TRINS: Towards Multimodal Language Models that Can Read

Ruiyi Zhang [1], Yanzhe Zhang [2], Jian Chen [3], Yufan Zhou [1], Jiuxiang Gu [1],
Changyou Chen [3], Tong Sun [1]
[1] Adobe Research [2] Georgia Institute of Technology [3] State University of New York at Buffalo
{ruizhang}@adobe.com

Abstract

Large multimodal language models have shown remarkable proficiency in understanding and editing images. However, a majority of these visually-tuned models struggle to comprehend the textual content embedded in images, primarily due to the limitation of training data. In this work, we introduce TRINS: a Text-Rich image[1] INStruction dataset, with the objective of enhancing the reading ability of the multimodal large language model. TRINS is built upon LAION [2] using hybrid data annotation strategies that include machine-assisted and human-assisted annotation process. It contains 39,153 text-rich images, captions, and 102,437 questions. Specifically, we show that the number of words per annotation in TRINS is significantly longer than that of related datasets, providing new challenges. Furthermore, we introduce a simple and effective architecture, called a Language-Vision Reading Assistant (LaRA), which is good at understanding textual content within images. LaRA outperforms existing state-of-the-art multimodal large language models on the TRINS dataset as well as other classical benchmarks. Lastly, we conducted a comprehensive evaluation with TRINS on various text-rich image understanding and generation tasks, demonstrating its effectiveness.

1. Introduction

Instruction tuning [9, 38] has shown a great generalization ability on unseen tasks and has contributed to the growing popularity of large language models (LLMs), such as ChatGPT [37]. Recently, multimodal language models benefit from visual instruction finetuning [1, 18, 19, 28, 63], and has shown great success in real-world applications. These models leverage visual encoders such as CLIP-ViT [12, 39] to empower LLMs with image comprehension ability. However, challenges arise in comprehension of textual information within images, which may stem from the prevalence of natural images in training datasets, such as Conceptual Captions [5] and COCO [26]), as highlighted by Liu et al. [30]. Recognizing the importance of visual textual understanding for effective collaboration between agents and humans, Zhang et al. [62] proposed enhancing end-to-end visual instruction-tuned models by introducing noisy Optical Character Recognition (OCR) annotations to improve vision-language alignment. In this work, we surpass existing achievements and collect a new Text-Rich image INStruction dataset named **TRINS**, which contains 39,153 text-rich images, captions and 102,437 questions.

TRINS is created in a semi-automatic manner for a more controllable and faithful collection. Specifically, we exploited large-scale pre-trained models such as CLIP [40] and GPT-4 in the annotation process. This semi-automatic process significantly reduces the time and resources required for manual annotation and surprisingly improves the overall quality of annotations. TRINS dataset is composed of three datasets for captioning, visual question answering (VQA) and image generation, respectively. Specifically, human-annotated captions for text-rich images are first collected because they can best translate text-rich images into texts. During this process, extracted OCR words and recognize-anything model tags are provided to the annotators for better and efficient annotations. With detailed image descriptions, VAQ data is built and fulfilled by large language models, such as GPT-4 [37] and LLaMA-2 [51]. In detailed statistics and analysis, we found that both annotated captions and collected question-answer pairs are more comprehensive and contain significantly more details than the existing dataset. Therefore, we show the superior advantage of TRINS compared to existing instruction fine-tuning datasets. As a by-product, high-quality image-caption pairs can serve as a good benchmark for text-rich image generation, which is still a very challenging task [6]. At the same time, we propose a new simple and effective multimodal language model architecture that includes OCR as a component. We call it Language-vision Reading Assistant (LaRA) and show that LaRA fine-tuned on TRINS brings the best text-rich image understanding ability. Our contributions are as follows:

[1] In this work, we use the phrase "text-rich images" to describe images with rich textual information, such as posters and book covers.
[2] Work done during Q3 2023.

- We introduce a novel dataset (TRINS) containing 39,153 captions and 102,437 high-quality text-rich image instruction pairs. TRINS is annotated with a novel semi-automantic annotation framework that is scalable and reliable.
- We develop several evaluation benchmarks for text-rich image understanding and generation tasks. Various methods are evaluated on TRINS, demonstrating the effectiveness of the dataset.
- The TRINS datasets are high-quality and comprehensive, which is reflected not only in the dataset statistics but also from the results of multiple baseline models. Especially, our proposed LaRA finetuned on TRINS outperforms existing state-of-the-art methods on text-rich image understanding tasks.

2. Related Work

Multimodal Instruction Tuning Multi-modal instruction tuning, including image, video [32, 59], and audio [16, 58] settings, has been an active research topic. MiniGPT-4 [63] uses ChatGPT to generate high-quality instruction-following data, while LLaVA [28] generates such data by prompting GPT-4 with captions and bounding boxes. LLaMA-Adapter [14, 60] aligns text-image features using COCO data, and mPLUG-owl [56] combines extensive image-text pairs for pretraining and a mixture of data for finetuning. Despite this, many models, according to Liu et al. [30], struggle with OCR tasks. InstructBLIP [10] addresses this by transforming 13 vision language tasks into an instruction-following format. mPLUGOwl [55, 56] apply multitask instruction funetuing using existing document datasets. A comprehensive survey is available in Li et al. [20]. LLaVAR [62] exploits GPT-4 to collect fine-tuning data without human annotations using OCR and captioning tools. It discovered that resolution plays a significant role in recognizing textual information and explored several options. Monkey [24] performed a surgery between simple text labels and high input resolution, enabling remarkable performance in visually-rich document images with dense text. TRINS exploits human-machine collaboration for data annotation and can provide more accurate information, reducing the problem of hallucination.

