DYSCA: A DYNAMIC AND SCALABLE BENCHMARK FOR EVALUATING PERCEPTION ABILITY OF LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Currently many benchmarks have been proposed to evaluate the perception ability of the Large Vision-Language Models (LVLMs). However, most benchmarks conduct questions by selecting images from existing datasets, resulting in the potential data leakage. Besides, these benchmarks merely focus on evaluating LVLMs on the realistic style images and clean scenarios, leaving the multi-stylized images and noisy scenarios unexplored. In response to these challenges, we propose a dynamic and scalable benchmark named Dysca for evaluating LVLMs by leveraging synthesis images. Specifically, we leverage Stable Diffusion and design a rule-based method to dynamically generate novel images, questions and the corresponding answers. We consider 51 kinds of image styles and evaluate the perception capability in 20 subtasks. Moreover, we conduct evaluations under 4 scenarios (i.e., Clean, Corruption, Print Attacking and Adversarial Attacking) and 3 question types (i.e., Multi-choices, True-or-false and Free-form). Thanks to the generative paradigm, Dysca serves as a scalable benchmark for easily adding new subtasks and scenarios. A total of 24 advanced open-source LVLMs and 2 close-source LVLMs are evaluated on Dysca, revealing the drawbacks of current LVLMs. The benchmark is released in anonymous github page `https://github.com/Benchmark-Dysca/Dysca`.

1 INTRODUCTION

Recent years have witnessed the great success of the Large Vision-Language Models (LVLMs) (Li et al., 2023d; Zhu et al., 2023; Dai et al., 2023; Liu et al., 2023b; Li et al., 2023a; Chen et al., 2023b; Zhang et al., 2023; Su et al., 2023; Gong et al., 2023; Sun et al., 2023b). These models leverage the powerful Large Language Models (LLMs) (Chung et al., 2022; OpenAI, 2022; Touvron et al., 2023; OpenAI, 2023; FastChat, 2023) as their brain and incorporate the state-of-the-art visual encoders (Radford et al., 2021; Fang et al., 2023; Dosovitskiy et al., 2020) as their eyes. Thanks to the alignment of visual feature with textual space and the development of visual instruction tuning techniques (Liu et al., 2023b), LVLMs showcase the impressive capability in terms of visual scene comprehension and multimodal instruction-following.

In order to comprehensively evaluate the capabilities of LVLMs, many benchmarks have been purposed (Antol et al., 2015a; Singh et al., 2019; Xu et al., 2023; Shao et al., 2023; Li et al., 2023c;b; Fu et al., 2023; Bai et al., 2023b; Yu et al., 2023; Yang et al., 2023b; Chen et al., 2024), where we categorize the current benchmarks into three types (Fu et al., 2023). The first type is the classical benchmarks, such as COCO Caption (Chen et al., 2015) and VQA (Antol et al., 2015a; Goyal et al., 2017; Marino et al., 2019). Although these benchmarks provide high-quality evaluation data, they also have notable limitations. On the one hand, they are inadequate for measuring the fine-grained capabilities of current LVLMs, offering the limited insightful feedback for the future improvement. On the other hand, since these classical benchmarks have been available as the open-source test data for a long time, it is hard to prevent the data leakage problem. The second type of benchmarks evaluate the LVLMs through a subjective manner (Yang et al., 2023b; Wu et al., 2023). Although the benchmarks reveal the insightful drawbacks of current models, their data scale is limited (i.e., less than 200 annotations) and they require manual evaluation by experts. The third type is built for objectively evaluating current LVLMs and the comparison between them are shown in Tab. 1. They provide an objective and automatic evaluation manner, giving the fine-grained evaluation for the LVLMs. However, these benchmarks conduct Vision-language QAs by selecting images from

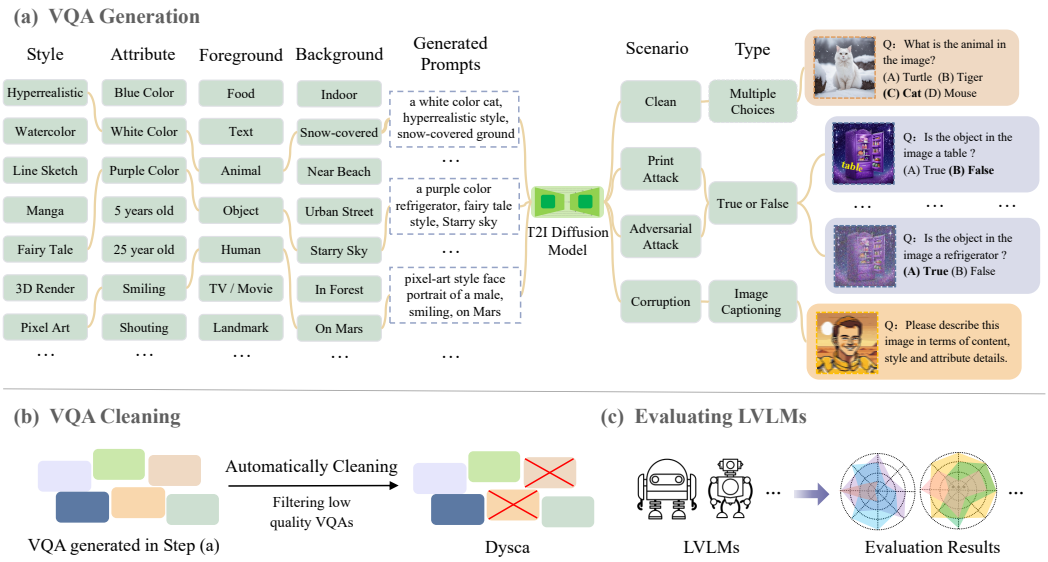

Figure 1: Overview of the automatic pipeline for generating Vision-language QAs, cleaning Vision-language QAs and evaluating LVLMs. ***(a)*** We first constructs prompts in terms of content, style and background, leveraging the Text-to-Image (T2I) diffusion model (e.g., SDXL (Podell et al., 2023)) to synthesis images to be asked. Then based on the scenarios and the question type, we post-process the synthesis images and generate the specific textual questions, respectively. ***(b)*** We further filter out low quality Vision-language QAs by utilizing trained models to form the final Dysca. ***(c)*** Finally, we evaluate LVLMs on our Dysca and feedback the fine-grained evaluation results.

existing dataset and annotate the textual questions. Although they claim that the questions are re-annotated, the previous work (Chen et al., 2024) has demonstrated that these benchmarks have Models unintentionally leaked into the training data of LLMs and LVLMs. Besides, most benchmarks focus on evaluating LVLMs in the realistic images and clean scenarios, leaving the multi-stylized images and noisy scenarios unexplored. While some works like MMCBench (Zhang et al., 2024b) and Typographic Dataset (Cheng et al., 2024) have investigated the robustness of LVLMs with corrupted and print-attacked images, respectively, they have not explored the effect of these noisy images on various perceptual tasks.

