Towards Improving Document Understanding: An Exploration on Text-Grounding via MLLMs

Anonymous ACL submission

Abstract

Recently, significant advances have been made in the fine-tuning of Multimodal Large Language Models (MLLMs) using instruction-following data. However, the potential of text-grounding capabilities within text-rich scenarios remains underexplored. In this paper, we investigate the impact of text-grounding on MLLM's understanding of text-rich images by examining whether Visual Question Answering (VQA) data contains text-grounding information. Empirical evidence implies that text-grounding provides positional information of the answers, thereby enhancing the model's understanding and interpretability of images. To this end, we curate a dataset containing 99,000 PowerPoint images sourced from the internet and formulate instruction tuning tasks, including text detection, recognition, and spotting, to facilitate cohesive alignment between the visual encoder and large language model. Moreover, we curate a collection of high-resolution text-rich images and prompt the GPT-4 to generate 12,000 high-quality conversations featuring textual locations. By integrating text-grounding data, our model adeptly discerns text locations during the VQA process. Despite our data being only 10% larger than that of LLaVAR, our method demonstrates notable improvement over LLaVAR in various benchmarks (14.2% point absolute improvement on the STVQA dataset and 11.46% point absolute improvement on the OCRVQA dataset), thereby validating the effectiveness of our method.

1 Introduction

Large Language Models (LLMs) have showcased remarkable generalization capabilities through fine-tuning with instruction-following data. Motivated by this, researchers have extended these capabilities into multimodal domains, especially vision and language (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2024b). However, due to the lack of relevant text-rich images data, these models struggle to comprehend text within images. To address this issue, recent studies (Zhang et al., 2023b; Ye et al., 2023a; Feng et al., 2023b; Li et al., 2024; Ye et al., 2023b; Hu et al., 2024; Feng et al., 2023a) have endeavored to enhance models with Optical Character Recognition (OCR) capabilities to improve their understanding of text-centric scenarios.

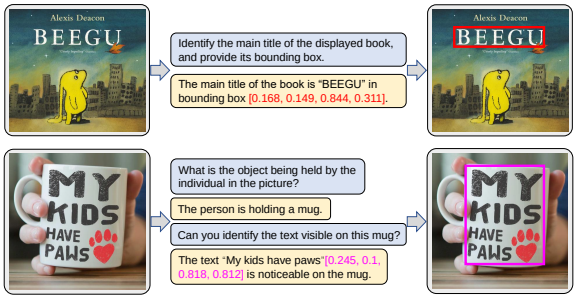

Figure 1: VQA in text-rich images with text-grounding. Given an image and a query, the model generates responses in an autoregressive way, and concurrently outputs the basis of reasoning, *i.e.*, the bounding box that delineates the location of the answer.

As the multimodal domain advances, researchers focus more on the development of general-purpose models. GPT-4V (Yang et al., 2023) holds a prominent position in the commercial model landscape, leading the industry. To keep pace, numerous open-source models (Bai et al., 2023; Wang et al., 2024; Chen et al., 2024; Dong et al., 2024; Lin et al., 2024; Xue et al., 2024) endeavor to achieve universal Artificial General Intelligence (AGI), marking significant milestones in the open-source field. These general-purpose models, along with certain specialized models (Chen et al., 2023b; Zhang et al., 2023a), possess visual grounding capabilities, allowing them to incorporate spatial information about objects in images. This advanced feature facilitates more effective user-agent interactions and mitigates ambiguity (for example, when two apples appear in an image, the user may be unsure which apple the model refers to).

We argue that grounding is an important tool for MLLMs to answer user questions accurately. Experiments from Shikra (Chen et al., 2023b) demonstrates that including center point locations for objects can alleviate hallucinations and improve the precision of its responses. However, the potential of text-grounding as a tool for MLLMs is still less explored. Generally, endowing text-grounding capabilities into a model can enhance its interpretability, making the generated answers more convincing. Moreover, within the embodied AI community, the integration of text-grounding improves the interactive experience between users and agents. As illustrated in Figure 1, our objective is to devise a model that can precisely answer a question related to an image while concurrently identifying and elucidating the specific text region within the image that corresponds to the rationale behind the response. We contend that text-grounding can further unearth the model's potential and unlock the wealth of world knowledge embedded within it.