Text-Rich Image Datasets Visual question answering or captioning datasets are widely used in task-specific finetuning and large multimodal model evaluation. TextCap [46] is the first text-rich image captioning dataset. Compared to TextCap, TRINS-Cap provides more detailed annotations that can fulfill the requirement of instruction finetuing. TextOCR [48] aims comprehend text in the context of an image, which is similar to our motivation but focuses more on text recognition in images instead of understanding. ST-VQA [13] uses spatial and textual information to answer visually grounded questions, effectively integrating visual and textual cues. OCR-VQA [36] focuses on incorporating optical character recognition (OCR) into visual question answering (VQA), which operates primarily on text within images. TextVQA [47] also takes advantage of the textual information present in the images to answer questions, but with an emphasis on open questions. DocVQA takes this one step further by applying VQA to document images, handling a variety of layouts and formats. InfoVQA [35] and ChartQA [33] focus on specific subdomains and aim to answer questions about information graphics and chart images, respectively. In summary, these related works provide datasets for leveraging spatial and textual cues. TRINS-VQA is a dataset that exploits the semi-automantic annotation process. It can be used for general domain instruction finetuning and model evaluations.

Dataset	Year	Size	Annotation Type
OCR-VQA	2019	200K	QA (🤖)
TextVQA	2019	45K	QA (👥)
TextCap	2020	140k	Caption (👥)
TextOCR	2021	145k	Text Bbox (👥)
DocVQA	2020	50k	QA (👥)
TRINS (**Ours**)	2023	40k	Caption (👥)
		100k	QA (🤖 + 👥)
		40k	Text Bbox (🤖)

Table 1. Comparison between TRINS and other related datasets.

3. Text-Rich Image Instruction Dataset

To equip multimodal language models with the ability to recognize text and relate it to its visual context, we have curated a new dataset named Text-Rich Instruction (TRINS). The ultimate goal is to enable these models to have spatial, semantic, and visual reasoning between multiple text tokens and visual entities. In this section, we present TRINS, a dataset crafted through a semi-automatic process. Specifically, we leverage large-scale pre-trained models like CLIP [40] and GPT-4 in the annotation process, offering potential advantages: (*i*) Significant reduction in annotation time and resources: using these models significantly reduces the time and resources required for manual annotation. (*ii*) Enhancement of annotation data quality through post-processing: the involvement of large models contributes to improving the overall quality of annotation data through subsequent post-processing. (*iii*) Functionality of large models as knowledge bases: large models can serve as effective knowledge bases, aiding in the annotation process by virtue of their extensive training in diverse datasets.

We first outline the document image collection process for TRINS, utilizing CLIP models. Then, we present data statistics to facilitate a comprehensive understanding. We delve into three distinct tasks derived from the TRINS dataset in detail: *i*) TRINS-Cap: Visual Captioning, *ii*) TRINS-VQA:

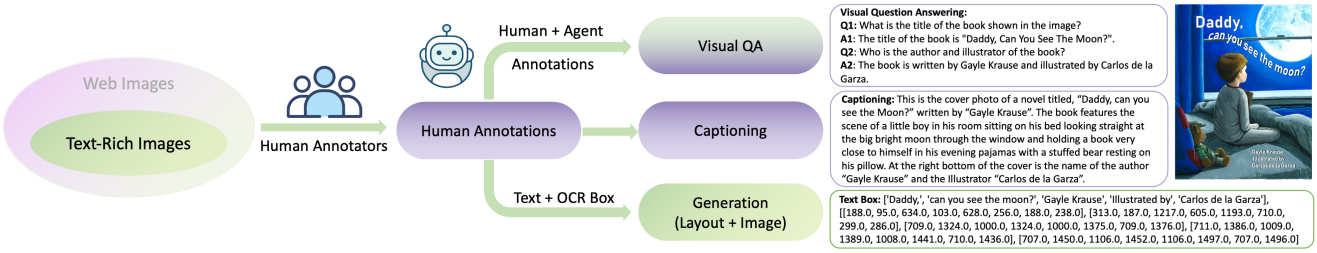


Figure 1. Overview of TRINS data collection process, which consists of three datasets. Text-rich images are first selected from web images and then ask annotators to describe the image in details. *i*) TRINS-Cap is extracted from human annotations with heuristic data processing for text-rich image summarization tasks. *ii*) TRINS-VQA is built upon human annotations and generate question-answer pairs for training by prompting text-only large language models. *iii*) TRINS-Gen combined human annotations and text boxes for text-rich image generation.