In this paper, aiming to address these challenges above, we propose Dysca which is a dynamic and scalable benchmark for evaluating the perception ability of LVLMs via various subtasks and scenarios. Inspired by the prior evaluation works for LLMs (Liang et al., 2023), we investigate on whether we could leverage the large-scale synthesized images for evaluating LVLMs. We display the overview of our pipeline in Fig. 1. Specifically, we leverage Stable Diffusion and design a rule-based method to dynamically generate novel images, questions and corresponding answers. We decouple the prompt into 4 part, i.e., attribute, foreground, style and background, and design pre-defined templates to dynamically generate prompts, as displayed in Fig. 3. Then we utilize the state-of-the-art text-to-image diffusion models (e.g., SDXL (Podell et al., 2023)) to generate the corresponding images. Since we already know the main information of the images through prompts, we easily generate question-answer textual pairs by the rule-based method. After that, in order to obtain the high quality Vision-language QAs, we employ CLIP (Radford et al., 2021) to perform data cleaning on the generated Vision-language QA pairs. Dysca focuses on assessing the fine-grained perception abilities, including recognizing human, animal, object, landmark, etc. Dysca evaluates LVLMs with **20 perceptual subtasks**, containing a total number of **51 different artistic styles**. Besides, to evaluate the robustness of the models across different scenarios and question types, we construct **4 testing scenarios** (clean, corruption, print attacking and adversarial attacking) and **3 question types** (multi-choices, true-or-false and free-form questions).

Compared to previous works in Tab. 1, we provide an end-to-end process from image to Vision-QA generation. The approach significantly reduces annotation costs compared to manually labeling images (e.g., MME (Fu et al., 2023)) while achieving the correctness for evaluating LVLMs. It also avoids the risk of hallucinate annotations that may occur when using ChatGPT for labeling based on image prompts (e.g., JourneyDB (Sun et al., 2023a)). This novel pipeline enables us to create a

Table 1: Comparisons between existing LVLM benchmarks. '✓✗' indicates that the benchmarks include both newly collected images / annotations and images / annotations gathered from existing datasets. '*' The scale of our released benchmark is 617K, however Dysca is able to generate unlimited data to be tested.

Benchmark	#Evaluation Data Scale	#Perceptual Tasks	Automatic Annotation	Collecting from Existing Datasets	Question Type	Automatic Evaluation
LLaVA-Bench	0.15K	-	✗	✓✗	Free-form	✓
MME	2.3K	10	✗	✓✗	True-or-false	✓
LVLM-eHub	-	3	✓	✗	Free-form	✗
tiny-LVLM-eHub	2.1K	3	✓	✗	Free-form	✓
SEED-Bench	19K	8	✓✗	✗	Multi-choices	✓
MMBench	2.9K	12	✗	✓✗	Multi-choices	✓
TouchStone	0.9K	10	✗	✓	Free-form	✓
REFORM-EVAL	50K	7	✓	✗	Multi-choices	✓
MM-BigBench	30K	6	✓	✗	Multi-choices	✓
MM-VET	0.2K	4	✓✗	✓✗	Free-form	✓
MLLM-Bench	0.42K	7	✗	✓✗	Free-form	✓
SEED-Bench2	24K	10	✓✗	✗	Multi-choices	✓
BenchLMM	2.4K	15	✗	✗	Free-form	✓
JourneyDB	5.4K	2	✓	✓	Free-form Multi-choices	✓
Dysca (Ours)	617K*	20	✓	✓	Free-form Multi-choices True-or-false	✓

benchmark that is easily scalable and adaptable for incorporating new subtasks and scenarios. In the end, Dysca consists of **617K** Vision-language QA pairs ($\times$20 larger than MM-BigBench (Yang et al., 2023a) and $\times$25 larger than Seed-Bench2 (Li et al., 2023b), with the most comprehensive evaluation perspectives and scenarios.

In summary, our work makes the following key contributions:

- **Dynamic and Scalable Benchmark:** We propose Dysca, a benchmark that is able to dynamically generate the test data that users need and is easily to scale up to to new subtasks and scenarios.
- **Multi-grained Perceptual Subtasks and Multi-scenarios:** Dysca evaluates the 20 perceptual subtasks performance of 26 mainstream LVLMs, including GPT-4o (Ope) and Gemini-1.5-Pro (Team et al., 2024), under 4 image scenarios (i.e., clean, corruption, print attacking and adversarial attacking) and 3 question types (i.e., multi-choices, true-or-false and free-form questions).
- **Analysis and Observations:** We demonstrate for the first time that evaluating LVLMs using large-scale synthetic data is valid. Experiments show the strong correlation coefficient between our evaluation rankings and the rankings obtained from non-synthetic benchmarks. The evaluation results also reveal the weakness of current LVLMs when facing different question types, image styles and image scenarios.

