

# BenchLMM: Benchmarking Cross-style Visual Capability of Large Multimodal Models

Rizhao Cai<sup>1\*</sup>, Zirui Song<sup>2,3\*</sup>, Dayan Guan<sup>1†</sup>, Zhenhao Chen<sup>4</sup>, Xing Luo<sup>5</sup>, Chenyu Yi<sup>1</sup>, Alex Kot<sup>1</sup>

<sup>1</sup>Nanyang Technological University <sup>2</sup>University of Technology Sydney <sup>3</sup>Northeastern University

<sup>4</sup>Mohamed bin Zayed University of Artificial Intelligence <sup>5</sup>Zhejiang University

## Abstract

*Large Multimodal Models (LMMs) such as GPT-4V and LLaVA have shown remarkable capabilities in visual reasoning with common image styles. However, their robustness against diverse style shifts, crucial for practical applications, remains largely unexplored. In this paper, we propose a new benchmark, BenchLMM, to assess the robustness of LMMs against three different styles: artistic image style, imaging sensor style, and application style, where each style has five sub-styles. Utilizing BenchLMM, we comprehensively evaluate state-of-the-art LMMs and reveal: 1) LMMs generally suffer performance degradation when working with other styles; 2) An LMM performs better than another model in common style does not guarantee its superior performance in other styles; 3) LMMs' reasoning capability can be enhanced by prompting LMMs to predict the style first, based on which we propose a versatile and training-free method for improving LMMs; 4) An intelligent LMM is expected to interpret the causes of its errors when facing stylistic variations. We hope that our benchmark and analysis can shed new light on developing more intelligent and versatile LMMs. The benchmark and evaluation code have been released in <https://github.com/AIFEG/BenchLMM>.*

## 1. Introduction

The dynamic landscape of computer vision has recently been shaped by the rise of Large Multimodal Models (LMMs). These models [10, 14, 32, 38, 69], incorporating visual and textual data, have emerged as cornerstones in the quest for building general-purpose assistants for visual reasoning, a widely-studied task that involves reasoning and answering questions based on images. Their growing prominence and the enthusiasm surrounding visual reasoning are evident from numerous studies like GPT-4V [60]



Figure 1. The motivation of our work. State-of-the-art LMMs, such as GPT-4V and LLaVA-1.5 can understand the image in a common style. However, these models have difficulties in reasoning images in other styles.

and LLaVA-1.5 [37]. As the interest in LMMs swells within the research community, understanding their depth, breadth, and limitations becomes imperative.

In pursuit of a comprehensive understanding of the visual reasoning capabilities of LMMs, numerous benchmarks [17, 21, 31, 34, 40, 65] have been proposed to systematically evaluate the linguistic capabilities of LMMs. These linguistic capabilities are assessed through various types of responses for visual reasoning tasks, including Yes/No answers [17], multi-choice answers [31, 40], single word or phrase answers [21, 24, 43]. However, while these benchmarks showcase textual diversity, they may fall short in terms of visual diversity as most images in these benchmarks are photographic images sourced from the internet. Despite providing valuable insights into LMMs' per-

\*Equal contribution

†Corresponding author

formance, there remains a gap in understanding how LLMs handle shifts in visual distribution. As illustrated in Figure 1, our preliminary study shows that LMMs tend to make mistakes in reasoning images with non-common styles.

To examine LMM’s cross-style visual capability, we propose BenchLMM, a novel benchmark that systematically assesses existing LMMs in three distinct styles of distribution shifts: artistic style, sensor style, and application style. Firstly, there are images of different artistic styles. Objects in the common style are realistic, but objects can have other artistic styles, such as Cartoon, Painting, Sketch, etc. Besides, we also identify the difference in sensor styles, as there are imaging sensors other than RGB cameras, such as infrared and x-ray, leading to the distribution shifts of the images. Moreover, we also identify the shift of knowledge required for different application styles. For example, the image recognition task requires knowledge about object shape and appearance, while the task of decision-making for robot action requires task-specific knowledge.

With BenchLLM, we evaluate cutting-edge LLMs through benchmark experiments and our designed error-reflection study. Through these comprehensive experiments, we observe that these state-of-the-art LMMs perform well on the data of common style but suffer significant performance drops in the other styles; Also, we observe that an LMM achieving better performance in common style than another model does guarantee its superior performance in other styles; In addition, we find that prompting an LMM to predict what is the style of the data and then answer the question can enhance LMM’s reasoning capability. Based on this, we propose Style Prompt Enhancement (SPE) as a training-free strategy to improve LMM’s performance.

Moreover, we analyze the failure case by conducting an error-reflection analysis by asking an LMM to interpret why it made mistakes in the Q&A. We find that a more intelligent LMM can understand the detailed reason why it makes mistakes and can learn to derive the correct answer, which is an important capability that deserves attention when developing LMMs. We hope that our benchmark and analysis can shed new light on developing more intelligent and versatile LMMs. Our contributions are summarized below:

- We propose BenchLMM, the first benchmark that can be used to assess LMMs’ capabilities, containing a diverse set of distribution shifts in artistic style, sensor style, and application style;
- We broadly benchmark the existing large multimodal models, including commercial GPT-4V with quantitative metrics, which is the first time;
- Our in-depth benchmark analysis, our error-reflection study, and the proposed SPE provide a new aspect of understanding LMMs, which are useful for future research.

