizza is not a sign of talent. Looking at the response does directly address the instruction and implies that the praise is sincere even though the pizza is burnt. Therefore, the response is at 6.

assistant (human authored)

Final Rating: 6

Fig. 20: The prompt used for extracting answers from ill-formatted ChatGPT-3.5 responses in a direct grading method.