2 Related Works

2.1 Large Vision-Language Models

The landscape of Large Vision-Language Models (LVLMs) has been significantly shaped by the pioneering success of Large Language Models (LLMs) such as GPTs (Radford et al., 2019; Brown et al., 2020; Ouyang et al., 2022) and LLaMA (Touvron et al., 2023), catalyzing advancements in multimodal content understanding and generation (Zhang et al., 2024a), including intricate tasks like image-text comprehension. At the forefront of these developments, BLIP-2 (Li et al., 2023d) introduces a lightweight Q-Former (Li et al., 2023d) that facilitates alignment between textual and visual representations through a cross-attention mechanism (Li et al., 2023d). InstructBLIP (Dai et al., 2023) takes a step further by incorporating textual instructions into the Q-Former, which

significantly improves performance. LLAVA (Liu et al., 2023b) employs GPT-4 (OpenAI, 2023) to transform data into multimodal instruction-following data and uses CLIP (Radford et al., 2021) and LLAMA (Touvron et al., 2023) for fine-tuning instructions, achieving advanced performance. LLAVA-1.5 (Liu et al., 2023a) extends this paradigm by integrating MLP projection and introducing academic task-specific Vision-language QA data. Recently, models like Otter (Li et al., 2023a), MiniGPT-4 (Zhu et al., 2023), Qwen-VL-Chat (Bai et al., 2023a) and XComposer-VL (Zhang et al., 2023) further unleash the cross-modal understanding capabilities of LVLMs. Besides, many powerful closed-source LVLMs, including Gemini-1.5-Pro (Team et al., 2024) and GPT-4o (Ope), have publicly released their APIs, promoting the development of downstream applications.

2.2 Benchmarks for LVLMs

The great progress of LVLMs triggers the development of benchmarks for evaluating these models, where we divide previous benchmarks into three categories. The first type is the classical benchmarks which focuses on evaluating LVLMs abilities via image caption (Lin et al., 2014) and VQA (Antol et al., 2015b;a). However, these benchmarks cannot provide the fine-grained feedback on how to improve the models. Besides, since these benchmarks have been the public resources for a long time, it is hardly to guarantee that the LVLMs have not use them for training. The second type subjectively evaluates LVLMs by experts (Yang et al., 2023b; Wu et al., 2023). Although these benchmarks reveal the insightful feedback of the LVLMs, their scale is limited (i.e., less than 200 annotations). The subjective manner also makes the evaluation expensive and hardly to expand the scale of the benchmarks.

The third type (Liu et al., 2023b; Fu et al., 2023; Xu et al., 2023; Shao et al., 2023; Li et al., 2023c;b; Liu et al., 2023c; Bai et al., 2023b; Li et al., 2023f; Yang et al., 2023a; Yu et al., 2023; Ge et al., 2023; Cai et al., 2023; Chen et al., 2024; Liu et al., 2024) focuses on evaluating LVLMs in an objective and large-scaled manner, where we list the detailed information of them in the Tab. 1. Some of them have been adopted by the community (Contributors, 2023) as the standard benchmarks for evaluating LVLMs (OpenAI, 2022; Li et al., 2023a; Zhang et al., 2023), like MME (Fu et al., 2023) and MMBench (Liu et al., 2023c). These benchmarks evaluate models through the objective answer types and most of them leverage the automatic annotation and evaluation manner for revealing the fine-grained drawbacks of current LVLMs. However, the previous benchmarks primarily concentrate on evaluating LVLMs using realistic images and clean scenario, leaving multi-stylized images and noisy scenarios unexplored. Moreover, many of them conduct QA by selecting images from publicly available datasets (e.g., (Lin et al., 2014; Russakovsky et al., 2014)). While they state that the questions have been re-annotated, they cannot guarantee that the LVLMs have not seen the image during training stage. The previous work (Chen et al., 2024) has proved that these benchmarks have unintentionally leaked into the training data of LLMs and LVLMs. One possible way to solve data leakage is using novel but synthesis images, where JourneyDB (Sun et al., 2023a) is the first work aiming to leverage synthesis images to evaluate current LVLMs. The prompts and the corresponding images are downloaded from Midjourney (mid) and ChatGPT (OpenAI, 2022) is leveraged to label the images. However, JourneyDB is a top-down framework where the number of images is fixed. Besides, the ChatGPT labeling may cause hallucinate annotations, leading to the unreliable evaluation results. Although 40 annotators have involved to clean the data, the data cleaning cost are expensive and it limits the data scale. In contrast, our Dysca serves as the bottom-up framework, allowing for dynamic and scalable generation for both images and evaluation questions. The rule-based question generation method also makes the annotations more accuracy. Besides, Dysca contains 20 subtasks which is more comprehensive than JourneyDB.

3 Dysca

3.1 Overview of Our Pipeline

The overview of our pipeline is shown in Fig. 1, containing data generation, data cleaning and LVLMs evaluation. For the data generation, our Dysca benchmark consists of four dimensions, i.e., (M, P, I, Q), where M means "Metadata", P means "Prompt", I means "Image" and Q means "Question-answer pair". We further decouple the metadata M into 4 parts, i.e., "style", "attribute", "foreground" and "background", and the combination of the four parts constitute the image prompts P. Then, given the prompt P and the selected scenario, we leverage the Text-to-Image (T2I) diffusion model (e.g., SDXL (Podell et al., 2023)) to synthesis image I and add the specific perturbation to the image I. After that, since the prompt already includes the question angle and the corresponding

Table 2: Key statistics of Dysca.

Statistic	#Number
Total questions	617K
- Clean	156K (25.2%)
- Print attacking	149K (24.1%)
- Adversarial attacking	156K (25.2%)
- Corruption	156K (25.2%)
Question type	
- Multi-choices	251K (40.6%)
- True-or-false	250K (40.5%)
- Free-form	116K (18.8%)
Image resolution	1024*1024
Unique number of images	289K
Unique number of questions	162K
Unique number of answers	31K
Average question length	37.8
Average answer length	2.7
Average choice number	3.0

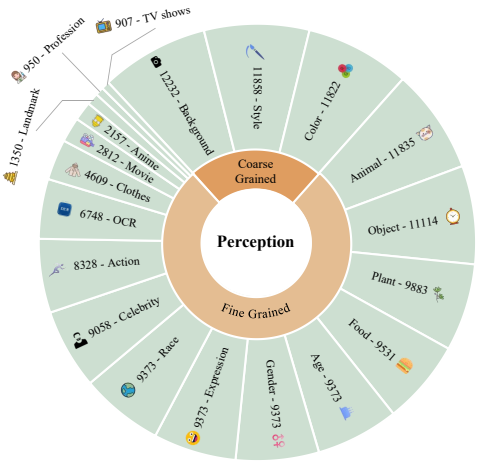

Figure 2: Overview of the dataset distribution of 20 perceptual tasks. The number in each subtask shows the corresponding amount of their annotation.

answer, we construct a rule-based approach to generate the Q. Three types of questions are considered, i.e., multi-choice, true-or-false and free-form. Multi-choice and true-or-false questions utilize a closed-ended manner to assess LVLMs, while free-form questions employ an open-ended manner through image captioning for evaluation. For the data cleaning, considering that the T2I diffusion model may generate unsuccessful outcomes, we then use CLIP (Radford et al., 2021) and PP-OCRv3 (Li et al., 2022) to automatically clean the whole dataset to obtain the final Dysca. Finally, we evaluate 14 open-sourced LVLMs and 2 closed-source LVLMs on our proposed Dysca.

