LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Vision-Language Models (LVLM) have recently played a dominant role in multimodal vision-language learning. Despite the great success, it lacks a holistic evaluation of their efficacy. This paper presents a comprehensive evaluation of publicly available large multimodal models by building an LVLM evaluation Hub (LVLM-eHub). Our LVLM-eHub consists of 8 representative LVLMs such as InstructBLIP and MiniGPT-4, which are thoroughly evaluated by a quantitative capability evaluation and an online arena platform. The former evaluates 6 categories of multimodal capabilities of LVLMs such as visual question answering and embodied artificial intelligence on 40 standard text-related visual benchmarks, while the latter provides the user-level evaluation of LVLMs in an open-world question-answering scenario. The study reveals several innovative findings. First, Instruction-tuned LVLM with massive in-domain data such as InstructBLIP may overfit many existing tasks, generalizing poorly in the open-world scenario. Second, Instruction-tuned LVLM with moderate instruction-following data may result in object hallucination issues (i.e., generate objects that are inconsistent with target images in the descriptions). It either makes the current evaluation metric such as CIDER for image captioning ineffective or generates wrong answers. Third, employing a multi-turn reasoning evaluation framework could mitigate the issue of object hallucination, shedding light on developing an effective metric for LVLM evaluation. The findings provide a foundational framework for the conception and assessment of innovative strategies aimed at enhancing zero-shot multimodal techniques. The evaluation pipeline will be available at vlarena page.

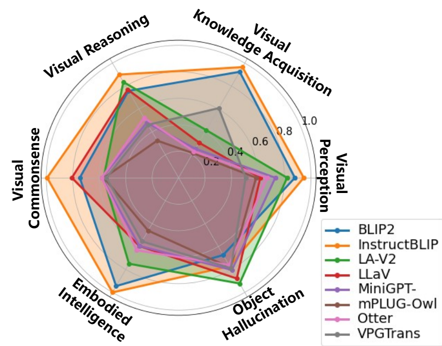
1 Introduction

Large Language Models (LLMs), such as LLaMA [1], GPT-3 [2], and Vicuna [3], have demonstrated remarkable progress in Natural Language Processing (NLP). These models leverage large-scale pre-training data and huge networks to achieve impressive results in NLP benchmarks. Recently, GPT-4 [4] further expanded the impact to the multimodal community, stimulating the rapid development of large vision-language models (LVLMs) and revolutionizing the landscape of artificial intelligence.

Large Vision-Language Models (LVLM) have achieved remarkable progress in multimodal vision-language learning for various multimodal tasks such as visual question answering and multimodal conversation. Specifically, LVLMs capitalize on the knowledge from LLMs and effectively align visual features with the textual space. Flamingo [5], a pioneering LVLM, integrates visual features into LLMs through cross-attention layers. Later studies proposed more efficient vision-text interactions [6], more efficient training methods [7, 8], and employing instruction tuning [9, 7, 9, 10, 11, 12, 13, 8].

However, despite the great success, few efforts have been made to provide systematic evaluations of LVLMs. But evaluation plays a critical role in understanding the strengths and weaknesses of LVLMs,

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) Quantitative Capability Evaluation

Ranking	Model	Score
1	mPLUG-Owl	1027.0
2	MiniGPT-4	1021.3
3	Otter	1013.2
4	LLaMA-Adapter V2	1010.2
5	LLaVA	1009.7
6	InstructBLIP	1003.7
7	VPGTrans	974.3
8	BLIP2	949.4

(b) LVLMs Arena Ranking

Figure 1: Comparative analysis of LVLMs within the LVLM eHub. (a) illustrates the variances in quantitative capability performance across six distinct aspects among LVLMs. (b) presents the Elo rating ranking of LVLMs within the LVLM Arena.

thereby guiding their future development. Recent work [14] presents a systematic investigation of object hallucination of LVLMs by proposing a polling-based object probing evaluation method. Moreover, ImageNetVC [15] studies how well LVLMs can master visual commonsense knowledge. Liu et al. [16] comprehensively evaluate the performance of LVLMs in visual recognition with text recognition, such as optical character recognition. GVT [17] evaluates LVLM's visual semantic understanding and fine-grained perception capabilities. Nevertheless, these studies only evaluate a portion of LVLMs on specific tasks, lacking an overall understanding of LVLM's capabilities.

In pursuit of a comprehensive evaluation of LVLMs, we build an LVLM Evaluation hub (LVLM-eHub) consolidating 8 representative LVLMs such as InstrucBLIP and MiniGPT-4. The detailed information about model configuration and training data is listed in Table 1. Our LVLM-eHub consists of a quantitative capability evaluation and an online arena platform, providing a thorough investigation of the selected LVLMs. Specifically, the quantitative capability evaluation extensively evaluates 6 categories of multimodal capabilities of LVLMs including visual perception, visual knowledge acquisition, visual reasoning, visual commonsense, object hallucination, and embodied intelligence (see Fig. 1 (a)), by collecting 40 standard text-related visual benchmarks. On the other hand, the online arena platform features anonymous randomized pairwise battles in a crowd-sourced manner, providing a user-level model ranking in the open-world question-answering scenario (see Fig. 1 (b)).