## 2. Related works

**Visual Reasoning.** Visual reasoning, also known as visual question answering, aims to reason and answer questions about visual information [3, 27, 28]. Early developments [5, 18, 29] focus on designing more complex fusion mechanisms instead of simple summation fusion [3]. By including modern attention mechanism [56], numerous works [8, 30, 44, 51] contributed to the creation of transformer-based vision-language models. Recent methods have proposed to improve other aspects of visual question answering, including avoiding shortcut learning [15, 23], out-of-distribution generalization [9, 53], addressing the linguistic bias [9, 53], improving transformer-based vision-language models [59, 68], external knowledge integration [16, 19], and consistency regularization [47, 52]. Lastly, the advent of large multimodal models has markedly enhanced the capabilities of visual reasoning.

**Large Multimodal Models.** Recent advancements in large language models (LLMs) [7, 13, 46, 54] have inspired the development of large multimodal models (LMMs) that enhance LLMs with vision-language capabilities. Existing LMMs can be categorized into three main classes: multimodal instruction tuning [10, 14, 32, 37, 38, 69] that involves the finetuning of pre-trained LLMs on multi-modal instruction-following data; multimodal in-context learning [2, 32, 55, 61] that derives insights from few-shot task-specific samples; multimodal chain-of-thought [20, 67] that encourages LMMs to articulate both the final answer and the underlying reasoning process. Additionally, some commercial LMMs like Google’s Bard [41] and Microsoft’s GPT-4V [60], with detailed information about their architecture remaining undisclosed, have shown impressive visual capabilities.

**Evaluation of LMMs.** To evaluate their visual capabilities, existing LMMs have been tested on numbers of vision-language datasets. These datasets focus on specific visual reasoning tasks by designing different kinds of questions and answers, and the specific task includes visual recognition [21, 43], image description [1, 12], scene text understanding [49, 50, 62], object hallucination evaluation [33], complex reasoning [38, 65], and commonsense reasoning [64, 66]. To facilitate a comprehensive comparison of LMMs, several benchmarks [17, 35, 58] have been established by encompassing various visual reasoning tasks with common-style images and diverse Q&A pairs. However, the above benchmarks do not consider different styles of visual distribution shift, which motivates us to propose our BenchLMM.

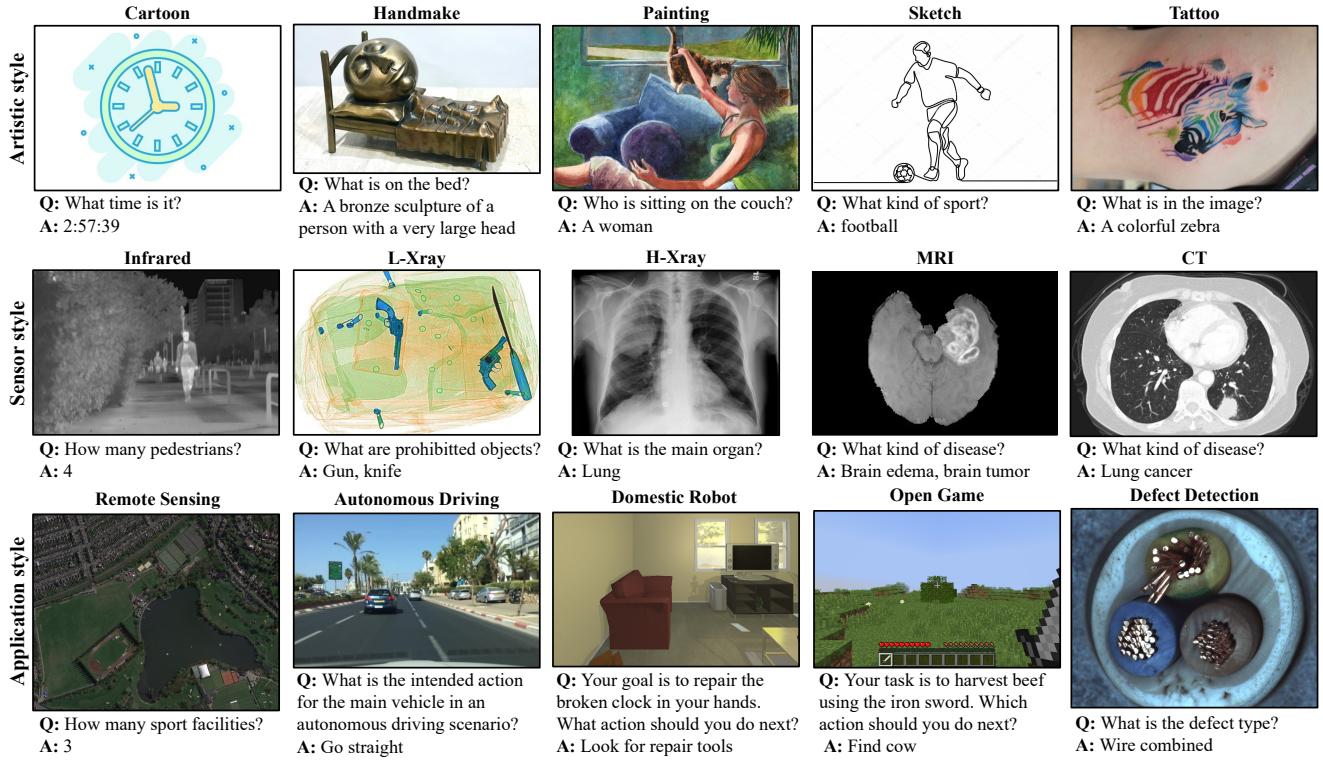


Figure 2. Benchmark examples. For each specific field, one example image and its corresponding Q&A have been shown. Note that, for a simple presentation, the questions in *Domestic Robot* and *Open Game* have been simplified from multiple-choice format. Please see the Appendix for more examples and detailed questions.