3.2 Perceptual Tasks

Evaluation dimensions. Perception is one of the most fundamental capabilities of LVLMs and previous works (Fu et al., 2023) have shown that the lack of perceptual ability may result in hallucination (Li et al., 2023e). In order to comprehensively evaluate LVLMs' perception capability, we collect and organize existing sub-dimensions from current evaluation datasets, resulting in a total of 20 assessment dimensions where we show all the subtasks and the corresponding amount of their annotation in the Fig. 2. We investigate on two types of perception dimensions, i.e., coarse-grained and fine-grained perception. Coarse-grained perception involves recognizing the style, background and color of images. Fine-grained perception involves recognizing the animal, object, plant, food, age, gender, expression, race, celebrity, action, text, clothes, movie, anime, landmark, profession and TV shows.

Data sources. For each perceptual subtask, we collect the textual data first to construct the metadata M. For the TV shows, Anime and Movie, we select the titles from the rating list of IMDb[1] based on the number of user reviews. For the styles, we utilize the style lists collected from the community[2] and remove those which have strong reflect on the image content like "architectural style" and "Pokemon style". Note that the style list does not include the style prompt associated with a particular artist's name. Besides, for the remaining contents, we select them from the label of current dataset (e.g., ImageNet (Russakovsky et al., 2014)). All the selected textual data above constitute the metadata M. We provide the detailed information of the metadata in the Appendix **??**.

3.3 Construction of Questions & Answers

Recall that the data generation for Dysca benchmark consists of four dimensions, i.e., (M, P, I, Q), denoting the metadata (M), prompt (P), image (I) and question-answer pairs (Q), respectively. The relationships between these parts and the process of constructing Dysca are shown in Fig. 3. The metadata M is the core of the whole Dysca, containing all the information for generating P, I and Q. The metadata M consists of foreground, attribute, background and style, and these information guide the generation of the prompt (P) through pre-designed templates. Then, we utilize the T2I diffusion model to generate the corresponding image using the prompt P. For generating

[1]https://www.imdb.com/

[2]https://stable-diffusion-art.com/sdxl-styles/

Table 3: Evaluation results on blind setting and 4 scenarios, where the darker colors represent higher scores. The top 1 result on each scenario are **bolded** and the value in brackets is the relative values with respect to the ones in the clean scenario. "PrintAtt" and "AdverAtt" means "Print Attacking" and "Adversarial Attacking", respectively. "*": the model is under white-box setting.

Model	LLM	Visual Encoder	Blind	Scenarios			
				Clean	Corruption	AdverAtt	PrintAtt
MiniGPT-4	Vicuna-7B	EVA-CLIP ViT-G	35.37	41.38	42.3 (+0.92)	34.42 (-6.96)	42.71 (+1.33)
MiniGPT-4	Vicuna-13B	EVA-CLIP ViT-G	35.21	50.17	49.63 (-0.54)	31.77 (-18.40)	47.55 (-2.62)
MiniGPT-4	$LLaMA_2$	EVA-CLIP ViT-G	34.77	56.61	55.7 (-0.91)	33.55 (-23.06)	49.78 (-6.83)
MiniGPT-2	$LLaMA_2$	EVA-CLIP ViT-G	35.28	58.46	58.06 (-0.40)	56.62 (-1.84)	52.96 (-5.50)
BLIP2	Flan-T5-XL	EVA-CLIP ViT-G	35.35	65.3	66.09 (+0.79)	32.55 (-32.75)	57.01 (-8.29)
BLIP2	OPT-3B	EVA-CLIP ViT-G	34.99	39.54	40.29 (+0.75)	30.62 (-8.92)	37.26 (-2.28)
BLIP2	OPT-7B	EVA-CLIP ViT-G	35.21	39.55	41.12 (+1.57)	31.76 (-7.79)	38.82 (-0.73)
InstructBLIP	Vicuna-7B	EVA-CLIP ViT-G	35.14	67.54	67.01 (-0.53)	34.42 (-33.12)	52.58 (-14.96)
InstructBLIP	Vicuna-13B	EVA-CLIP ViT-G	34.37	64.89	64.68 (-0.21)	31.77 (-33.12)	53.53 (-11.36)
InstructBLIP	Flan-T5-XL	EVA-CLIP ViT-G	34.51	66.54	67.58 (+1.04)	32.95 (-33.59)*	52.09 (-14.45)
InstructBLIP	Flan-T5-XXL	EVA-CLIP ViT-G	34.82	68.65	69.79 (+1.14)	32.95 (-35.70)	57.73 (-10.92)
LLava-1.5	Vicuna-7B	CLIP ViT-L	34.63	51.27	51.7 (+0.43)	49.62 (-1.65)	47.27 (-4.00)
LLava-1.5	Vicuna-13B	CLIP ViT-L	35.21	59.23	59.58 (+0.35)	56.87 (-2.36)	51.69 (-7.54)
Otter	LLaMA-7B	CLIP ViT-L	35.19	54.9	56.02 (+1.12)	51.42 (-3.48)	37.78 (-17.12)
Shikra	LLaMA-7B	CLIP ViT-L	34.96	62.24	63.06 (+0.82)	58.78 (-3.46)	49.56 (-12.68)
Xcomposer-VL	InternLM-7B	EVA-CLIP ViT-G	32.33	71.4	72.08 (+0.68)	30.28 (-41.12)	64.71 (-6.69)
Xcomposer2-VL	InternLM2-7B	CLIP ViT-L	32.76	79.13	78.64 (-0.49)	76.6 (-2.53)	66.34 (-12.79)
Qwen-VL-Chat	Qwen-7B	OpenClip ViT-bigG	33.06	62.18	61.05 (-1.13)	59.85 (-2.33)	51.94 (-10.24)
Emu2-Chat	LLaMA-33B	EVA2-CLIP-E	35.14	63.64	62.81 (-0.83)	61.9 (-1.74)	54.82 (-8.82)
GLM-4V	GLM-4-9B-Chat	EVA2-CLIP-E	35.08	**82.09**	**81.95 (-0.14)**	**80.72 (-1.37)**	52.09 (-30.00)
MiniCPM-V2.5	Llama3-Instruct 8B	SigLIP SoViT-400m	34.99	78.75	77.41 (-1.34)	75.44 (-3.31)	60.77 (-17.98)
Yi-VL	Yi-6B-Chat	OpenClip ViT-H	35.01	75.71	74.94 (-0.77)	72.53 (-3.18)	64.97 (-10.74)
mPLUG-Owl-2	LLaMA-7B	CLIP ViT-L	35.03	74.09	72.85 (-1.24)	69.76 (-4.33)	**72.85 (-1.24)**
Phi-3-Vision	Phi-3	CLIP ViT-L	34.74	73.23	72.11 (-1.12)	69.66 (-3.57)	57.78 (-15.45)
GPT-4o	/	/	35.02	75.69	75.52 (-0.17)	73.47 (-2.22)	56.34 (-19.35)
Gemini-1.5-Pro	/	/	34.55	77.79	77.12 (-0.67)	75.89 (-1.90)	61.05 (-16.74)