Our LVLM-eHub comprehensively evaluates LVLMs, revealing several innovative findings. (1) Instruction-tuned LVLM with massive in-domain data suffers from overfitting and generalizes poorly in open-world scenarios, such as InstructBLIP (see Fig. 1 (a)). (2) With moderate instruction-following data, Instruction-tuned LVLM may cause object hallucination issues, generating objects that are inconsistent with target images in the descriptions. This leads to incorrect answers or renders current evaluation metrics, such as CIDER for image captioning, ineffective. (3) We find that a multi-turn reasoning evaluation pipeline can mitigate the issue of object hallucination, indicating that developing an effective metric for LVLM evaluation is urgent.

The contributions of our work are summarized follows. (1) We propose LVLM-eHub which is the first comprehensive evaluation benchmark for large vision-language models, to our best knowledge. (2) LVLM-eHub provides extensive evaluation on 6 categories of multimodal capabilities of LVLMs in more than 40 text-based visual tasks. (3) LVLM-eHub builds an online arena platform for LVLMs, which features anonymous randomized pairwise user-level comparison in a open-world scenario. (4) Our evaluation results reveal several innovative findings, providing a foundational framework for the assessment of innovative strategies aimed at enhancing zero-shot multimodal techniques.

2 LVLM Evaluation Hub

In this section, we introduce representative LVLMs, multimodal capabilities of interest, and evaluation methods. The whole LVLM Evaluation Hub is illustrated in Fig. 2. Our LVLM evaluation hub

Model	Model Configuration						Image-Text Data		Visual Instruction Data	
	VE	LLM	Adapt	ToP	TuP	# Token	Source	Size	Source	Size
BILP2	ViT-g/14†	FlanT5-XL†	Q-Former	4B	107M	32	CC*-VG-SBU-L400	129M	-	-
LLaVA	ViT-L/14†	Vicuna	FC layer	7B	7B	256	CC3M	595K	LLaVA-I	158K
LA-V2	ViT-L/14†	LLaMA†	B-Tuning	7B	63.1M	10	L400	200M	LLaVA-I+G4L	210K
MiniGPT-4	BLIP2-VE†	Vicuna†	FC layer	7B	3.1M	32	CC-SBU-L400	5M	CC+ChatGPT	3.5K
mPLUG-Owl	ViT-L/14	LLaMA†	LoRA	7B	1.1B	65	CC*-CY-L400	204M	LLaVA-I	158K
Otter	ViT-L/14†	LLaMA†	Resampler	9B	1.3B	64	-	-	LLaVA-I	158K
InstructBLIP	ViT-g/14†	Vicuna†	Q-Former	7B	107M	32	-	-	QA*	16M
VPGTrans	ViT-g/14†	Vicuna†	Q-Former	7B	107M	32	COCO-VG-SBU	13.8M	CC+ChatGPT	3.5K

Table 1: **Comparison of Different LVLMs.** 'VE', 'Adapt', 'ToP', 'TuP', and '# Token' represent the visual encoder, adaption module, number of total parameters, tuning parameters, and visual tokens fed into the text encoder, respectively. † indicates that the model is frozen. CC* consists of COCO [18], CC3M [19], and CC12M [20]. CC, VG, SBU CY, and L400 indicate Conceptual Caption [19, 20], Visual Genome [21], COYO-700M [22] and LAION 400M [23], respectively. LLaVA-I and G4L represent 158K multimodal instruction-following data in LLaVA [9] and data generated by GPT-4 for building an instruction-following LLMs [24]. QA* denotes 13 question-answering datasets in InstructBLIP [13]. We count all the data and tuning parameters needed to convert the pretrained vision model and LLM into a visual instruction model. The average score is obtained by normalizing over each row and taking the average of each column.

compromises 8 representative models including BLIP2 [6], LLaVa [9], LLaMA-Adapter V2 [7], MiniGPT-4 [10], mPLUG-Owl [11], Otter [12], InstructBLIP [13], and VPGTrans [8]. All models boost vision-language representation learning by utilizing pre-trained image encoders and large language models (LLM). But they differ in training data scale and model configuration as shown in Table 1. For a fair comparison between LVLMs, we collect their checkpoints with parameter sizes less than 10B. The detailed descriptions of these models are in the Appendix.A.

2.1 Quantitative Capability Evaluation

We aim to evaluate LVLMs' capability comprehensively. In particular, we summarize 6 categories of capabilities and collect corresponding benchmarks for quantitative evaluation (see Fig.2). Please see our supplementary materials for more statistics and details of the collected benchmarks.

Visual Perception. Visual perception is the ability to recognize the scene or objects in images, the preliminary ability of the human visual system. We evaluate this capability of models through image classification (ImgCLs) using the ImageNet1K [25], CIFAR10 [26], Pets37 [27] and Flowers102 [28] benchmarks, multi-class identification (MCI) and object counting (OC) using the GVT [29] benchmark. ImgCLs and MCI measure how well an LVLM grasps high-level semantic information, while OC assesses the recognition ability for fine-grained objects.