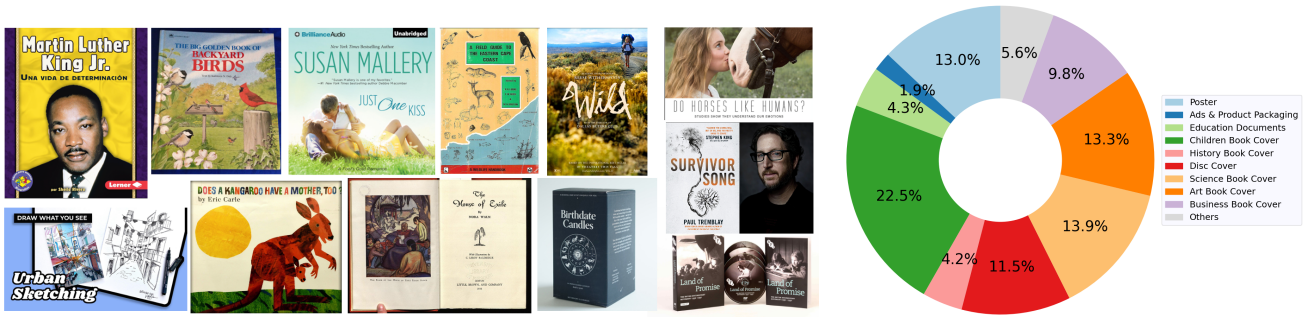


Figure 2. CLIP-based categorization of our collected images and selected representative data samples from each category.

Visual Question Answering and *iii*) TRINS-Gen: Text-to-Image Generation. The overview of the TRINS data collection process is illustrated in Figure 1. To provide a succinct overview, TRINS-Cap undergoes full annotation by human annotators, while TRINS-Gen and TRINS-VQA are constructed based on TRINS-Cap with the help of pre-trained models. A summary of the various TRINS datasets is presented in Table 7 of Appendix A.

3.1. Machine-Assisted Text-rich Image Selections

Beginning with the LAION-5B dataset[3] [44], our objective is to selectively retain images that exhibit a significant presence of text. Recognizing that document images typically contain substantial textual content, we initially formed a binary classification dataset by combining natural images with document data. Subsequently, we trained an image classifier using a DiT [21] base backbone, fine-tuned on the RVL-CDIP dataset [15]. The purpose of this classifier was to predict whether an image contains text. Then a subset was constructed by selecting images with a predicted probability greater than 0.8, while also adhering to the criteria $p(\text{watermark}) < 0.8$ and $p(\text{unsafe}) < 0.5$, where both probabilities are derived from the metadata of the LAION dataset. Acknowledging the noise introduced due to the classifier's limitations, we further refined the dataset by incorporating human judgment. A random sample of 20,000 images

(a) (b)

Figure 3. Word clouds of (a) predicted tags and (b) detected words from the text-rich images of TRINS.

from the filtered LAION-5B was clustered into 50 groups based on CLIP-ViT-B/32 visual features. After inspecting the clustering results, one cluster was meticulously chosen, encompassing diverse text-rich images such as posters, covers, advertisements, and educational documents. This cluster model then served as the filtering mechanism for collecting images that comprise the TRINS dataset. For reference, we present a CLIP-based categorization [39] in Figure 2 to depict the distribution of images in the collected data. The major class is book cover images, further categorized on the basis of book themes and contents. To enhance our understanding of text-rich images, we employed a Recognize Anything Model (RAM) [17, 61] to extract tags from TRINS images. Figure 3a displays word clouds of RAM tags, where "book" and "poster" emerge as major keywords. Additionally, we utilize the Azure Read API [4] and PaddleOCR to extract

[3] https://huggingface.co/datasets/laion/laion-high-resolution

[4] https://azure.microsoft.com/en-us/updates/computer-vision-v3-preview-6/

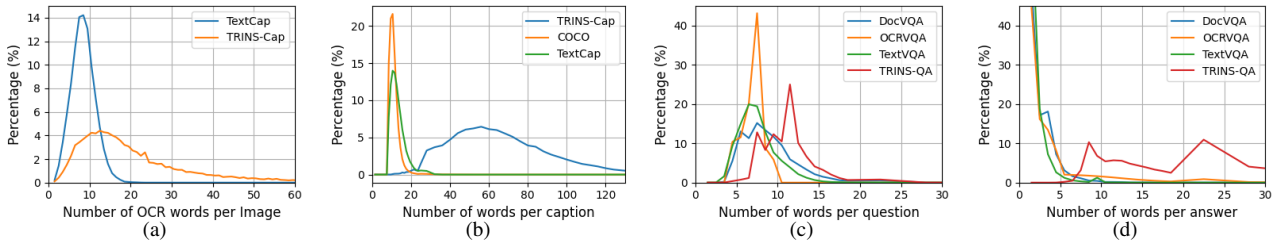


Figure 4. OCR word (a), Caption (b), Question (c) and Answer (d) statistics for TRINS.

text within TRINS images. The word cloud of the extracted texts is presented in Figure 3b. Figure 4a illustrates the distribution of OCR words per image, indicating that most of the images in TextCap have fewer than 10 words, while the TRINS images average 31.4 OCR words. Recognizing texts within TRINS images is more challenging because of the presence of numerous small words. In summary, TRINS images encompass rich visual content, seamlessly integrating textual information into the image context.