In this paper, we focus on investigating the importance of text-grounding in understanding text-rich scenarios and propose our model, TGDoc. Note that recent models generally follow a common structure of $visual\ encoder \rightarrow adaptor \rightarrow LLM$, a blueprint that we continue to follow. We compile data for our experiments. Specifically, during the pre-training stage, we curate 99,000 PowerPoint presentations from the internet for image-text alignment, cultivating the model's proficiency in text detection and recognition across diverse scenarios. During fine-tuning stage, we prompt GPT-4 with OCR results and image captions to produce 12,000 high-quality, multi-turn conversations with text-grounding. Additionally, we incorporate all data used by LLaVAR (Zhang et al., 2023b) during both two stages. Experimental results indicate that using instruction tuning with the collected instruction-following data, our method shows improved performance on understanding text-rich scenes and exhibits text-grounding capability.

Our contributions can be summarized as follows:

- We conduct an exploration of text-grounding in MLLMs, and substantiate its critical role in understanding text-rich images.
- We develop an data collection methods to curate text-grouding dataset, and the instructing-following dataset comprising 99,000 PowerPoint images and 12,000 text-rich images with high-quality conversations.
- Extensive experiments show that our method achieves superior results on several benchmarks, validating the importance of text-grounding capability in MLLMs.

2 Related Work

LLMs through instruction tuning have demonstrated impressive zero-shot capabilities for new tasks (Ouyang et al., 2022; Wang et al., 2022; Chung et al., 2024; Xu et al., 2023; Touvron et al., 2023). LLaMA (Touvron et al., 2023), a prominent open-source model, has drawn attention in research initiatives (Taori et al., 2023; Chiang et al., 2023). However, it is noteworthy that these studies exclusively accept the text as input. Recently, people have extended LLMs to the multimodal domain.

LLaVA (Liu et al., 2024b) and MiniGPT-4 (Zhu et al., 2023) employ the $visual\ encoder \rightarrow adaptor \rightarrow LLM$ paradigm to align image modalities with text embeddings. Subsequently, other MLLMs (Ye et al., 2023c; Chen et al., 2023a; Liu et al., 2023a; Ye et al., 2024) have adopted this paradigm with some modifications to explore various aspects of MLLMs. Concurrently, Qwen-VL series (Bai et al., 2023; Wang et al., 2024), InternVL series (Chen et al., 2024; Dong et al., 2024), and other MLLMs (Lin et al., 2024; Xue et al., 2024) aim to achieve a general-purpose to excel the multilingual question-answering tasks. Additionally, the most powerful model, GPT-4V (Yang et al., 2023), has shown unprecedented proficiency in handling intricate multi-modal inputs, inlcuding text, image, and audio.

With the impressive performance of MLLMs in visual-language tasks (Tiong et al., 2022; Driess et al., 2023; Li et al., 2022), the field of document understanding has shifted from supervised methods (Xu et al., 2020b; Wang et al., 2021; Xu et al., 2020a; Yu et al., 2023; Koo, 2016; Li and Tang, 2016; Yang et al., 2020; Wang and Yoon, 2021; Li et al., 2021) to generative approaches (Zhang et al., 2023b; Ye et al., 2023a). Currently, it has been observed that higher resolution significantly bolsters the comprehension of MLLMs for text-dense documents. Many works (Li et al., 2024; Ye et al., 2023b; Feng et al., 2023a; Hu et al., 2024) have explored this field and have showcased remarkable efficiencies in document understanding.

The exploration of grounding capabilities within MLLMs has garnered significant attention from researchers in recent years (Chen et al., 2023b; Peng et al., 2023; You et al., 2023; Zhang et al., 2023a).

Some studies have focused on introducing new referential dialogue tasks, while others have aimed to enhance the perception and recognition of objects within images, improving object boundary detection and mitigating the performance decline typically observed in grounding following chat. However, the integration and exploration of text-grounding capability within MLLMs have been relatively limited in scope. In this study, we delve into the analysis of detailed location information for text within MLLMs. Our empirical findings indicate that incorporating text-grounding can boost the understanding and interpretability of the model in text-rich scenarios.