Visual Knowledge Acquisition. Visual knowledge acquisition entails understanding images beyond perception to acquire knowledge. This evaluation is conducted through Optical Characters Recognition (OCR) using twelve benchmarks (including IIIT5K [30], IC13 [31], IC15 [32], Total-Text [33], CUTE80 [34], SVT [35], SVTP [36], COCO-Text [37], WordArt [38], CTW [39], HOST [40], WOST [40]), Key Information Extraction (KIE) using the SROIE [41] and FUNSD [42], and Image Captioning (ImgCap) using two benchmarks (including NoCaps [43] and Flickr30K [44]). The OCR task measures whether a model can accurately identify and extract text from images or scanned documents. The KIE task further poses challenges in extracting structured information from unstructured or semi-structured text. Finally, ImgCap assesses whether a model can generate a good natural language description of the content of an image.

Visual Reasoning. Visual reasoning requires a comprehensive understanding of images and related texts. To evaluate the visual reasoning ability of LVLMs, we utilize three tasks including visual question answering (VQA), knowledge-grounded image description (KGID), and visual entailment SNLI-VE [45]), two benchmarks (i.e. ScienceQA [46] and VizWiz [47]) and one benchmark (i.e. SNLI-VE), respectively. These three tasks are in VQA form in different domains. A capable LVLM should be able to understand the objects and scenes in an image and can reason to generate answers that are semantically meaningful and relevant to the question asked.

Visual Commonsense. Visual commonsense refers to the general visual knowledge commonly shared across the world, as opposed to the visual information specific to a single image. This evaluation tests

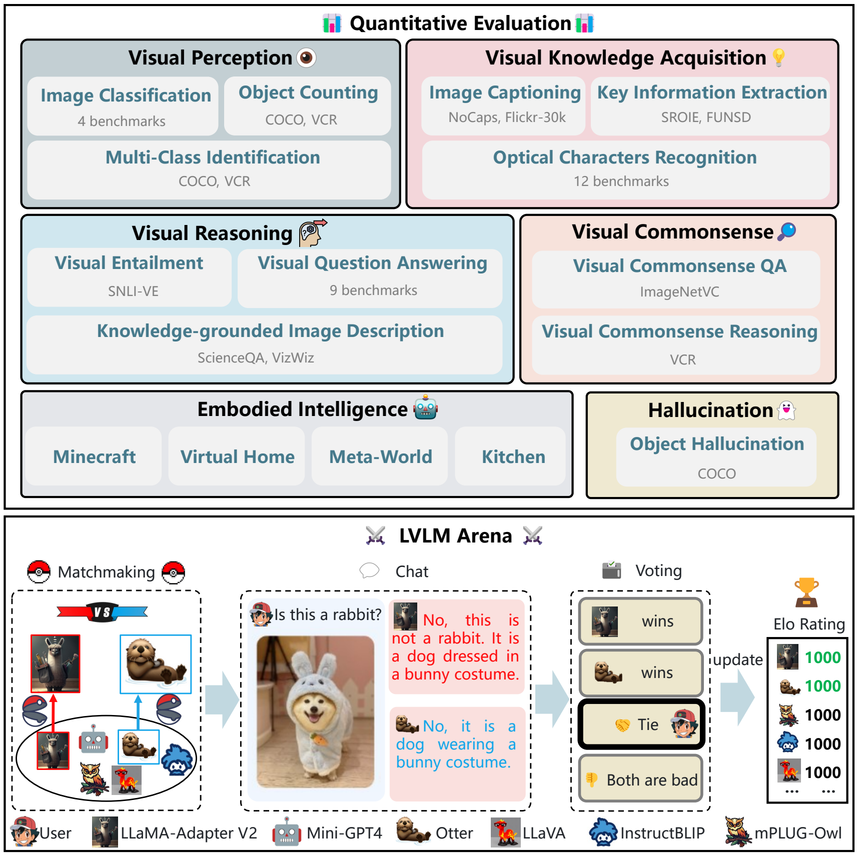


Figure 2: Our evaluation encompasses quantitative evaluation and online LVLM Arena. Plentiful benchmarks are employed to comprehensively evaluate the six critical capabilities of the models in the quantitative evaluation. In the LVLM Arena, an online platform, users can participate in an online evaluation by chatting with two anonymous models and choosing their preferred model.

the model's understanding of commonly shared human knowledge about generic visual concepts using ImageNetVC [15] and visual commonsense reasoning (VCR) [48]. Specifically, ImageNetVC is utilized for zero-shot visual commonsense evaluation, such as color and shape, while VCR covers various scenes, such as spatial, casual, and mental commonsense.

Embodied Intelligence. Embodied intelligence aims to create agents, such as robots, which learn to solve challenging tasks requiring environmental interaction. Recently, LLM and LVLM exhibited exceptional effectiveness in guiding the agent to complete a series of tasks. In this evaluation, we utilize high-level tasks as in EmbodiedGPT [49] and employ Minecraft [50], VirtualHome [51], Meta-World [52], and Franks Kitchen [52] as benchmarks.

Object Hallucination. It is known that LVLM suffers from the object hallucination problem, i.e., the generated results are inconsistent with the target images in the descriptions [14]. Evaluating the degree of object hallucination for different LVLMs help understand their respective weaknesses. To this end, we evaluate the object hallucination problem of LVLMs on the MSCOCO dataset [18].

2.2 Online Evaluation with LVLM Arena

Designing quantitative evaluations for LVLM to satisfy all capabilities is challenging, as evaluating LVLM responses constitutes an open-ended problem. Inspired by FastChat [53], we introduce the LVLM Arena, an online evaluation framework for LVLMs' pairwise battle with human judgment.