the image with specific text on it for the OCR subtask, we leverage TextDiffusion2 (Chen et al., 2023a), which is the state-of-the-art text rendering method. For the rest of images, we leverage Stable Diffusion XL (Podell et al., 2023). Subsequently, based on the different question types we select, i.e., multi-choices, true-or-false and free-form, we generate the corresponding VQA pairs in Dysca.

Besides, in order to evaluate the model performance under various scenarios, we conduct experiments on 4 scenarios, i.e., clean, corruption, print attacking and adversarial attacking. For the print attacking, followed by (Cheng et al., 2024), we add the deceptive text on the image, where the text is a wrong option. Besides, to comprehensively evaluate the performance of LVLMs under corruption scenario, we add more typographic factors to original settings (i.e., different font orientations and font positions). For the adversarial attacking, we leverage PGD (Madry et al., 2017) to generate the adversarial image. We use InstructBLIP (Dai et al., 2023) as the proxy model and regard others as the black box models. The reason why we choose InstructBLIP is that it has shown superior performance in clean scenario. Besides, the black-box setting better reflects the robustness of the models when they face the real-world adversarial attacks. For the corruption, we leverage the image corruption methods collected from (Zhang et al., 2024b). We remove some hard corruptions as they significantly impact the quality of the image, leading to human failure in judging the style and content of the image. The detailed examples are shown in Appendix **??**.

Data Clean. To ensure the quality of Dysca, four steps are adopt: 1) First, we manually remove difficult-to-generate foregrounds and attributes, along with backgrounds and styles that could heavily affect image content. We believe this process can serve as a coarse-grained method to eliminate samples that are highly likely to be generated incorrectly. 2) Then, we leverage the off-the-shelf models, i.e., PP-OCRv3 (Li et al., 2022) and CLIP-L-14 (Radford et al., 2021), to clean the data. PP-OCRv3 (Li et al., 2022) is leveraged as the filter to exclude the failure image that TextDiffusion2 (Chen et al., 2023a) generates the wrong text on the image. For the other images, we use CLIP-L-14 (Radford et al., 2021) with a threshold of 0.75 to filter out the images with low text-image consistency. We find that using 0.75 as the threshold achieves a good balance between image correctness and data scale. 3) After that, We select the top six performing models and eliminate any question-answer pairs where the models either answer incorrectly or indicate that the answer was not included among the options. we observe that nearly 100% of the samples filtered out by these models are incorrect. 4) Finally, we analyze the patterns in these incorrect samples, removing the associated vocabulary from our metadata and discarding all related samples. By meticulously refining the metadata manually and utilizing automated tools to assist in question filtering, Dysca ensures high-quality data synthesis. In the end, we filter out nearly 40% of low quality samples. The final statistics of our released

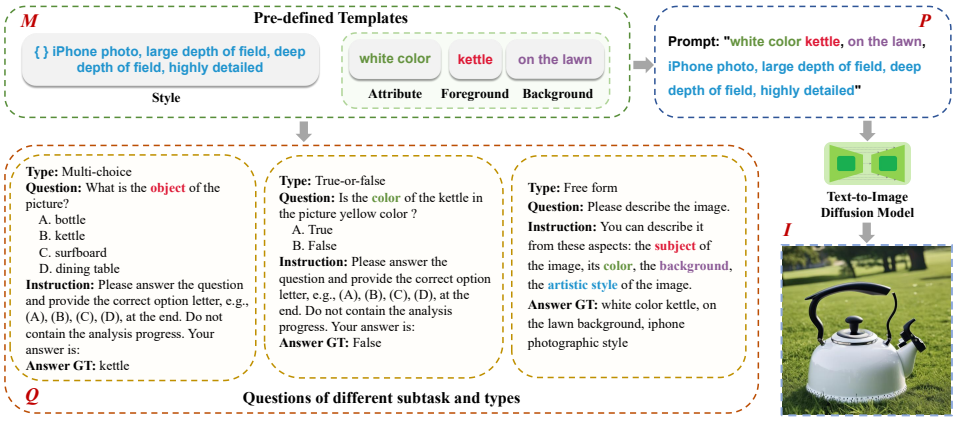

Figure 3: The process of generating the prompt (P), image (I) and question-answer pairs (Q) from the metadata (M).

Dysca are shown in Tab. 2. Note that the OCR subtask does not involve print attacking scenario as misidentifying adversarial text does not indicate poor OCR robustness of the LVLMs. Therefore, there are 7K fewer questions in the print attacking scenario. Besides, for the free-form question type, since it allows to assess the model's perception abilities across multiple subtasks at the same time, we reduce the number of free-form questions for achieving a balanced data distribution.