3 Methodology

3.1 Model Architecture

Following LLaVA (Liu et al., 2024b), the overall architecture of our model consists of three components: a vision encoder, a linear projection layer, and a large language model. Specifically, we employ CLIP (Radford et al., 2021) as vision encoder, configured to process images at two resolutions: 224×224 and 336×336. The alignment module contains a single projection layer and is designed to transform the image features into the word embedding space compatible with the language decoder. For the large language model, we select the Vicuna-7B (Chiang et al., 2023), an instruction-tuned variant based on LLaMA (Touvron et al., 2023) to enhance our language understanding capabilities.

More specifically, given an image $\boldsymbol{I}^{H\times W\times C}$, we initially process it using a vision encoder to extract the image features $\boldsymbol{F}^{256\times 1024}$. In this paper, we opt for grid features before the last transformer layer to serve as our image features. These features, as opposed to those from the last layer, exhibit an enhanced ability to discern intricate details in the image (Liu et al., 2024b). This attribute is especially beneficial for the recognition of text in text-rich environments. Subsequently, these image features $\boldsymbol{F}^{256\times 1024}$ are transformed through a linear projection layer into the word embedding space, resulting in transformed features $\boldsymbol{F}_{trans}^{256\times 4096}$ to align with the language decoder. Moreover, user queries are tokenized within the same word embedding space and seamlessly concatenated with the visual embeddings to form uniform input vectors for the large language model. Consistent with the previous methods (Zhang et al., 2023b; Ye et al., 2023a; Feng et al., 2023b; Chen et al., 2023b), we continue to train the model with the next-token prediction task, *i.e.*, generating the next token based on the sequence of preceding tokens.

3.2 Grounded Input Representations

We adopt the natural language formatted notation $[x_{min}, y_{min}, x_{max}, y_{max}]$ to represent the bounding boxes coordinates. In this context, $[x_{min}, y_{min}]$ signifies the top-left corner of the text's minimum bounding rectangle, while $[x_{max}, y_{max}]$ corresponds to the bottom-right corner. We normalize the bounding boxes to $[0, 1]$ relative to the image size and maintain a three-decimal precision for each coordinate. The bounding box notion can be integrated seamlessly into both the prompts and responses, encapsulated as "<text>$[x_{min}, y_{min}, x_{max}, y_{max}]$". We employ the model to predict the coordinates in a manner analogous to natural language prediction, without introducing additional tokens or detection modules.

3.3 Data for Pre-training

We indicate that the data collection criteria stipulate not only textual richness but also diversity in content. We select PowerPoint slides as our instruction-following training data, due to their multiple advantages. Firstly, the slides are rich in structured text, featuring diverse fonts and artistic texts. Unlike text from natural scenes, the content in PowerPoint slides is clearer and more organized, which markedly improves its compatibility with Optical Character Recognition (OCR) technique. This clarity reduces the need for manual corrections, thus simplifying the data preprocessing phase. Secondly, PowerPoint slides encompass various non-textual elements like graphs and landscapes, enriching the diversity of image types. Thirdly, the slides contain interleaved visual and text information, such as scene photographs and product images, which embed the text within an narrative context and foster the model's comprehension of complex image-text scenarios.

In the construction of PowerPoint datasets, we consciously avoid automated generation techniques to prevent the risk of data homogeneity. We choose to source data from the website, SlideShare [1], a public platform for sharing presentations and multimedia materials. Our collection spans various domains such as business, education, and athletics. We apply the MD5 hashing algorithm (Rivest,

[1] https://www.slideshare.net

Captions generated by BLIP-2
"babies are from airports", "babies come from airports", "the cover of babies come from airports".
PaddleOCR Results
"COME"[0.504,0.049,0.711,0.130], "BABIES"[0.084,0.067,0.496,0.190], "AIRPORTS"[0.350,0.138,0.934,0.298], "FROM"[0.150,0.193,0.367,0.270], "Arrivals="[0.128,0.318,0.381,0.400], "Written by Erin Dealey lllustrated by Luciana Navarro Powel"[0.136,0.909,0.887,0.930].
EasyOCR Results
"BABI-S CONE"[0.063,0.032,0.733,0.208], "FROM"[0.144,0.186,0.371,0.279], "Writien"[0.133,0.907,0.225,0.933], "Erin Dealey"[0.261,0.902,0.408,0.938], "Illustrated by Luciana Navarro Powell"[0.438,0.905,0.893,0.938], "AIRPORTS"[0.353,0.124,0.940,0.304], "Arrivals"[0.120,0.310,0.340,0.398], "by "[0.219,0.902,0.268,0.939].