Figure 2 illustrates the LVLM Arena, comprising three primary components: matchmaking, chat, and voting. Initially, two models are sampled from the model zoo. Users then converse side-by-side with the models, who remain anonymous. Subsequently, users vote for the superior model.

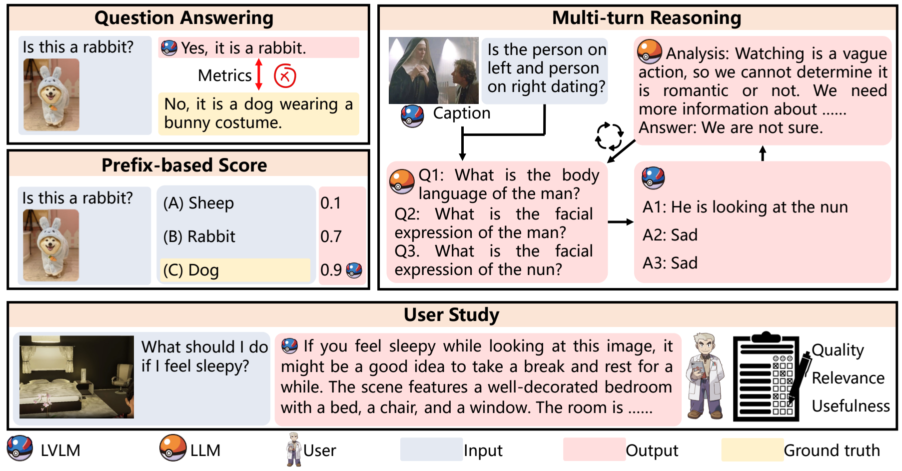


Figure 3: Illustration of our adopted evaluation methods. To evaluate the zero-shot performance of LVLMs on diverse downstream tasks, we employ four methods including question answering, prefix-based score, multi-turn reasoning, and user study.

Matchmaking. The matchmaking module samples two models in a tournament style based on their Elo rating. However, due to the currently limited size of the model hub, we employ random sampling.

Chat. Users chat side-by-side with two sampled models (which remain anonymous) using images or text inputs. Different from quantitative evaluation, users can chat about anything. Our existing online platform supports only single-round chats due to multi-round chats' high computational and memory demands. Future updates will address this constraint.

Voting. After the chat session, users vote for their preferred model. Four options are available: Model A, Model B, Tie, and Both are bad. The Elo rating is subsequently updated using voting results.

In contrast to limited quantitative evaluations, the LVLM Arena provides an open-world evaluation framework that enables users to chat with models about anything, emulating real-world conditions. Besides, users serve as the judge for the battle, which brings more convincing evaluation results than traditional evaluation metrics.

2.3 Zero-shot Evaluation

LVLMs are capable of capturing a wide range of multimodal patterns and relationships. We evaluate the above 6 categories of capabilities of LVLMs by investigating their zero-shot performance on various tasks. Zero-shot evaluation allows us to evaluate the LVLMs' ability to generalize to new tasks without training the model, which is competent for large-scale evaluation. To be specific, we treat the zero-shot evaluation as various forms of prompt engineering for different tasks (see Fig. 3) as presented in the following.

- *Question Answering.* Prompting with visual question answering can be used to solve many downstream tasks, which assess how well an LVLM understands the underlying language and visual features. We design proper prompts to ensure that the LLM can produce meaningful results. For example, text prompts of OCR can be "*what is written in the image?*". Then, we evaluate the answers generated by the LLM using the corresponding metric such as accuracy.
- *Prefix-based Score.* For multi-choice QA tasks, we can utilize a visual encoder to obtain visual prompts for a given image. Then, the visual prompts are prefixed into the text embeddings, which are fed into the LLM. The likelihood of image-text pair can be generated, which is referred to as a prefix-based score. We can obtain a prefix-based score for each text prompt of the candidate's answer. The answer with the largest prefix-based score is selected as the final answer. We provide the formulation in Sec. A.3 of Appendix.
- *Multi-turn Reasoning.* Following IdealGPT [16], we use a multi-turn reasoning framework to evaluate complex visual reasoning tasks. Specifically, we utilize an LLM such as ChatGPT to generate sub-questions for a given question, an LVLM to provide corresponding sub-answers, and

Datasets		BLIP2	InstructBLIP	LA-V2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans	S-SOTA
ImgCls	ImageNet1K [54]	23.71	24.51	25.89	23.50	21.58	**26.81**	19.29	15.60	91.10 [55]
	CIFAR10 [26]	58.20	67.24	64.86	**67.96**	61.17	53.09	65.42	53.11	99.70 [56]
	Pets37 [27]	34.83	**39.17**	24.56	9.05	19.81	33.66	5.91	8.56	96.70 [57]
	Flowers102 [28]	30.90	**32.79**	32.05	11.99	29.74	20.15	10.41	10.46	99.64 [58]
OC	COCO	**48.90**	46.65	38.50	20.56	20.86	27.51	46.14	25.46	-
	VCR	25.05	29.29	26.51	24.60	25.26	8.99	**41.06**	18.03	-
MCI	COCO	86.06	**87.81**	82.90	49.66	72.70	35.39	51.03	50.98	-
	VCR	66.59	**76.49**	50.66	66.90	66.02	19.12	51.60	47.13	-
Avg.		0.879	**0.946**	0.820	0.617	0.731	0.753	0.669	0.507	-

Table 2: Evaluation results of visual perception capability of LVLMs on tasks of Image Classification (Imgcls), Object Counting (OC), and Multi-class Identification (MCI). The **best** result is **bold** while the second is underlined. S-SOTA indicates the supervised state-of-the-art results

another LLM to reason to assess sub-answers' quality. Such a pipeline iteratively proceeds until a satisfactory answer is obtained. We provide the formulation in Sec. A.3 of Appendix.