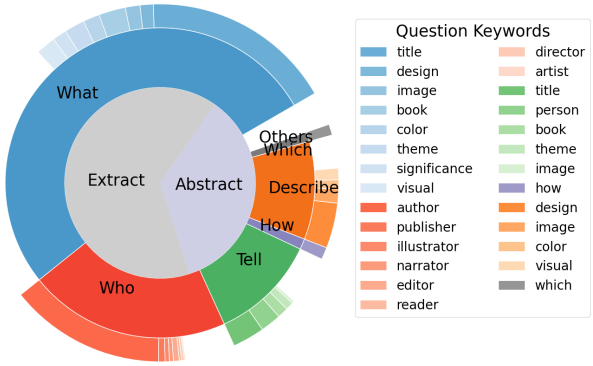


Figure 5. Question type statistics based on key words.

3.2. Annotation Details

Annotator Selections All annotators (with the tag *100% Job Success* and *Top Rated Plus*) are native English speakers and have experience with document annotations. We first asked all annotators to annotate a 200-image set and provide them with detailed annotation guidelines with multiple examples. In addition, we use Labelbox as an annotation tool and set quality control questions.

Heuristic Filters We first use EasyOCR to extract texts from images and retain text phrases with more than three characters, a height greater than 5% of the canvas height, and a confidence score greater than 0.1. For each phrase retrieved, we employ an edit distance-based string matching algorithm (due to potentially erroneous OCR results) to search for its optimal matching substrings within the human-generated caption. The average score for all extracted phrases serves as a metric.

Manual Reviews We accept annotations with high metric scores and reject the lowest for rework. We manually review other annotations.

Sensitive Images We combined neural models with human efforts to filter the images. The first step involves an initial data filtering by 2-3 individuals to filter out sensitive images for training. The second step involves hiring additional people to perform a further screening on the data. We engaged annotators from various countries to check the selected images.

3.3. TRINS-Cap: Text-Rich Image Captioning

Annotation Process TRINS-Cap is a dataset fully annotated by human annotators. We hired 20 native English speakers with experience in document annotation through Upwork. The annotation process, conducted in LabelBox, involved a total of 2,079 hours to annotate 40,576 text-rich images, with an additional 159 hours allocated for result review. After filtering low-quality annotations and addressing missing images, we obtained a final set of 39,153 image-annotation pairs. The dataset is partitioned into train, validation, and test splits with sizes of 29,153, 5,000, and 5,000, respectively. All annotations undergo an initial automated review that involves matching the OCR words with the annotations. Subsequently, human evaluators conduct a thorough review, rejecting annotations with errors, and prompting annotators to rework them for enhancement. We provide comprehensive annotation instructions to all annotators to ensure that each annotation includes: (*i*) detailed descriptions of visual components. (*ii*) describe texts' location, attributes, and put texts into annotations. (*iii*) optional insights or abstract descriptions.

Statistics and Analysis The primary objective of the annotation process is to facilitate a human or machine's full comprehension of the information conveyed in the image without direct viewing. Consequently, the average annotation length for TRINS is 65.1 words, significantly exceeding that of COCO (10.6 words) and TextCaps (12.4 words). Figure 4b shows the caption length distributions for TRINS-Cap, COCO, and TextCap, demonstrating the comprehensive nature of the TRINS-Cap annotations. TRINS with more contexts can generally provide a better description of complex images, where short captions are insufficient. Hence, LLMs finetund on TRINS can better understand images with complex texts and layouts, which has been further verified in Section 5.

3.4. TRINS-VQA: Multimodal Question Answering

The annotation process for question answering is inherently complex, primarily due to the necessity for annotators to generate high-quality questions. Creating an effective question is more challenging than providing an answer. As a result, annotators frequently gravitate toward formulating concrete and extractive questions (*e.g.*, "Who is the author of this book?") rather than abstract ones (*e.g.*, "How does the design of the book cover reflect the content of the book?"). We introduce semi-automatic annotation methods to generate high-quality visual question-answering data for TRINS-VQA. This dataset is designed to train general vision language assistants through instruction fine-tuning, and its benefits on model performance are evaluated in Section 5.1. TRINS-Cap, on the other hand, serves as human-assisted annotations, offering a comprehensive but non-instructive dataset for fine-tuning. To utilize the wealth of high-quality annotations available, we incorporated semi-automatic annotation by using large language models (LLMs) such as OpenAI's GPT-4 [37] and Llama-70B [50, 51] to enhance our data annotation pipeline. OCR results and detailed descriptions of each image are provided to LLMs. Furthermore, high-quality human-crafted demonstrations and detailed annotation rules are provided to LLMs. One demonstration focused on extract questions, while the other emphasized abstract questions, creating a more balanced dataset.

Human Annotations To facilitate a robust evaluation of model performance, we hired 10 Upwork annotators, whose native language is English, to annotate the test dataset, following a methodology similar to previous work [13, 34, 47]. The test dataset comprises 5,000 images with 18,764 question-answer pairs. These data collected are used exclusively for evaluation purposes.