3.4 Evaluation Strategy

Instruction Design. We design two types of instructions to improve the instruction-following result of LVLMs. For the multi-choices and true-or-false questions, we design the questions followed by the description "Please answer the question and provide the correct option letter, e.g., (A), (B), (C), (D), at the end. Do not contain the analysis progress. Your answer is: ". For the free-form questions, recalling that the prompt P contains four part, i.e., the style, attribute, foreground and background, we instruct the model to caption these four dimensions by "Please describe the image. You can describe it from these aspects: {}", where "{}" includes the specific template we design for each part. We display the sample in the Fig. 3 and more examples can be found in the Appendix **??**.

Evaluation Metric. For the multi-choices and true-or-false questions, we use accuracy as the evaluation metric. We randomly shuffle the order of choices to prevent evaluation results from being influenced by the model's tendency towards specific choices (Zong et al., 2023). The random accuracy of the two types are equal to 25% and 50%, respectively. We use regular expressions to extract the model's answer choices. For cases where the extraction is fail, we calculate the Levenshtein distance between the answer string and the choice string, and select the option with the minimum distance as the model's answer. As our answers rarely exist in forms that can be represented in multiple ways (e.g., "six" and "6" in reasoning or counting tasks), with selectively designed question prompt and the answer pool, the answers of LVLMs can be effectively extracted. For the free-form questions, we test the model's image caption capability where the ground truth is the prompt of the image. Followed by (Xu et al., 2023), we use SentenceTransformer (Thakur et al., 2021) to compute the text similarity with prompt P and the caption output of the LVLMs. The final score of each question type is the average score of subtasks.

4 Results and Analysis

In this section, we report the evaluation results and make insightful analysis. A total of 26 LVLMs are evaluated on Dysca benchmark, including BLIP2 (Li et al., 2023d), InstructBLIP (Dai et al., 2023), LLavA (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2023), Otter (Li et al., 2023a), XComposer-VL (Zhang et al., 2023), Qwen-VL-chat (Bai et al., 2023a), Shikra (Chen et al., 2023b), Emu2-Chat (Sun et al., 2024), GLM-4V (GLM et al., 2024), MiniCPM-v2.5 (Yao et al., 2024), Yi-VL (AI et al., 2024), mPLUG-Owl-2 (Ye et al., 2023), Phi-3-Vision (Abdin et al., 2024), GPT-4o (OpenAI, 2023), Gemini-1.5-pro (Team et al., 2024). Each model is evaluated with all the 20 perception subtasks under 4 scenarios. The detailed rankings for each subtask can be found in the Appendix **??**.

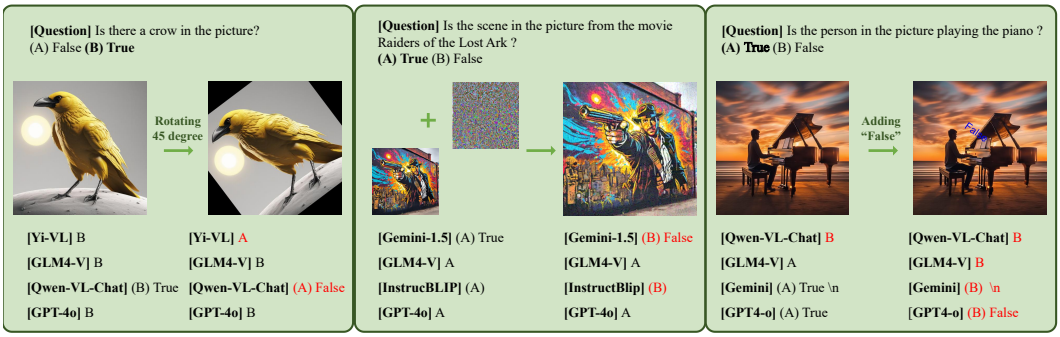

Figure 4: The failure cases for the noisy scenarios. From left to right are: corruption scenario, adversarial attacking scenario, and print attacking scenario.

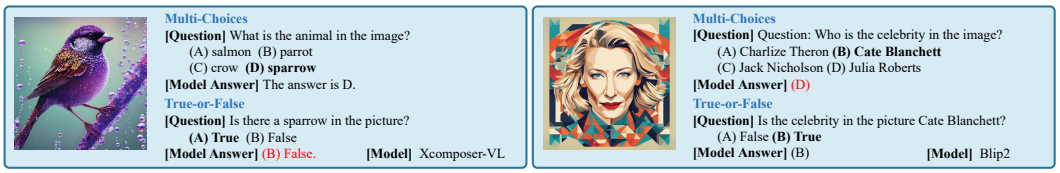

Figure 5: Models exhibit different performance when facing the same image but different question types.

4.1 Main Results

Blind Setting. We first evaluate LVLMs when only textual questions are provided. As shown in the "Blind" column of Tab. 3, all LVLMs yield consistent results on the Dysca and perform comparable to random guessing. This outcome demonstrates that the generated paradigm employed by Dysca effectively mitigates the potential impact of data leakage (Chen et al., 2024), thereby enhancing the fairness of the evaluation results. Additional comparisons can be found in Appendix **??**.

Clean Scenario. The evaluation results of various LVLMs in different perceptual subtasks under clean scenarios are presented in the "clean" column of Tab. 3. We calculate the average score of 3 question types. As can be seen, GLM-4V (GLM et al., 2024) outperforms other LVLMs, achieving top-1 performance. MiniCPM-v2.5 (Yao et al., 2024), Xcomposer2-VL (Dong et al., 2024) and Gemini-1.5-pro (Team et al., 2024) also perform well. It is evident that for the latest large models, their scores remain below 90. The results highlight that all existing LVLMs still struggle to provide accurate responses to questions formulated by Dysca, being capable of uncovering drawbacks present in existing LVLMs.