- *User Study.* Evaluating the quality of the text generated by an LVLM requires a thorough understanding of the underlying language and context. In embedded artificial intelligence tasks, the LVLM generates a plan for the given instruction, which should be evaluated through various aspects such as recognition accuracy and conciseness in answers. It is hard to implement such an evaluation using an existing metric. Thus, user studies are conducted to assess the quality, relevance, and usefulness of the text generated by the LVLM in a specific context. To maintain evaluation fairness, we randomly shuffle the model's output order and anonymize outputs during evaluation.
  Note that our user study does not involve direct interactions with human participants and does not involve potential risks to participants, such as the collection of personal information, or any other aspects that could impact the participants' rights or well-being. Currently, we do not include an IRB Approval. We are dedicated to addressing the ethical and moral considerations regarding the user evaluation method with thoroughness and commitment, while also providing effective solutions.

3 Experiment and Analysis

In this section, we perform a zero-shot evaluation to assess the 6 kinds of capabilities of LVLMs. Specifically, visual perception ability, visual knowledge acquisition, visual Reasoning, visual commonsense understanding, visual object hallucination, and embodied intelligence are assessed in Sec. 3.1 $\sim$ Sec.3.6, respectively. The LVLM arena evaluation result is presented in Sec.3.7. More quantitative results can be found in Appendix C.

3.1 Results on Visual Perception

Visual perception is an important ability of LVLMs. As presented in Sec. 2.1, we evaluate through image classification (ImgCls), multi-class identification (MCI), and object counting (OC). The evaluation details of tasks are demonstrated in Appendix.B.1. The evaluation results are reported in Table 2. We have three observations. (1) mPLUG-Owl and LLaVA perform best on coarse-grained classification tasks (*i.e.,* ImageNet1K and CIFAR10). The commonality is that they update LLM with 158K instruction-following data. (2) InstructBLIP presents good perception ability in fine-grained ImgCls, OC, and MCI tasks. The main reason is that InstructBLIP may be fine-tuned on various existing VQA datasets, which may make it overfit on these tasks. (3) The performances of LVLMs on ImgCls are significantly inferior to supervised SOTA, indicating plenty of room for LVLM's perception ability.

3.2 Results on Visual Knowledge Acquisition

Visual knowledge acquisition involves going beyond image perception to acquire deeper understanding and knowledge. In our study, we evaluate the acquisition of visual knowledge through various tasks, namely Optical Character Recognition (OCR), Key Information Extraction (KIE), and Image Captioning, all performed in a Visual Question Answering (VQA) fashion. The evaluation details of tasks are demonstrated in Appendix.B.2. Table 3 shows the zero-shot performance in visual knowledge acquisition, and we have the following observations. First, BLIP2, InstructBLIP, and

Datasets		BLIP2	InstructBLIP	LA-V2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans	S-SOTA
OCR	IIIT5K	80.17	**83.90**	36.30	31.57	25.13	26.50	17.57	51.50	99.2[59]
	IC13	81.13	**82.08**	20.87	16.39	16.75	14.86	09.67	61.67	98.4[60]
	IC15	66.68	**73.57**	29.40	26.58	21.43	21.14	18.49	42.00	91.4[59]
	Total-Text	68.31	**71.51**	30.93	24.51	18.65	21.08	14.81	43.60	90.5[61]
	CUTE80	85.07	**86.11**	35.76	36.46	33.33	34.03	18.75	62.85	99.3[59]
	SVT	85.78	**86.86**	20.40	18.55	17.47	13.45	10.51	51.16	98.3[59]
	SVTP	77.34	**80.93**	31.01	27.44	19.69	20.78	19.22	47.13	97.2[59]
	COCO-Text	53.62	**58.25**	20.94	18.05	12.05	13.50	11.30	27.00	81.1[59]
	WordArt	73.66	**75.12**	38.98	35.87	31.57	32.36	21.05	53.30	72.5[38]
	CTW	67.43	**68.58**	18.13	16.73	15.14	12.91	10.05	40.80	88.3[61]
	HOST	57.28	**61.22**	16.60	15.94	14.57	11.92	10.14	32.20	77.5[59]
	WOST	68.83	**73.26**	21.73	20.49	17.47	14.45	12.29	37.91	87.5[59]
KIE	SROIE	0.08	**0.09**	0.02	0.01	0.01	0.01	0.01	0.06	97.81[62]
	FUNSD	1.02	1.03	**2.16**	1.93	1.20	0.41	1.91	1.27	89.45[63]
Image Captioning	NoCaps	**48.60**	46.61	33.69	1.56	5.84	0.26	11.56	36.20	124.77[64]
	Flickr-30k	46.65	**50.69**	23.85	2.23	2.66	0.02	7.12	23.41	-
Average Score		0.924	**0.965**	0.416	0.307	0.253	0.215	0.231	0.607	-

Table 3: Comparison of Zero-shot Performance for Large-scale Vision and Language Models (LVLMs) on OCR, KIE, and Image Captioning Tasks. Evaluation metrics include word accuracy for OCR datasets, entity-level F1 score for KIE datasets, and CIDEr score for image captioning datasets.