### 3. BenchLMM

#### 3.1. Motivation

We propose BenchLMM to investigate the cross-style capability of Large Multimodal Models (LMMs). This includes their proficiency in tasks like analyzing images with diverse styles, processing images acquired from non-RGB cameras, and interpreting images sourced from specific application knowledge. In this section, we will elaborate on how we build the benchmark that encompasses these diverse styles.

**Cross-artistic style** Existing benchmarks for LMMs commonly use photo images for evaluation, which we name as ‘Common’ style, because of their widespread use in applications such as education and e-commerce. However, there are diverse image styles, including Cartoon, Sketch, etc. In Figure 2, we visually compare ‘Common’ style images with various other styles. The object in other artistic styles have different distributions of object characteristics. For instance, a colorful zebra, common in painting and tattoo styles, is rare in ‘Common’ style images. While humans easily interpret such characteristics, the extent of LMMs’ understanding of the other styles remains uncertain. This uncertainty arises from potential insufficient training on datasets encompassing these diverse styles. Therefore, we propose an evaluation to assess LMMs’ robustness and proficiency in analyzing images across various artistic styles.

**Cross-sensor style.** Photographic images are typically obtained through RGB cameras for information sensing. However, alongside RGB cameras, other sensor types are also important and employed as complementary means of image capture. For instance, infrared cameras are frequently utilized in surveillance systems to capture images during nighttime, as RGB cameras rely on adequate illumination for optimal image quality. Additionally, X-ray sensors are used in airport security checks, enabling the detection of prohibited items within luggage.

These sensors feature distinct imaging processes compared to RGB cameras. Consequently, images acquired through various sensors exhibit unique intrinsic properties. The question of whether the performance of LMMs remains remarkable when applied to other sensors has not been thoroughly investigated. This issue will be addressed in our benchmark study.

**Cross-application style** Existing LMMs are predominantly assessed in general application scenarios where questions can be answered using common knowledge [17, 21]. For instance, identifying the object on a table in an image is a task that can be accomplished with general knowledge, and LMMs perform well in such cases. However, there exist specific application scenarios that demand specialized domain knowledge to address and solve. For example, when an autonomous driving control center processes the road im-

age to decide the next actions: going straight, turning left, or turning right), specific driving knowledge is required.

As existing LMMs are primarily evaluated with questions and answers (Q&A) that rely on common knowledge, the extent to which an LMM can remain effective in situations requiring domain-specific knowledge has not been comprehensively explored. We also investigate this aspect in our proposed benchmark.

### 3.2. Benchmark Construction

We construct our benchmark by reorganizing publicly available datasets and relabeling them with VQA annotations.

**Artistic style data.** To construct a benchmark characterized by distributional shifts pertaining to artistic styles, we curated a diverse set of images encompassing Cartoon, Painting, Sketch, Handmade, and Tattoo styles from the COCO-O dataset [42]. We systematically sampled 100 images for each stylistic category, resulting in a total of 500 images extracted from the COCO-O dataset. This dataset comprises a rich array of object types, providing a comprehensive evaluation ground for assessing the image understanding capabilities of the existing LMMs. Accordingly, the questions are prepared as asking the models to reason and answer questions about images, and the sample images with Q&A are shown in the first row of Figure 2.

**Sensor-style data.** Other than RGB cameras, Infrared sensors, X-ray sensors, Magnetic Resonance Imaging (MRI) sensors, and Computed Tomography (CT) sensors are often used to collect image data for human beings to perceive and analyze. Our benchmark examines whether the LMMs can still have the comparative reasoning capability toward these images that are intrinsically different from RGB photo images. In the case of Infrared data, we collected a set of 200 images from the datasets that focus on street scenes analysis toward with pedestrians, buildings, and vehicles [22, 25, 26, 39]. Questions for LMMs mainly revolve around analyzing the contents of these street scenes. Specifically, we design some thermal-related questions such as determining the hottest content. An illustrative example is provided in the second row of Figure 2.

For Low-energy X-ray (L-Xray) data from the SiXray-D dataset [45], we gathered 200 images, including 160 with prohibited objects (knife, scissors, wrench, gun, hammer) and 40 without any prohibited objects. Questions are designed to assess the ability to detect security risks in X-ray images, such as identifying forbidden objects. A sample question is presented in the second row of Figure 2. In the medical application, high-energy X-ray (H-Xray), MRI, and CT scanning images (200 each) were extracted from the [36]. Questions for LMMs in this context involve tasks like recognizing organs or identifying diseases in the respective images. By evaluating and comparing inference results be-

tween different sensor types, our benchmark aims to uncover challenges associated with cross-sensor analysis.

**Application-style data.** In order to establish a comprehensive cross-application-style benchmark, we systematically integrated diverse task domains into our evaluation framework. We consider four representative applications: remote sensing, autonomous driving, agent action prediction, and defect detection. Specifically, for the remote sensing application, we utilized 200 images sourced from the DOTA dataset [57] and annotated them in the VQA format from a satellite perspective. For the autonomous driving task, we extracted 250 images from the BDD100K dataset [63], featuring various weather conditions. Our primary focus was on designing questions and answers aimed at predicting driving actions such as going straight, turning left, or turning right. An illustrative example is provided in the third row of Figure 2.