Statistics and Analysis Building upon prior research [28, 49, 53], we provide visualizations of instructions in Figure 5 based on question keywords. The inner cycle illustrates the distribution of the first word in the questions, while the outer cycle presents extracted keywords determined by carefully designed heuristics. Types of questions are categorized according to keywords found in questions.

In Figure 4c and 4d, we present statistics on the number of words per question-answer pair, comparing them with previous work. Generally, the average length of questions for TRINS-VQA is 10.5, surpassing that of DocVQA (8.3), OCR-VQA (6.5), and TextVQA (7.1). Surprisingly, the average answer length for TRINS is 23.9, significantly longer than related datasets (all less than 4). This discrepancy arises from TRINS containing more abstract questions that typically have longer answers. Similarly, the dataset is divided into train, validation, and test splits. For extract questions, the accuracy of the answers is calculated similarly to Liu et al. [28], while for abstract questions, generation metrics such as BLEU scores are used to assess the quality of the answers.

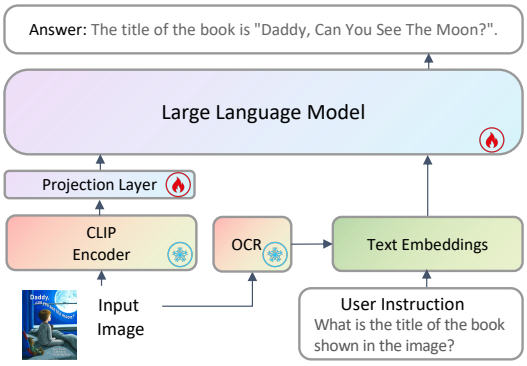


Figure 6. Model overview of the LaRA. the CLIP model processes the input image to generate patch-wise features. These features then serve as input to a projection layer, yielding visual tokens. Concurrently, an OCR tool extracts textual data from the image, which is then merged with the user instruction.

The question-answering and instruction data we obtain are extensive, encompassing a balanced mix of extract and abstract questions. This includes detailed descriptions, summaries, question-answer pairs, tasks that promote creativity and the generation of novel thoughts, and conversational tasks. The dataset spans a diverse range of concepts, ranging from visual presentation and visual language relations to intricate reasoning tasks. Compared to previous methods [28], captions generated by models such as BLIP-2 [22] exhibit less informativeness than detailed human annotations, as demonstrated in Section 5.2. Additionally, captioning models may be susceptible to hallucinations [42], a concern mitigated in TRINS, which provides more comprehensive and reliable visual descriptions. Although OCR tools are robust, they can introduce noisy information. To address this, we utilize both the Azure Read API and PaddleOCR to extract text information. We added the potential unreliability in our system prompts to LLMs, instructing them to generate questions with assured answers. Ultimately, we directly leverage the responses from large language models (LLMs) to construct TRINS-VQA. The quality of the instruction data can be further enhanced through self-alignment [23] or by seeking verification from human annotators. Although this has not been explored in this work, we leave it as a potential direction for future research.

4. The Language-vision Reading Assistant (LaRA) Model

To verify our data, we adopt a simple architecture based on LLaVA [28] for language-vision reading assistant, as illustrated in Figure 6. A low-resolution visual encoder is a significant bottleneck for extracting textual information from images [62]. On the contrary, a high-resolution

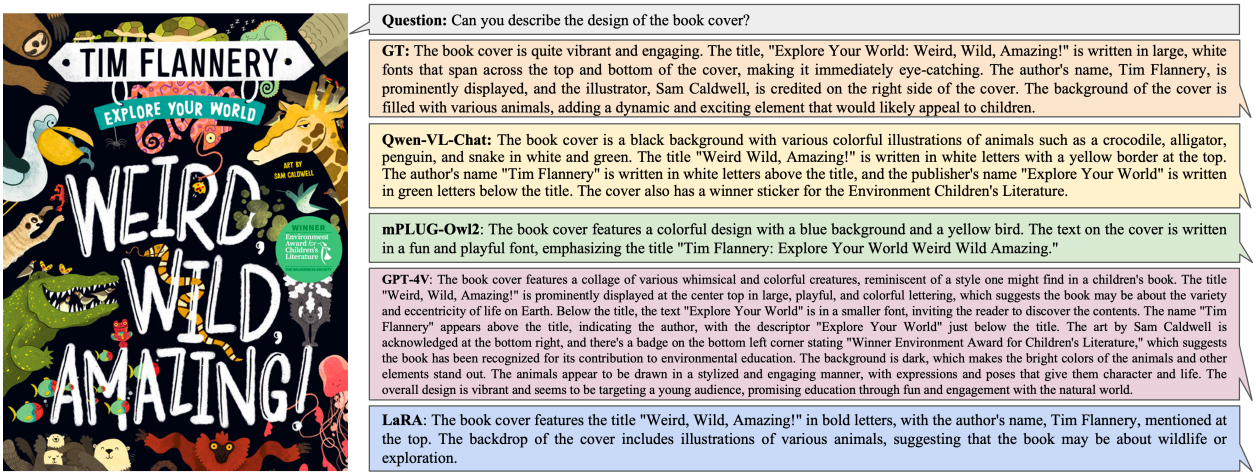


Figure 7. Examples generated by different multimodal language models on the abstract TRINS-VQA benchamrk.