Datasets		BLIP2	InstructBLIP	LLaMA-Adapter-v2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans	S-SOTA
VQA	DocVQA	4.75	5.89	**8.13**	6.26	3.57	2.24	3.44	2.64	54.48[65]
	TextVQA	31.98	39.60	**43.76**	38.92	21.78	38.76	21.52	17.52	73.1[66]
	STVQA	20.98	28.30	**32.33**	28.40	12.20	8.30	15.23	12.88	-
	OCR-VQA	38.85	**60.20**	38.12	23.40	16.15	3.40	19.50	16.97	-
	OKVQA	44.93	**60.52**	55.93	54.36	30.06	22.89	49.01	45.31	-
	GQA	45.53	**49.96**	43.93	41.30	27.03	12.60	38.12	38.54	72.1[67]
	Visdial	10.73	**45.20**	12.92	14.66	7.97	13.34	11.67	12.10	68.92[68]
	IconQA	**62.82**	56.25	41.83	42.95	28.20	09.12	26.77	25.73	83.62[69]
	VSR	**63.63**	41.28	50.63	51.24	41.04	10.11	06.40	37.00	70.1[70]
KGID	ScienceQA IMG	**60.73**	46.26	54.19	49.33	20.18	2.80	27.22	20.43	92.53[71]
	VizWiz	**65.44**	65.31	62.07	62.42	40.76	11.14	50.04	11.99	73.3[66]
VE	SNLI-VE	34.00	56.20	56.80	**57.00**	52.60	55.00	56.60	47.60	-
Average Score		0.758	**0.900**	0.835	0.769	0.481	0.324	0.523	0.462	-

Table 4: Comparison of Zero-shot Performance for LVLM Models on VQA, KGID, and VE Tasks. For VQA and KGID tasks, Mean Reciprocal Rank (MRR) is used for the Visdial, while top-1 accuracy is employed for the remaining tasks.

VPGTrans achieve dominant performance in all tasks. This may be because these models use a large visual encoder (i.e., ViT-g/14) and Q-Former updated with massive image-text pairs. A stronger visual encoder and adaption module can extract better tokens entailed with the global and local context, leading to remarkable improvement in visual knowledge acquisition. Second, InstructBLIP presents consistently the best results on all tasks. The main reason is that InstructBLIP overfits these tasks by fine-tuning massive VQA data.

3.3 Results on Visual Reasoning

Visual reasoning encompasses the ability to comprehensively understand images and perform cognitive tasks. In this section, we evaluate the visual reasoning ability of LVLMs on various tasks, including Visual Question Answering (VQA), Knowledge-Grounded Image Description (KGID), and Visual Entailment (VE) tasks. The evaluation details of tasks are demonstrated in Appendix.B.3. Table 4 shows the zero-shot performance in visual reasoning, and we have the following observations. First, compared with BLIP2, InstructBLIP again presents better results overall because it overfits many tasks by fine-tuning massive VQA data. Second, compared with BLIP2, instruction-tuned LVLMs, except for InstructBLIP, generally perform worse than BLIP2. The common words in the instruction data often influence the generated content, which can not be evaluated by the current metrics (see Appendix C). Third, instruction-tuned LVLMs consistently surpass BLIP2 on SNLI-VE where the final answer is obtained by multi-turn reasoning. It shows that instruction-following fine-tuning can produce promising content once a good evaluation scheme is employed.

3.4 Results on Visual Commonsense

The visual commonsense evaluation aims to evaluate the model's comprehension of commonly shared human knowledge about generic visual concepts. We use two challenging visual commonsense

Datasets		BLIP2	InstructBLIP	LA-v2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans	S-SOTA
ImageNetVC	Color	44.60	**67.79**	23.16	41.92	26.57	25.56	26.21	24.72	44.70[15]
	Shape	40.14	**59.06**	28.16	38.74	22.88	30.72	34.19	24.69	40.50[15]
	Mater.	61.49	63.58	32.51	**64.91**	29.50	34.24	35.81	27.21	61.90[15]
	Compo.	53.86	**83.25**	50.38	58.53	59.96	49.47	50.72	57.21	54.00[15]
	Others	51.50	**68.37**	32.64	59.06	38.86	35.11	34.39	36.39	51.70[15]
	Avg	50.30	**68.41**	33.37	52.63	35.55	35.02	36.26	34.04	50.50[15]
VCR	VCR	36.80	45.60	**46.20**	**46.20**	44.40	39.40	39.60	39.60	-
Average Score		0.747	**0.994**	0.567	0.807	0.581	0.564	0.581	0.546	-

Table 5: Comparisons of Zero-shot visual commonsense Performance for LVLM Models on VCR and ImageNetVC datasets. Top-1 accuracy is employed for the two datasets.