To introduce the agent action prediction application, we meticulously selected 100 images from the Domestic Robot dataset [11] and an additional 117 images from the Open Game dataset [11]. The associated questions center around predicting the next action to be taken under the current circumstances. Furthermore, for evaluating defect detection capabilities, we curated 100 images from the Defect Detection dataset [6], with questions focused on identifying defect types and determining the presence of defects. This diverse set of applications and datasets collectively forms a diverse benchmark, allowing for a comprehensive assessment of the proposed approach across various application domains.

### 3.3. Style Prompt Enhancement

It is generally expected that providing additional prior information to LMMs' inputs can enhance performance. Nevertheless, such contextual priors, such as knowledge of artistic styles, might be unavailable in automated inference scenarios. We are inspired by human cognitive behavior that humans often comprehend the style of an image before delving into object recognition, akin to appreciating a painting by Van Gogh, where recognition of the impressionist style precedes identification of objects within the artwork. Other contextual priors, like sensor or application styles, can also enhance LMMs' performance. For example, identifying people in infrared images is easier when LMMs know that the images are captured using infrared thermal cameras, since humans typically show consistent temperature.

With this insight, we investigate the efficacy of prompting LMMs to predict the style of an image before addressing associated questions. To this end, we devised a Style Prompt Enhancement (SPE) by introducing a text prompt, instructing the model to "Determine the style of the image before answering the question." SPE aims to leverage stylistic cues as a precursor to visual question answering, poten-

Table 1. Evaluations of public LMMs on cross-style BenchLMM. Note that Average\* represents the average accuracy computed over five artistic-style benchmarks. The best and the second best results are highlighted in **bold** and underline, respectively. All the numbers are presented in % and the full score is 100%.

LMMs	Visual Encoder	Language Model	Common	Cartoon	Handmake	Painting	Sketch	Tattoo	Average*
GPT-4V [60]	-	-	<b>81.5</b>	<b>63.3</b>	<b>58.7</b>	<u>58.2</u>	<b>64.7</b>	<b>68.3</b>	<b>62.6</b>
LLaVA-1.5-13B [37]	CLIP-ViT-L/14	Vicuna-13B	<u>74.6</u>	<u>62.0</u>	56.6	57.6	<u>63.3</u>	57.0	<u>59.3</u>
InstructBLIP-13B [14]	EVA-ViT-G	Vicuna-13B	72.7	<u>59.0</u>	<u>57.1</u>	<b>59.3</b>	57.5	<u>61.4</u>	58.9
LLaVA-13B [38]	CLIP-ViT-L/14	Vicuna-13B	56.6	49.9	45.4	48.0	53.7	47.0	48.8
MiniGPT4-13B [69]	CLIP-ViT-L/14	Vicuna-13B	56.9	57.8	28.4	37.8	42.6	32.3	40.0
LLaVA-1.5-7B [38]	CLIP-ViT-L/14	Vicuna-7B	71.2	51.9	53.3	43.4	62.0	54.6	53.0
InstructBLIP-7B [14]	EVA-ViT-L	Vicuna-7B	73.9	58.1	52.0	55.8	55.2	55.5	55.3
MiniGPT4-v2-7B [10]	EVA-ViT-G	LLaMA2-7B	66.4	28.7	37.8	40.9	41.7	28.6	35.5
MiniGPT4-7B [69]	CLIP-ViT-L/14	Vicuna-7B	45.6	35.0	37.0	38.5	36.4	37.7	36.9
Otter-7B [32]	CLIP-ViT-L/14	MPT-7B	64.8	45.1	49.0	39.8	44.1	42.4	44.1

Table 2. Evaluations of public LMMs on cross-sensor BenchLMM. Note that GPT-4V denied to test medical images (H-Xray/MRI/CT).

LMMs	Visual Encoder	Language Model	Infrared	L-Xray	H-Xray	MRI	CT	Average
GPT-4V [60]	-	-	<b>55.0</b>	50.0	-	-	-	-
LLaVA-1.5-13B [37]	CLIP-ViT-L/14	Vicuna-13B	52.1	<u>51.5</u>	<u>55.9</u>	<u>33.7</u>	43.1	<b>47.3</b>
InstructBLIP-13B [14]	EVA-ViT-G	Vicuna-13B	40.8	24.6	<u>56.6</u>	20.9	43.4	37.3
LLaVA-13B [38]	CLIP-ViT-L/14	Vicuna-13B	<u>53.7</u>	46.0	39.1	28.4	39.7	41.4
MiniGPT4-13B [69]	CLIP-ViT-L/14	Vicuna-13B	32.9	47.2	24.6	25.3	23.2	30.6
LLaVA-1.5-7B [38]	CLIP-ViT-L/14	Vicuna-7B	46.5	<b>53.2</b>	49.4	29.6	<b>46.8</b>	<u>45.3</u>
InstructBLIP-7B [14]	EVA-ViT-L	Vicuna-7B	42.2	35.4	<b>59.5</b>	19.4	38.5	<u>39.0</u>
MiniGPT4-v2-7B [10]	EVA-ViT-G	LLaMA2-7B	30.1	43.0	31.5	17.4	17.8	28.2
MiniGPT4-7B [69]	CLIP-ViT-L/14	Vicuna-7B	35.1	44.1	23.7	16.8	15.3	27.0
Otter-7B [32]	CLIP-ViT-L/14	MPT-7B	30.2	46.4	39.1	<b>47.8</b>	<u>46.3</u>	42.0

tially enhancing the model’s ability of visual reasoning. Despite its simplicity, SPE’s versatility allows it to be applied directly to images of various styles without any additional prompt tuning or model training, clearly boosting the overall performance of LMMs, as detailed in Section 4.3.