visual encoder generates an excessive number of patches, drastically impacting model efficiency. Our conviction is that a low-resolution visual encoder is designed to capture visual information more effectively, including layouts. In contrast, a considerably smaller OCR tool is employed to extract text from high-resolution images. Instruction-tuning on TRINS could serve as an effective method of training LLMs to better align OCR texts and extracted visual features. In addition, LLMs can autonomously rectify errors generated by OCR tools. For the visual encoder V, we employ `CLIP-ViT-L/14-336` at a resolution of 336×336. The grid features before the last transformer layer are then mapped into the word embedding space of the language decoder using a trainable projection matrix W. Regarding the language decoder D, we utilize Vicuna-1.5-13B [8], a language model tuned for instructions through LLaMA 2 [50].

We follow the two-stage training design of LLaVA but adopt the pre-trained projection layer of LLaVAR [62]. Training targets remain similar: generating ***output responses*** ($<res>$) for the ***input instructions*** ($<ins>$), alongside OCR results ($<ocr>$). The transformed image tokens ($<img>$) are introduced before or after the first input instruction randomly when building the instruction finetuning data. During the finetuning stage, both the projection matrix W and the language decoder D are trained. We consolidate our nearly 90K visual question-answering data with the 158K instruction-following data from LLaVA to form the training set. It should be noted that the visual encoder remains frozen throughout the training period. Compared with previous approaches, LaRA incorporates OCR words as part of the input, a simple way to enhance visual text understanding.

5. Experiments

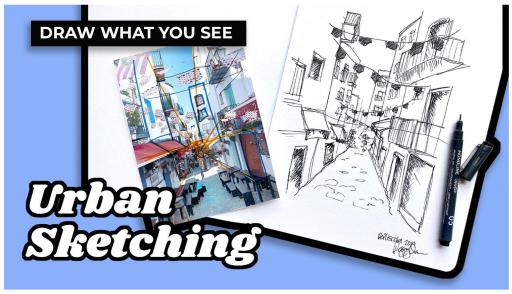
In this section, we present three downstream tasks based on the TRINS dataset and outline their evaluation metrics. The proposed method LaRA is used in both the text-rich image summarization and visual question-answer tasks. All experiments were conducted on NVIDIA A100 80GB GPUs. In fine-tuning, we use a cosine annealing schedule with an initial learning rate of $2e^{-5}$ and a batch size of 32.

Method	Recog.	VQA^S	VQA^D	KIE	Final Score
Gemini	**215**	**174**	128	134	**651**
GPT-4v	167	163	**146**	**160**	636
Monkey	174	**161**	**91**	88	514
mPLUG-Owl2	153	153	41	19	366
LLaVAR	186	122	25	13	346
LLaVA1.5-13B	176	129	19	7	331
mPLUG-Owl	172	104	18	3	297
MiniGPT-V2	124	29	4	0	157
LaRA	**211**	147	85	**105**	**548**

Table 2. Results of LMMs on OCRBench. Recog. represents text recognition, VQA^S represents Scene Text-Centric VQA, VQA^D represents Document-Oriented VQA.

5.1. TRINS-VQA: Text-Rich Image Visual Question Answering

We first performed experiments to evaluate the zero-shot performance of LaRA on classical benchmarks [28, 54]. The results are reported in Table 2 and Table 10. The proposed LaRA model exhibits significant performance improvement across all the datasets compared to other models. Even without OCR, LaRA outperforms other models in most cases, highlighting its robustness and effectiveness in handling visual question-answering tasks. The inclusion of OCR in LaRA further enhances performance, suggesting that Large Language Models (LLMs) can proficiently utilize textual information. However, the ability to directly extract text from images remains limited. LaRA, with its simple structure, significantly boosts model performance on text-rich images, offering an alternative solution to overcome the limitations of pre-trained image encoders. We further evaluate different



GT: There are big white words "Urban Sketching" at the bottom left of the cover. In the upper left corner of the cover, it reads "Draw What You See." In the background of the picture, there is a drawing book and pen. It shows a signature in the bottom right corner of the book.

LLaVA: The image is an urban sketching project displayed on a tabletop, featuring a variety of hand-drawn city scenes. Several people can be seen in the drawing, walking around and engaging in different activities. Among the drawings, there is an image of an alleyway with a prominent Umbrella on the right side.

LLaVAR: The main subjects in this image are a notebook with a pen placed on top of it, and a photograph of a city street. The notebook and pen appear to be the focal point, with the urban scene in the background.

Instruct-BLIP: The main subjects in the image are a series of sketches and drawings depicting urban scenes.

LaRA: This is a blog image. Black text in the upper left corner reads "Draw What You See". Large white text in the bottom left reads "Urban Sketching". The image background is a blue spiral notebook with a pen in the upper right and a picture of a colorful urban scene in the middle.

Figure 8. Examples generated by different multimodal language models on the TRINS-Cap benchamrk.