Datasets		BLIP2	InstructBLIP	LA-V2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans	S-SOTA
MSCOCO	Random	82.21	**88.83**	74.44	51.52	52.58	40.65	61.40	47.92	-
	Popular	80.10	**84.15**	56.82	50.00	49.31	38.82	49.56	47.64	-
	Adversarial	78.52	**81.95**	60.52	50.00	49.62	38.04	50.68	45.95	-
Average Score		0.945	**1.00**	0.750	0.595	0.594	0.461	0.633	0.555	-

Table 6: Evaluation results of POPE [14] performance of LVLMs on MSCOCO. The accuracy is used to assess the performance.

benchmarks in a zero-shot setting, including ImageNetVC and Visual Commonsense Reasoning (VCR). The evaluation details of tasks are demonstrated in Appendix.B.4. As shown in Table 5, we can find that all those LVLMs represent their abilities to solve visual commonsense problems. First, InstructBLIP performs best (68.41%) among those LVLMs on the ImageNetVC dataset. The main reason is that it is fine-tuned on 1.6M fine-grained VQA data, making it adapt to answer visual common questions. Second, LLaMA-Adapter V2 (46.20%) and LLaVA (46.20%) show the same best performance among those LVLMs on the VCR dataset. The main reason is that instruction-flowing data is used to update the LLM. Note that the final answer of VCR is obtained by multi-turn reasoning. It also shows the significant role of a good evaluation scheme in producing promising content for instruction-tuned models.

3.5 Results on Object Hallucination

Although LVLMs have made significant progress, they still struggle with the issue of hallucination, which refers to their tendency to produce objects that do not align with the descriptions provided in the target images. In this section, we focus on evaluating such object hallucination problems on MSCOCO captioning dataset. Following POPE [14] evaluation pipeline which is a multi-step QA procedure, we prompt LVLMs with multiple Yes-or-No questions. For example, '*Is there a person in the image?*'. We use accuracy as the evaluation metric. From Table 6, we could come to the following conclusions. InstructBlip performs best in the hallucination problem, followed by BLIP2, whose average accuracy both reached more than 80%. We find that instruction-tuned models, except for InstructBLIP, perform worse than BLIP2 because they tend to answer 'Yes' to the question, which shows that LVLMs are prone to generate objects frequently occurring in the instruction data. Such object hallucination problem can be alleviated by a multi-turn reasoning pipeline shown in the experiments on SNLI-VE and VCR.

3.6 Results on Embodied Intelligence

In this section, we present the evaluation results focusing on embodied intelligence. To appraise the effectiveness of planning outputs using the given image, we conducted a user study involving 15 participants. The study comprised 6 household scenarios carefully selected from VirtualHome [51]. Specifically, the participants rated the generated plans from different LVLM models using a scoring system similar to [49]. The evaluation comprised five dimensions with scores ranging from 1 to 5. These dimensions included object recognition accuracy, spatial relationship understanding, level of conciseness in the response, reasonability of the planning, and executability of the planning. The resulting average scores for the different models among the participants are presented in Table 7 below. Furthermore, in the Appendix C, we present quantitative evaluation results for Franka Kitchen [52], Minecraft [50], and Meta-World [72]. Based on the evaluation results, we observe that visual

Dataset		BLIP2	InstructBLIP	LA-V2	LLaVA	MiniGPT-4	mPLUG-Owl	Otter	VPGTrans
VirtualHome	Object Recon.(↑)	2.03	3.08	<u>3.81</u>	**3.88**	3.70	3.42	3.38	3.43
	Spatial Relation.(↑)	1.68	2.78	**3.71**	<u>3.61</u>	3.47	3.22	3.10	3.22
	Conciseness (↑)	**3.25**	<u>2.48</u>	2.04	1.86	1.62	1.48	1.86	1.76
	Reasonability(↑)	2.78	3.20	**4.04**	<u>3.70</u>	3.54	3.44	3.07	3.35
	Executability(↑)	2.88	3.10	**4.08**	<u>3.82</u>	3.11	3.54	3.12	3.35
Average Score		0.674	0.772	**0.922**	<u>0.879</u>	0.805	0.785	0.761	0.789

Table 7: Generated planning quality evaluation on embodied tasks. Five dimensions including object recognition, spatial relationship, conciseness, reasonability, and executability are used to assess the performance.

instruction data is essential for embodied tasks. BLIP2 lacked visual instruction tuning, which greatly affected its capability to produce reasonable and executable plans.

3.7 Results on Online Arena Evaluation

The arena features anonymous and randomized pairwise battles in a crowd-sourced manner. We have collected 634 pieces of evaluation data since we launch the LVLM arena. The collected data shows almost the same number of battle outcomes for 'Model A wins' and 'Model B wins.' Moreover, 21.8% battle outcomes are voted as 'both are bad,' implying that the current LVLMs still struggle to generate good answers for open-world visual questions. Furthermore, we rank the selected 8 LVLMs with Elo rating [73] using the collected data by following Fastchat [53] and [74]. As shown in Fig. 1 (b), mPLUG-Owl, MiniGPT-4, and Otter, which are fine-tuned with amounts of instruction-following data with updating many parameters, are the top-3 best models in the open-world VQA scenario, indicating the significance of instruction-following tuning and effective parameter update. Moreover, InstructBLIP perform best on in-domain capability evaluation, while being much worse than many instruction-tuned models, implying severe overfitting issue, as shown in Fig. 1.