## 4. Experiment

### 4.1. Experimental Setup

In this paper, we gather and assess a selection of existing open-source LMMs, which include LLaVA [37, 38], MiniGPT4 [10, 69], InstructBLIP [14], Otter [32], and OpenFlamingo-7B [4]. We initialize these models with pre-trained weights and conduct evaluations using our dataset. In addition, we use the variants of these models, where the language encoders and vision encoders are changed to investigate how different encoders have impacts on the cross-field capabilities. Moreover, we acknowledge the recent release of the latest commercial LMM, GPT-4V. We perform a quantitative evaluation of GPT-4V with our dataset and compare its performance with the aforementioned open-source LMMs, and such a quantitative evaluation is the first time.

In our performance evaluation, we adhere to previous works [37, 38] methodology, employing the ChatGPT API

to gauge the proximity of answers predicted by the LMMs to ground-truth answers. Here, we designate the output of the LMM as  $y$ , and the ground-truth label as  $\hat{y}$ . The evaluation utilizes the prompt “Compare the similarity between  $y$  and  $\hat{y}$  with a correctness score ranging from 0.0 (totally wrong) to 1.0 (totally right). The middle score provides the percentage of correctness.”

### 4.2. Evaluations of public LMMs

#### 4.2.1 Cross-artistic capability

In most existing works [17, 21, 31, 34, 40, 65], LMMs are predominantly evaluated using images in the ‘Photo’ style, leading to a gap in understanding their performance across diverse artistic styles. We extend the evaluation scope by examining LMMs’ performance with various artistic styles beyond the common ‘Photo’ style. Results, as detailed in Table 1, reveal a notable decline in LMMs’ effectiveness when processing these artistic styles. This trend suggests a potential overfitting of LMMs to the ‘Photo’ style, highlighting their limited adaptability to varied artistic styles, a capability that humans typically possess. Interestingly, GPT-4V, despite being a robust commercial model, exhibits similar limitations in handling diverse styles.

Besides, LLaVA-1.5-13B [37] is improved from LLaVA-

Table 3. Evaluations of public LMMs on cross-task BenchLMM.

LMMs	Visual Encoder	Language Model	Remote Sensing	Autonomous Driving	Domestic Robot	Open Game	Defect Detection	Average
GPT-4V [60]	-	-	<b>69.7</b>	<u>35.8</u>	<u>56.0</u>	<b>74.0</b>	<b>64.4</b>	<b>60.0</b>
LLaVA-1.5-13B [37]	CLIP-ViT-L/14	Vicuna-13B	<u>65.6</u>	28.6	49.0	21.4	<u>60.3</u>	<u>45.2</u>
InstructBLIP-13B [14]	EVA-ViT-G	Vicuna-13B	<u>63.6</u>	32.6	32.0	19.5	47.0	38.9
LLaVA-13B [38]	CLIP-ViT-L/14	Vicuna-13B	<u>38.7</u>	31.7	<b>59.0</b>	<u>43.8</u>	28.2	40.3
MiniGPT4-13B [69]	CLIP-ViT-L/14	Vicuna-13B	<u>33.3</u>	22.6	48.0	34.2	31.7	34.2
LLaVA-1.5-7B [38]	CLIP-ViT-L/14	Vicuna-7B	61.7	31.3	44.0	15.4	59.6	42.2
InstructBLIP-7B [14]	EVA-ViT-L	Vicuna-7B	<u>63.5</u>	<b>36.7</b>	42.0	24.8	30.8	39.6
MiniGPT4-v2-7B [10]	EVA-ViT-G	LLaMA2-7B	<u>34.5</u>	17.6	40.0	22.2	18.8	26.6
MiniGPT4-7B [69]	CLIP-ViT-L/14	Vicuna-7B	27.9	20.8	41.0	40.1	13.6	28.7
Otter-7B [32]	CLIP-ViT-L/14	MPT-7B	44.8	29.7	36.0	18.8	27.3	31.3