Noisy Scenarios The evaluation results of various LVLMs under noisy scenarios (i.e., corruption, print attacking and adversarial attacking) are presented in last 3 columns in Tab. 3. The value in the brackets shows the relative values with respect to the ones in the clean scenario. As can be seen, GLM-4V (GLM et al., 2024) still takes a lead on corruption and adversarial attacking scenarios. For the print attacking scenario, mPLUG-Owl-2 (Ye et al., 2023) performs the best. Here, we present a failure case sample for each of the three different scenarios on Fig. 4.

4.2 Analysis

4.2.1 Key Observations

(1) For LVLMs, the capacity of the language model plays a crucial role. When using the same visual encoder, models that adopt a language model with larger parameter sizes (e.g., Vicuna-13B models generally outperform Vicuna-7B models by 8%) or models with a stronger but different architecture (e.g., GLM-4V (GLM et al., 2024) with GLM-4-9B-Chat shows a 20% performance increase compared to Emu2 (Sun et al., 2024) with LLaMA-33B) tend to achieve better overall performance.

(2) Models exhibit performance inconsistency when dealing with multiple-choice and true-or-false question types. We present two examples in Fig. 5. In the left example, the XComposer-

VL (Zhang et al., 2023) recognizes the sparrow in the image under a multiple-choice setting but fails to identify the sparrow in the same image under a true-or-false setting. This inconsistency is also observed in other models, and we report detailed results in Appendix **??**. We hypothesize that the observed inconsistency may result from an imbalance in the training dataset, where certain question types, such as multiple-choice or true-or-false questions, are more frequently represented. This imbalance could lead the model to perform better on these overrepresented question types, while struggling with others that are less common.

(3) The perceptual performance of individual models varies significantly across different subtasks. For instance, Qwen-VL-Chat (Bai et al., 2023a) achieves 95.9% accuracy in the landmark recognition task for multiple-choice questions (4.1% below the top score), but only 68.12% accuracy in the TV show recognition task (28.94% below the top score). These results suggest that Qwen-VL-Chat may require further fine-tuning in certain recognition tasks, especially in underperforming areas such as TV show identification. Analyzing model performance across different subtasks can provide insights for targeted improvements. Detailed results are provided in Appendix **??**.

(4) **Each model shows robustness in the corruption scenario, but experiences significant degradation in both attack scenarios.** In the image corruption scenario, all models demonstrate minimal score variations (less than 1%), indicating consistent performance under non-targeted disruptions. However, under print attacks, performance drops are notable. For example, two closed-source models experience substantial declines: Gemini-1.5-pro (Team et al., 2024) drops by 21.5% (from 77.79 to 61.05), and GPT-4o (OpenAI, 2023) suffers a 25.8% decrease (from 75.69 to 56.10). Among the advanced open-source models, GLM-4V's (GLM et al., 2024) performance decreases sharply by 36.5% (from 82.09 to 52.09), and Phi-3-Vision (Abdin et al., 2024) records a 21.39% drop (from 73.23 to 57.78). Notably, mPLUG-Owl-2 (Ye et al., 2023) shows the highest robustness, with only a 1.7% reduction (from 74.09 to 72.84), while the XComposer-VL (Zhang et al., 2023; Dong et al., 2024) series also exhibits strong resistance to print attacks. In the adversarial attack scenario, where the attack algorithm directly targets the image encoder, LVLMs with shared encoder architecture (e.g., Blip2 (Li et al., 2023d), InstructBLIP (Dai et al., 2023), and XComposer-VL (Zhang et al., 2023), all using EVA-CLIP (Fang et al., 2022) as their image encoder) suffer significant performance drops, with some models even performing worse than random chance. For instance, XComposer-VL (Zhang et al., 2023) experiences a 57.6% drop (from 71.40 to 30.28). Models with alternative image encoders also experience degradation, ranging from 1% to 5%. The greater impact of adversarial noise compared to corruption noise suggests that adversarial attacks exhibit a certain degree of transferability across different model architectures, meaning that attack strategies effective on one model can potentially compromise others. More detailed results are available in Appendix **??**.

4.2.2 The Validity of Dysca

In this section, we investigate on the evaluation gap between Dysca and non-synthesis benchmarks. We calculate the Spearman's rank correlation coefficient (Spearman, 1904) ρ and the Kendall rank correlation coefficient (KENDALL, 1938) τ between the evaluation ranking of Dysca under clean scenario with the non-synthesis benchmark's evaluation ranking, i.e., MMBench (Liu et al., 2023c), OCRBench (Liu et al., 2024) and SeedBench-2 (Li et al., 2023b). Both coefficient generate a score in the range of [-1,1], where 1 represents a perfect positive correlation, -1 represents a perfect negative correlation, and 0 represents no correlation. These coefficients are typical tools for measuring the correlation between variables in statistics. When the absolute value of either coefficient exceeds 0.6, it is considered to indicate a significant correlation Akoglu (2018). Specifically, we intersect our Dysca with current benchmarks based on the perceptual subtasks, evaluation models and evaluation question types. We then calculate the correlation of model evaluation rankings within this intersection. The results are shown in the first row of Tab. 4. For the MMbench (Liu et al., 2023c) and OCRBench (Liu et al., 2024), both metrics show the high correlation, with ρ and τ higher than 0.6. However, the correlation for SeedBench-2 (Li et al., 2023b) is not as strong. Considering that SeedBench-2 only contains realistic images, we conduct additional experiments using the evaluation ranks on our realistic style images only. As shown in the second row of Tab. 4, the correlation results of SeedBench-2 significantly improve (i.e., 0.46 vs. 0.64 for

Table 4: The correlation results on three benchmarks, where $\rho \in [-1, 1]$ and $\tau \in [-1, 1]$.