3.8 Takeaway Analysis

We can conclude some actionable insights from our evaluation results. *First*, the quality of visual instruction data matters more than quantity in the open-world VQA. We observe that MiniGPT-4, which is tuned by only 3.5K high-quality visual instruction data performs much better than InstructBLIP tuned on visual instruction data adapted from various existing VQA datasets in our Multi-Modality Arena. *Second*, a strong visual encoder can help extract detailed information from the image, leading to good performance in OCR tasks. For instance, we see that BLIP2, InstructBLIP, and VPGTrans achieve better performance than the remaining 5 LVLMs. This may be because the visual encoder ViT-g/14 used in BLIP2, InstructBLIP, and VPGTrans is more powerful than ViT-L/14 employed in the remaining LVLMs. *Third*, multi-turn reasoning helps alleviate the hallucination issue, indicating that the evaluation method with critical thinking can induce the correct prediction from the model. We find that LVLM with multi-turn reasoning can determine whether an object exists in the image more accurately than single-turn reasoning. Hence, multi-turn reasoning is appropriate to assess the full potential of the model. *Fourth*, LVLMs tuned with high-quality instruction-following data present more promising planning ability than models without being tuned with instruction data as demonstrated in Table 7.

4 Conclusion

This paper proposes a comprehensive evaluation benchmark for large vision-language models called LVLM-eHub that incorporates both quantitative performance evaluation and human feedback evaluation. For the quantitative evaluation, we employ 16 tasks spanning over 40+ text-related visual datasets to assess the six essential capabilities of LVLM models. Additionally, we have established an online LVLM Arena to gather human feedback on LVLM models continually. This arena serves as an invaluable resource, providing an Elo rating rank that offers LVLMs ranking in the open-world scenario. Our evaluation results reveal several important findings, stimulating the future development of LVLMs. We will make ongoing efforts to build a platform for LVLM evaluation as discussed in Sec. A.4.

References

[1] 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] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[4] OpenAI. Gpt-4 technical report, 2023.

[5] 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.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[7] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[8] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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 arXiv:2305.06500*, 2023.

[14] 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.

[15] Heming Xia, Qingxiu Dong, Lei Li, Jingjing Xu, Ziwei Qin, and Zhifang Sui. Imagenetvc: Zero-shot visual commonsense evaluation on 1000 imagenet categories. *arXiv preprint arXiv:2305.15028*, 2023.

[16] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

[17] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*, 2023.

[18] 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.

[19] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[20] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

[23] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[24] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[27] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[29] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*, 2023.

[30] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2687–2694, 2012.

[31] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazàn Almazàn, and Lluís Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.

[32] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160, 2015.

[33] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 935–942, 2017.

[34] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014.

[35] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Song Gao, and Jinlong Hu. End-to-end scene text recognition using tree-structured models. *Pattern Recognition*, 47(9):2853–2866, 2014.

[36] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *2013 IEEE International Conference on Computer Vision*, pages 569–576, 2013.

[37] Andreas Veit, Tomas Matera, Lukás Neumann, Jiri Matas, and Serge J. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *ArXiv*, abs/1601.07140, 2016.

[38] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. 2022.

[39] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recogn.*, 90(C):337–345, jun 2019.

[40] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021.

[41] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.

[42] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[43] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8947–8956, 2019.

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 02 2014.

[45] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

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

[47] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[48] 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.

[49] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[50] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[51] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018.

[52] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

[53] Wei-Lin Chiang Hao Zhang Joseph E. Gonzalez Lianmin Zheng, Ying Sheng and Ion Stoica. Fastchat. `https://github.com/lm-sys/FastChat`, 2023.

[54] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[55] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023.

[56] H M Dipu Kabir. Reduction of class activation uncertainty with background information, 2023.

[57] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[58] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, pages 1–14, 2023.

[59] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model, 2023.

[60] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196, Cham, 10 2022. Springer Nature Switzerland.

[61] Tao Sheng, Jie Chen, and Zhouhui Lian. Centripetaltext: An efficient text instance representation for scene text detection. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[62] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL-IJCNLP 2021*, January 2021.

[63] Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. Geolayoutlm: Geometric pre-training for visual information extraction. *CoRR*, abs/2304.10759, 2023.

[64] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

[65] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[66] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. 2023.

[67] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4566, 2022.

[68] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048, 2019.

[69] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021.

[70] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023.

[71] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[72] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[73] Arpad E Elo. The proposed uscf rating system. its development, theory, and applications. *Chess Life*, 22(8):242–247, 1967.

[74] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[75] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online, June 2021. Association for Computational Linguistics.

[76] 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*, pages 19358–19369, 2023.

[77] 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.

[78] 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*, pages 8748–8763. PMLR, 2021.

[79] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[80] 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)*, pages 8309–8318, 2019.

[81] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570, 2019.

[82] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952, 2019.

[83] 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)*, pages 3190–3199, 2019.

[84] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019.

[85] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

[86] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[87] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[88] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.