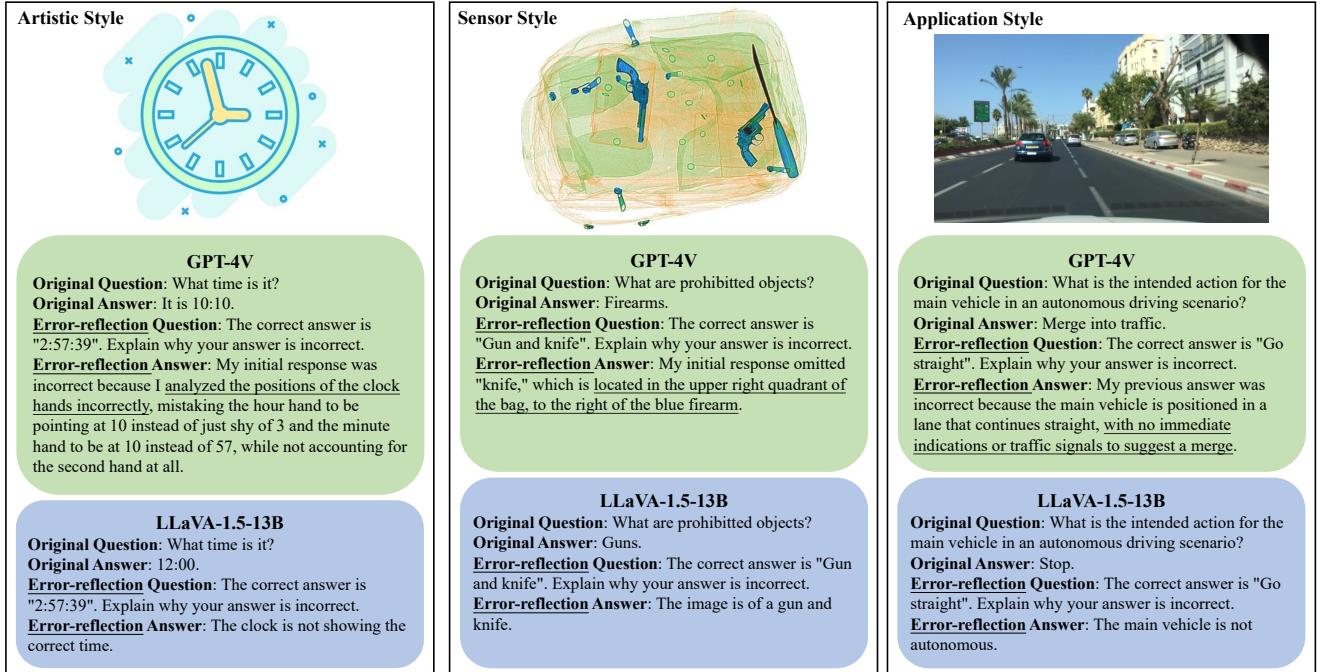


Figure 3. Error-reflection capability comparisons: Unlike LLaVA-1.5-13B, GPT-4V possesses the ability to analyze its errors when provided with the correct answers.

13B [38] by introducing more training tricks and data, and LLaVA-1.5-13B outperform LLaVA-13B by a clear margin on the ‘Common’ and other artistic styles. This comparison reveals the importance of data and training strategy as LLaVA-1.5-13B and LLaVA-13B use the same visual encoder and language encoder. However, we also find that a better performance on ‘Common’ does not necessarily indicate a better performance on other styles. For example, compared with InstructBLIP-7B, InstructBLIP-13B demonstrates better performance across five artistic-style benchmarks in terms of average accuracy, though it shows comparatively lower performance in the ‘Common’ style. Therefore, such results convey that it is important to conduct comprehensive evaluations to fully validate an LMM’s

capability rather than only focus on the ‘Common’ style, which shows the importance of our proposed benchmark.

#### 4.2.2 Cross-sensor capability

We benchmark the LMM models’ cross-sensor capability in Table 2. Because GPT-4V denied to provide medical diagnostics or identify potential abnormalities in medical images like H-Xray/MRI/CT, GPT-4V is only evaluated on the Infrared and L-Xray sensor data, and achieves the results are 55.0% and 50.0% respectively, which are far lower than the ‘Common’ style (81.5%). Meanwhile, other LMM models’ performance on images captured by X-ray, MRI, CT, and Infrared dropped significantly compared to the performance on ‘Common’ captured by the RGB camera sensor. There-

Table 4. Results of our proposed SPE on the BenchLMM benchmark

LMMs	Cartoon	Handmake	Painting	Sketch	Tattoo	Average
LLaVA-1.5-13B [37] + SPE (Ours)	62.0 <b>67.6</b>	56.6 <b>57.7</b>	57.6 <b>67.9</b>	63.3 <b>68.8</b>	57.0 <b>63.5</b>	59.3 <b>65.1 (+5.8)</b>
InstructBLIP-13B [14] + SPE (Ours)	59.0 <b>61.1</b>	57.1 <b>62.3</b>	59.3 <b>64.8</b>	57.5 <b>60.9</b>	61.4 <b>63.3</b>	58.9 <b>62.5 (+3.6)</b>
LMMs	Infrared	L-Xray	H-Xray	MRI	CT	Average
LLaVA-1.5-13B [37] + SPE (Ours)	52.1 <b>58.7</b>	51.5 <b>62.8</b>	55.9 <b>56.2</b>	33.7 33.6	43.1 <b>47.7</b>	47.3 <b>51.8 (+4.5)</b>
InstructBLIP-13B [14] + SPE (Ours)	40.8 <b>43.5</b>	24.6 <b>24.8</b>	56.6 <b>57.7</b>	20.9 <b>25.0</b>	43.4 42.6	37.3 <b>38.7 (+1.4)</b>
LMMs	Remote Sensing	Autonomous Driving	Domestic Robot	Open Game	Defect Detection	Average
LLaVA-1.5-13B [37] + SPE (Ours)	65.6 <b>68.0</b>	28.6 <b>29.4</b>	49.0 <b>51.0</b>	21.4 <b>22.7</b>	60.3 <b>61.7</b>	45.2 <b>46.6 (+1.4)</b>
InstructBLIP-13B [14] + SPE (Ours)	<b>63.6</b>	32.6	<b>32.0</b>	<b>19.5</b>	47.0	38.9
	63.5	<b>38.8</b>	31.0	19.0	<b>49.4</b>	<b>40.3 (+1.4)</b>

fore, the intrinsic difference in the imaging process between RGB cameras and other sensors is also a crucial challenge to LMMs.

With respect to the MRI, H-Xray, and CT images, the questions may require specific medical knowledge to answer, and it may be argued that the performance drop is due to the knowledge gap. It is also observed from the L-Xray, where the objects (*i.e.*, knife, gun, hammer, etc.) can be recognized with only common knowledge, but the LMMs’ performance on the L-Xray data is significantly lower than ‘Common’, which further validates the challenge brought by the sensor difference.