Style	Method	MMBench	OCRBench	SeedBench-2
All	ρ	0.70	0.90	0.46
	τ	0.60	0.80	0.43
Realistic	ρ	0.70	1.00	0.64
	τ	0.60	1.00	0.62

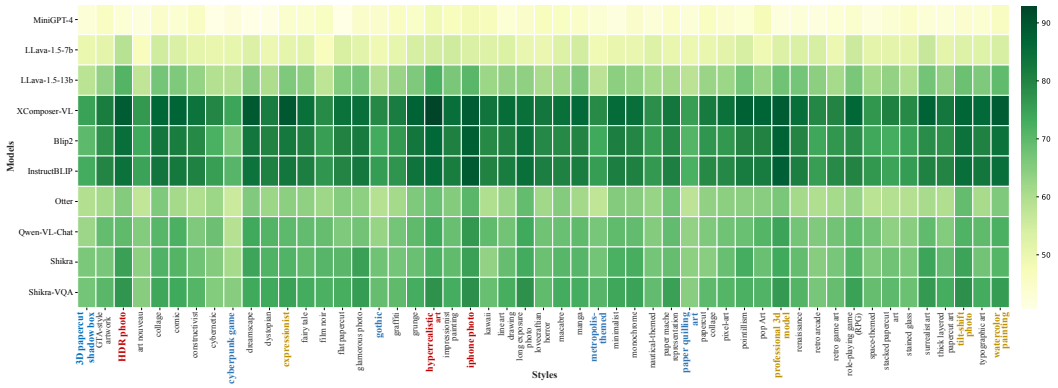

Figure 7: Illustration of each model's performance across 51 image styles, where the darker colors represent higher scores. The representative styles are colored with non-black font. Realistic styles are shown in red font, unrealistic but common styles are displayed in yellow font, and unrealistic and less common styles are represented in blue font.

ρ and 0.43 vs. 0.62 for τ). The correlation with OCRBench also improves to 1, demonstrating the validity of using synthetic datasets for evaluation LVLMs.

To further explore the the impact of image styles on evaluation results, we present the average scores across all subtasks for each of the 51 styles in Fig. 7. We observe slight score differences across all styles. In the case of realistic styles such as "iPhone photo", all LVLMs perform better compared to other image styles. The LVLMs also exhibit better performance on unrealistic but common styles like "expressionist". However, for unrealistic and less common styles such as "gothic", all models show relatively poor performance. The results reveal that the gap between Dysca and non-synthesis benchmarks primarily stems from the more diverse range of image styles, making Dysca a more comprehensive benchmark for assessing the perception ability compared to previous benchmarks.

Besides, we calculate the data distribution distance between each benchmark to prove the low distance distribution between Dysca and non-synthesis benchmarks. We select CCBench (Liu et al., 2023c), COCO-Val, MMVet (Yu et al., 2023), MMBench (Liu et al., 2023c), MME (Fu et al., 2023), MMStar (Chen et al., 2024), OCRBench (Liu et al., 2024) and ScienceQA (Lu et al., 2022). The reason why we choose these benchmark is that they have been widely used in evaluating LVLMs. We use Kernel Maximum Mean Discrepancy (KMMD) (Schölkopf et al., 2007) to measure the distribution distance. Specifically, we randomly sample 3,000 images from each benchmark (if the scale of the benchmark less than 3000, we use all that data) and utilize CLIP (Radford et al., 2021) to encode these images. Then, we calculate the KMMD value using an RBF kernel between each pair. The results are shown in Fig. 6, with darker colors indicating larger distances. As can be seen, the differences between Dysca and other real datasets are not significant. In fact, the overall distribution distance is even smaller than other datasets compared with MME does. This demonstrates that the distribution between our dataset and real datasets is minimal, indicating that the evaluation results can effectively reflect the model's performance in real-world scenarios.

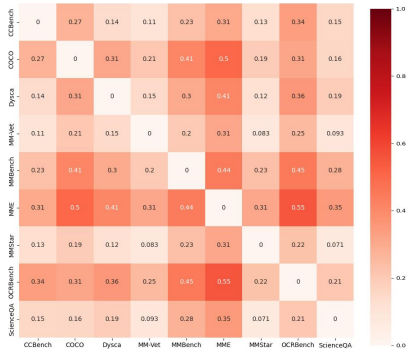

Figure 6: The KMMD distance between each benchmarks, with darker colors indicating larger distances.

5 Conclusion

In this paper, we purpose Dysca, a dynamic and scalable benchmark for evaluating perception ability of Large Vision Language Models (LVLMs). Dysca consists of 617K Vision-language QA pairs, covering 20 perceptual subtasks, 4 image scenarios and 3 question types. We conduct the experiment on 24 advanced open-source LVLMs and 2 closed-source LVLMs, revealing the insightful weakness of current LVLMs when facing different question types, image styles and image conditions. Experiments demonstrate the validity on evaluating LVLMs by using synthesis images.

REFERENCES

Openai. hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Join the midjourney discord server! https://discord.com/invite/midjourney.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, and Nguyen Bach et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, and Guanwei Zhang et al. Yi: Open foundation models by 01.ai. *arXiv preprint arXiv:2403.04652*, 2024.

Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18 (3):91–93, 2018. ISSN 2452-2473.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015a. doi: 10.1109/ICCV.2015.279.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015b.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023a.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023b.

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 (NeurIPS)*, 33:1877–1901, 2020.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*, 2023.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *arXiv preprint arXiv:2311.16465*, 2023a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *arXiv preprint arXiv:2402.19150*, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

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. *arXiv preprint:2305.06500*, 2023.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, and Xilin Wei et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2022.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, June 2023.

FastChat. Vicuna. https://github.com/lm-sys/FastChat, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, and Benyou Wang. Mllm-bench, evaluating multi-modal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*, 2023.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, and Diego Rojas et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017. doi: 10.1109/CVPR.2017.670.

M. G. KENDALL. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81.

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:2305.03726*, 2023a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023b.

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*, 2023c.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023d.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023e.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, Xuanjing Huang, and Zhongyu Wei. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*, 2023f.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, September 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

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*, 2023c.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2024.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019. doi: 10.1109/CVPR.2019.00331.

OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems (NeurIPS)*, 35:27730–27744, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. In *2014 IEEE International Conference on Computer Vision (ICCV)*, 2014.

Bernhard Schölkopf, John Platt, and Thomas Hofmann. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, pp. 513–520, 2007.

Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, and Ping Luo. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019. doi: 10.1109/CVPR.2019.00851.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49659–49678. Curran Associates, Inc., 2023a.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023b.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, and Anja Hauth et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:312.11805*, 2024.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 296–310. Association for Computational Linguistics, June 2021.

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.

Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v(ision). *arXiv preprint arXiv:2310.16534*, 2023.

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.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*, 2023a.

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*, 2023b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

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.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*, 2024b.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*, 2023.