Method	Resolution	Extract Accuracy	Abstract						
			B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr
Instruct-BLIP [11]	224^2	43.9	13.5	9.2	6.7	5.2	8.9	16.7	23.5
Mini-GPTv2		15.3	29.4	19.4	13.8	10.4	17.2	28.7	44.3
mPLUG-Owl2 [57]	448^2	61.0	35.9	25.5	19.3	15.2	17.7	38.0	97.2
Qwen-VL [3]		**63.6**	40.2	27.3	19.9	15.2	20.4	36.4	79.4
LLaVA [28]	336^2	23.7	33.1	20.9	14.4	10.5	18.0	31.8	48.2
LLaVAR [62]		51.7	38.0	25.9	19.1	14.7	17.4	35.6	84.9
LLaVAR w/ OCR		58.1	40.3	28.0	20.9	16.3	18.7	37.3	97.3
LLaVA 1.5 [27]		38.8	29.8	20.9	15.8	12.6	20.8	34.9	40.7
LLaVAR (finetuned)	336^2	61.2	**45.3**	31.9	23.9	18.7	20.4	39.6	104.7
LaRA		62.8	45.1	**34.2**	**27.3**	**22.6**	**21.9**	**46.5**	**186.6**

Table 3. Results of different models on TRINS-VQA for text-rich image question-answering tasks.

methods on the TRINS-VQA dataset, as shown in Table 3. For extraction questions, we use the same metric as Wu et al. [54]. For abstract questions, where the answer is typically a longer sentence, we evaluate them based on text similarity metrics such as BLEU [25], ROUGE [25], and CIDEr [52]. In zero-shot inference, LLaVAR with OCR exhibits the best performance, reinforcing the importance of extracting textual information. Furthermore, mPLUG-Owl2 and Qwen-VL perform well and represent the best methods in extract question evaluations, showing that a high-resolution encoder can significantly improve model performance. Instruct-BLIP demonstrates good performance on extract questions, but did not fare as well on abstract questions, given that the answers provided are usually short and concise. Figure 7 shows an example of the responses of different methods on the abstract TRINS-VQA dataset, and more examples can be found in the Appendix D. Qwen-VL includes all details but does not provide high-level insights, such as the ground-truth annotation. Both mPLUG-Owl2 and GPT-4V suffer from hallucination issues.

5.2. TRINS-Cap: Text-rich image Captioning

In our experiments on TRINS-Cap, we ask large multimodal models to generate summaries based on text-rich images. The data set was divided into train, validation, and test sets. We compared LaRA with popular baselines, including InstructBLIP [11], Mini-GPT4 [63], LLaVA [28], LLaVAR [62], mPLUG-Owl2 [56] and Qwen-VL [3]. Given that BLIP-2 faced challenges in generating comprehensive and meaningful results for text-rich images, we considered InstructBLIP as an alternative. For all methods, we randomly selected three prompts from ten as instructions for the model (details provided in the Appendix C.1).

Table 4 presents the results of different methods in terms of classical captioning metrics. Models with enhanced visual text understanding generally outperform general multimodal models, such as LLaVA, Mini-GPT4, and Instruct-BLIP. LaRA (zero-shot) refers to the LaRA model fine-tuned on the TRINS-QA dataset and demonstrates improved performance. Comparison of fine-tuned LaRA variants indicates that text recognition ability is still limited for OCR-free methods, suggesting that the CLIP encoder or feature projection process may cause visual information loss. Addressing this limitation may involve employing a better trained encoder on text-rich images or designing a more carefully crafted architecture, a direction we leave for future exploration. When fine-tuned with TRINS-Cap, LaRA exhibits much better performance, underscoring the importance of high-quality human-annotated data. Figure 8 shows examples of different models on the TRINS-Cap benchmark. It shows the great capability of LaRA in recognizing text and relating it to its visual contexts, demonstrating the effectiveness of the TRINS dataset.

Method	Backbone	B@1	B@2	B@3	B@4	METEOR	ROUGE	CIDEr
mPLUG-Owl2 [57]	Llama-2-7B	4.9	3.2	2.3	1.7	8.7	19.6	4.5
Instruct-BLIP [11]	Vicuna-7B	13.4	8.5	5.7	4.1	9.2	17.9	5.7
Qwen-VL [3]	Qwen-7B	28.8	18.4	12.2	8.6	14.1	24.3	16.6
Instruct-BLIP [11]	Vicuna-13B	15.9	9.9	6.4	4.4	9.6	18.6	8.0
Mini-GPT4 [63]		31.1	16.1	8.6	5.0	11.4	20.8	6.3
Mini-GPT-v2 [7]		27.9	14.7	8.0	4.8	10.8	20.9	7.3
LLaVA [28]		35.1	18.2	9.5	5.4	13.2	22.2	8.8
LLaVAR [62]		18.9	11.2	7.2	5.0	10.8	20.1	11.4
LLaVA 1.5 [27]		31.1	16.4	9.5	6.1	11.9	21.8	15.4
LLaVAR w/ OCR	Vicuna-13B	21.4	13.1	8.8	6.4	12.0	22.2	13.0
LaRA (zero-shot)		29.3	19.1	12.9	9.2	14.8	26.3	21.3
LLaVAR (fine-tuned)	Vicuna-13B	36.5	25.4	18.0	13.4	17.8	32.4	35.7
LaRA		**37.7**	**26.4**	**18.9**	**14.2**	**18.4**	**33.2**	**46.7**

Table 4. Results of different models on text-rich image captioning tasks.