#### 4.2.3 Cross-application capability

In Table 3, we evaluate the different LMMs on various applications, finding GPT-4V generally outperforms others, except in autonomous driving where its accuracy drops to 35.8%. This decrease is likely due to the challenges in the Autonomous Driving dataset, such as recognizing low-resolution traffic signs and requiring specific driving knowledge, where both GPT-4 and open-source LMMs have poor performance. We also observe that LLaVA-1.5-13B achieves comparable performance as GPT-4V on the Remote Sensing and Defect Detection datasets, which seems that they have comparable cross-task capability. However, we surprisingly find that LLaVA-1.5-13B merely achieves 21.4% but GPT-4V achieves up to 74.0%. As GPT-4V is close-source, the technical details are not disclosed and we conjecture that the training data for GPT-4V may involve data similar to the Open Game dataset.

#### 4.2.4 Summary of the BenchLMM

As we can summarize from the above experiments, 1) existing LMMs generally suffer performance degradation when processing cross-style images. 2) Achieving better performance on the ‘Common’ data, which is of majority, does not guarantee better performance on other styles. 3) It is necessary to conduct a comprehensive evaluation of different dimensions because in some tasks, such as medical or security, the error information will lead to a significant loss.

#### 4.2.5 Error-reflection Capability

In order to understand the LMMs further, we let the LMM do an error-reflection that uses itself to parse the error information. When a model gives wrong answers, we warp up the conversation: the previous question input and predicted answer output. Specifically, we add “The correct answer is [GT Answer]. Explain why your answer is incorrect. Use a single sentence.” to the conversation, which will be the new input of LMMs.

Although both GPT-4V and LLaVA-1.5-13B produce incorrect answers, GPT-4V demonstrates the ability to reason the correct answer through a detailed inference process. As illustrated in Figure 3, GPT-4V can offer insightful analysis beyond the ground-truth answer, including the positions of clock hands for estimating time, the location of prohibited objects, and the traffic signals for driving planning.

By contrast, the error-reflection capability of LLaVA-1.5-13B is inferior to that of GPT-4V, as it merely rephrases the information from the correct answers provided. Therefore, we reveal that another important capability of a large multimodal model is whether the model can reflect how to derive the correct answer. Studying such capability is use-

Table 5. Results of our proposed SPE on style-transfer images

LMMs	Cartoon	Handmake	Painting	Sketch	Tattoo	Average
LLaVA-1.5-13B [37] + SPE (Ours)	62.1 <b>67.4</b>	61.1 <b>62.1</b>	52.4 <b>59.4</b>	68.8 <b>69.4</b>	41.2 <b>45.5</b>	57.1 <b>60.8 (+3.7)</b>
InstructBLIP-13B [14] + SPE (Ours)	69.3 <b>73.0</b>	70.8 <b>71.6</b>	61.3 <b>64.2</b>	53.8 <b>62.1</b>	65.0 <b>65.3</b>	64.0 <b>67.2 (+3.2)</b>

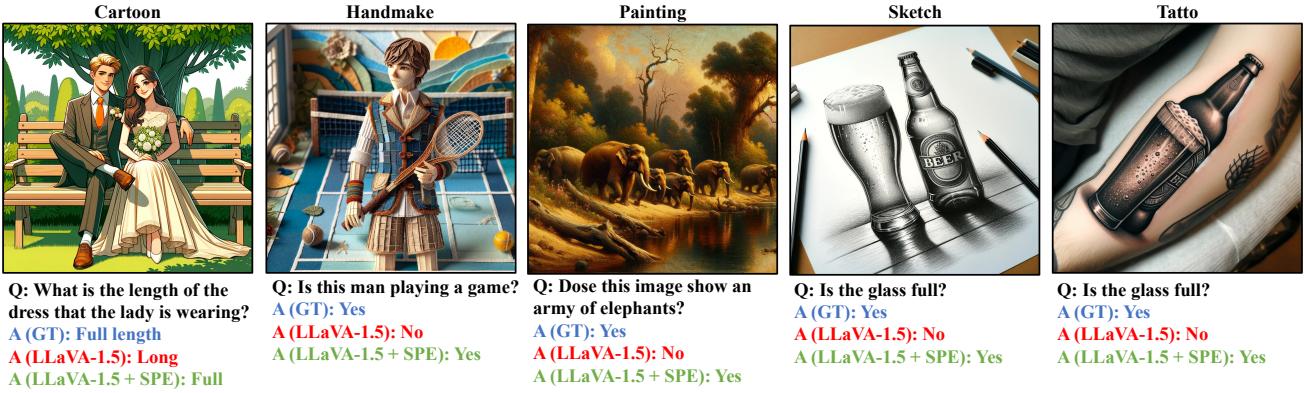


Figure 4. Qualitative comparisons between LLaVA-1.5-13B and LLaVA-1.5-13B + SPE (Ours) on style-transfer images.

ful to common users. For example, when users use LMMs for self-learning, the user can provide reference answers and LMMs can provide a more detailed inference process, which can improve the learning effect. Moreover, the answer of GPT-4V can also be useful materials used to train open-source LMMs in the future.

#### 4.3. Experiments of Style Prompt Enhancement

In this section, we aim to study if a large multimodal model can be enhanced by prompting the model to estimate the style first and then answer the question, based on which we propose Style Prompt Enhancement (SPE) method and conduct experiments with our BenchLMM. Specifically, we examined our SPE with LLaVA-1.5-13B [37] and InstructBLIP-13B [14] because these two models are representative with better average performance than MiniGPT4 [69] and Otter [32].