5.3. Additional Experiments

Performance on general visual tasks after TRINS fine-tuning. We adopted the evaluation protocols of MiniGPT-v2 [7] and compared LaRA with LLaVA [28] on traditional visual question answering benchmarks in table 5. LaRA shows a comparable performance on knowledgeability and better performance on reasoning and spatial awareness. This further verifies the effectiveness of the TRINS dataset, demonstrating that fine-tuning on text-rich images does not degrade performance on natural images, but instead enhances the results.

	OKVQA	GQA	VSR	VizWiz
LLaVA [28]	57.8	41.3	51.2	45.0
LaRA	**58.1**	**42.4**	**53.0**	**53.1**

Table 5. Quantitative Results on the public visual benchmarks.

Metrics	ControlNet	DeepFloyd	TextDiffuser
FID (↓)	51.59	**49.96**	51.26
CLIP Score (↑)	0.3717	**0.3917**	0.3707
OCR Acc. (↑)	0.4241	0.2192	**0.5027**

Table 6. Empirical Results on TRINS-Gen (easy) benchmark.

Text-to-document generation Diffusion-based text-to-image generation has shown great success, while precise textual renderings remain a big challenge. TextDiffuser introduced the MARIO-Eval benchmark, drawing from works such as DrawBench [43] and DrawTextCreative [29]. However, most text prompts in MARIO-Eval are short and cannot serve as a good evaluation dataset to handle complex real-world human instructions. We take advantage of human annotations from TRINS-Cap and build the TRINS-Gen benchmark. It is still difficult to render too many words in a single image [6]. In response to this, we filter out images with more than 20 OCR words, resulting in a curated set of 2,104 images. We divide these images into two sets (easy and difficult) based on the number of OCR words and the length of the longest OCR string per annotation, where all text prompts in the easy set have less than 9 OCR words. We evaluated existing methods using their public checkpoints and reported the results in Table 6 and detailed results in Table 9 (Appendix C.3).

Figure 9. Examples generated by different text-to-image models on the TRINS-Gen benchamrk.

6. Conclusions

Despite the challenges posed by the prevalence of natural images in training data, the significance of visual textual understanding cannot be understated. In this paper, we introduce TRINS, a Text-Rich Image INStruction dataset, comprising a diverse collection of text-rich images, captions, and questions. This dataset, created through a semi-automatic process leveraging large-scale pre-trained models, not only significantly reduces annotation time but also elevates annotation quality. Furthermore, we propose a novel multimodal language model architecture, LaRA, which incorporates OCR as a pivotal enhancement for text-rich image understanding. We anticipate that continued progress in multimodal language model architectures, fine-tuning techniques, and the expansion of diverse, text-rich datasets like TRINS will push the boundaries of visual textual understanding. This, in turn, will facilitate more efficient collaboration between humans and agents, potentially revolutionizing numerous real-world applications.

References

[1] 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. 1

[2] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, 2023. 22

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7, 8

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 12

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. 1

[6] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023. 1, 8, 12

[7] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 8

[8] 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, 2023. 6

[9] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. 1

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023. 7, 8

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. 1

[13] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. 2, 5

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023. 2

[15] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. Evaluation of deep convolutional nets for document image classification and retrieval, 2015. 3

[16] Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jia-Bin Huang, Jinglin Liu, Yixiang Ren, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. *ArXiv*, abs/2304.12995, 2023. 2

[17] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023. 3

[18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1

[19] Chunyuan Li. Large multimodal models: Notes on cvpr 2023 tutorial. *ArXiv*, abs/2306.14895, 2023. 1

[20] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1, 2023. 2

[21] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Dit: Self-supervised pre-training for document image transformer. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 3

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 5, 22

[23] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023. 5

[24] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 2

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7, 8

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 5, 6, 7, 8, 22

[29] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*, 2022. 8, 12

[30] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Cheng lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models, 2023. 1, 2, 22

[31] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently, 2023. 12

[32] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2023. 2

[33] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2

[34] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2020. 5

[35] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2

[36] 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. IEEE, 2019. 2

[37] OpenAI. Gpt-4 technical report, 2023. 1, 5

[38] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 1

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 3

[40] 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. 1, 2

[41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 12

[42] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 5

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 8, 12

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[45] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd. https://github.com/deep-floyd/if, 2023. 12

[46] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 2

[47] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[48] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 2

[49] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 5

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 5, 6

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 5

[52] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7

[53] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2022. 5

[54] Zizhang Wu, Xinyuan Chen, Jizheng Wang, Xiaoquan Wang, Yuanzhu Gan, Muqing Fang, and Tianhao Xu. Ocr-rtps: an ocr-based real-time positioning system for the valet parking. *Applied Intelligence*, pages 1–15, 2023. 6, 7

[55] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 2

[56] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 2, 7, 22

[57] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, 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. 7, 8, 22

[58] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023. 2

[59] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[60] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023. 2

[61] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 3

[62] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 1, 2, 5, 6, 7, 8, 22

[63] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1, 2, 7, 8, 22