We conduct experiments on our BenchLMM with the data in different artistic, sensor, and application styles, and the results are presented in Table 4. Our proposed SPE can benefit the LMMs by achieving general improvement over different styles, which means that reasoning the style first is an effective prompting method to improve LMMs' visual reasoning capability. Besides, in the experiments of crossing to different applications (Remote Sensing, Autonomous Driving, etc.), we observe that its benefit is smaller than others. For example, LLaVA-1.5-13B gains 5.8% and 4.5% improvement in the experiments of cross-artistic and cross-sensor styles, while the improvement is merely 1.4% when crossing applications. We conjecture that specific application knowledge could be inferred from only the style,

which limits the benefits of SPE. Moreover, we observe that LLaVA-1.5-13 generally obtains more benefits than InstructBLIP-13B, which means that the stronger the LMM, and more benefits from SPE, which is because a stronger LMM has stronger inference capability and can better utilize prior information.

We employ the latest commercial style transfer model, DALL-E3 [48], to transfer the ‘Common’ style images used in Table 1 to different artistic styles (*i.e.*, cartoon, handmake, painting, sketch, and tattoo). In Figure 4, we provide visual examples of images in these different styles. These images share similar structures and layouts between different styles. As shown in the Table 5, the performance of the transferred artistic styles also shows significant degradation, which further validates that LMMs suffer from difficulties in recognizing images of different styles. Furthermore, our SPE significantly enhances the performance of LMMs in this challenging task.

#### 5. Conclusion

In this work, we introduce BenchLMM, a novel benchmark for quantitatively evaluating large multimodal models (LMMs) across varied visual distribution shifts, including artistic, sensor, and application styles. Our evaluations on numerous existing LMMs highlight two key findings: 1) LMMs typically under-perform with non-common image styles. 2) Our style prompt enhancement approach, inspired by human perception of art, significantly improves LMMs' visual reasoning without extra fine-tuning. Additionally, our error-reflection study shows that stronger LMMs can

self-diagnose errors by providing insights not given by humans, whereas less capable models merely restate correct answers without new insights. Our benchmark offers a comprehensive tool for LMM research, emphasizing the importance of error-reflection capabilities in future developments.

**Acknowledgement.** This research is supported in part by the Rapid-Rich Object Search (ROSE) Lab of Nanyang Technological University and the NTU-PKU Joint Research Institute (a collaboration between NTU and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation). We are deeply grateful to Yaohang Li from the University of Technology Sydney for his invaluable assistance in conducting the experiments, and to Jingpu Yang, Helin Wang, Zihui Cui, Yushan Jiang, Fengxian Ji, and Yuxiao Hang from NLULab@NEUQ (Northeastern University at Qinhuangdao, China) for their meticulous efforts in annotating the dataset. We also would like to thank Prof. Miao Fang (PI of NLULab@NEUQ) for insightful suggestion during discussion on this project.

## References

- [1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. [2](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. [2](#)
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [2](#)
- [4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. [5](#)
- [5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. [2](#)
- [6] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. [4](#)
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. [2](#)
- [8] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998, 2019. [2](#)
- [9] Qingxing Cao, Wentao Wan, Keze Wang, Xiaodan Liang, and Liang Lin. Linguistically routing capsule network for out-of-distribution visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1614–1623, 2021. [2](#)
- [10] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. [1, 2, 5, 6](#)
- [11] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*, 2023. [4](#)
- [12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. [2](#)
- [13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. [2](#)
- [14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. [1, 2, 5, 6, 7, 8](#)
- [15] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021. [2](#)
- [16] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukae: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2022. [2](#)
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. [1, 2, 3, 5](#)
- [18] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal com-

- pact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2
- [19] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022. 2
- [20] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhan. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*, 2023. 2
- [21] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2, 3, 5
- [22] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multispectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 4
- [23] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021. 2
- [24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1
- [25] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 4
- [26] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llivip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 4
- [27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2
- [28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE international conference on computer vision*, pages 2989–2998, 2017. 2
- [29] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 2
- [30] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018. 2
- [31] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 5
- [32] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1, 2, 5, 6, 8
- [33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2
- [34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 5
- [35] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021. 2
- [36] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 4
- [37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 5, 6, 7, 8
- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 5, 6
- [39] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, 2022. 4
- [40] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 5
- [41] James Manyika. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2023. 2
- [42] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6339–6350, 2023. 4
- [43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings*

- of the IEEE/cvpr conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1, 2
- [44] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307, 2017. 2
- [45] Hong Duc Nguyen, Rizhao Cai, Heng Zhao, Alex C Kot, and Bihan Wen. Towards more efficient security inspection via deep learning: a task-driven x-ray image cropping scheme. *Micromachines*, 13(4):565, 2022. 4
- [46] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 2
- [47] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658, 2019. 2
- [48] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*, 2020. 8
- [49] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 2
- [50] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2
- [51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2
- [52] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Logical implications for visual question answering consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6725–6735, 2023. 2
- [53] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1417–1427, 2021. 2
- [54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2
- [55] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Es-lami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [57] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Be- longie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liang- pei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 4
- [58] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 2
- [59] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021. 2
- [60] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 1, 2, 5, 6
- [61] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2
- [62] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2
- [63] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020. 4
- [64] Samuel Yu, Peter Wu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Pacs: A dataset for physical audiovisual commonsense reasoning. In *European Conference on Computer Vision*, pages 292–309. Springer, 2022. 2
- [65] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1, 2, 5
- [66] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 2
- [67] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2
- [68] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for

- visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2074–2084, 2021. 2
- [69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1, 2, 5, 6, 8