Final conversation
Q1: Identify the main title displayed in the image.
A1: The main title in the image is "BABIES COME FROM AIRPORTS"[0.084, 0.049, 0.934, 0.298].

Q2: Are there any distinct sections denoted in the background? If so, please specify and support your response with the text and its bounding box.
A2: Yes, it seems that the babies are from airports and the airport stop says "Arrivals"[0.12, 0.31, 0.34, 0.398].

Q3: Can you name the creators of the content depicted in the image? Provide the relevant text and its bounding box for justification.
A3: The creators are "Written by Erin Dealey"[0.133, 0.902, 0.408, 0.938] and "Illustrated by Luciana Navarro Powell"[0.438, 0.905, 0.893, 0.938].

Figure 2: The construction of the instruction-following dataset during the fine-tuning stage. Given an image, we employ BLIP-2 (Li et al., 2023) to produce three captions. In parallel, we extract text and bounding boxes from the image using two OCR engines. These captions and OCR results are then combined to form prompts, guiding GPT-4 in generating multi-turn conversations. To maintain data quality, these conversations are manually reviewed and refined, removing any content unrelated to the image.

1992) to eliminate any duplicates. We then utilize PaddleOCR [2], an open-source tool, to extract both text and spatial bounding box data from each slide. To ensure high data quality, we impose a threshold requiring that each slide must feature at least one block of text accounting for a minimum of 5% of the slide's total image area. This rigorous process yields data consisting of 99,000 image-text pairs.

With respect to the accumulated PowerPoint slides data, we develop three distinct types of tasks for each image: recognition, detection, and spotting. The instructions for these tasks are: "Identifying all text along with its corresponding bounding box in the image", "Locating the box containing the word <text>", and "Please recognize and provide the text contained within the specified bounding box $[x_{min}, y_{min}, x_{max}, y_{max}]$", respectively. To broaden the diversity of instructions, we utilize GPT-4 for query augmentation, thereby bolstering the model's zero-shot generalization ability (see Appendix A for more instructions).

3.4 Data for Fine-tuning

To improve the model's proficiency in specific tasks and deepen its understanding and execution of standard VQA tasks, we carefully design a total of 12,000 high-resolution images with multi-turn conversations. Similarly to the pre-training stage, this dataset integrates textual content with associated

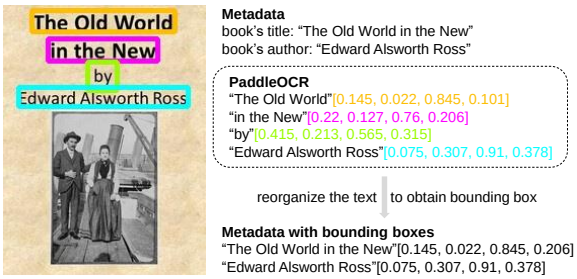

Figure 3: Acquisition of bounding boxes for the metadata of book cover. We employ the PaddleOCR tool to extract texts and position information from the image. Subsequently, the texts are reorganized to obtain the final bounding boxes corresponding to the metadata.

bounding box. The data comprises two components: book covers derived from public digital libraries and high-resolution text-rich images from LAION-400M (Schuhmann et al., 2021).

Book covers. We manually download 11,000 book covers from the openly accessible Project Gutenberg digital library [3]. For each cover, we carefully extract pivotal metadata, such as the title and author. Subsequently, we utilize PaddleOCR [2] to analyze each image, from which we reorganize the extracted text to constitute the metadata to obtain the final bounding boxes. The entire process is shown in Figure 3. We utilize GPT-4 to design conversations revolving around the book's metadata, simultaneously requiring the model to explain its

[2] https://github.com/PaddlePaddle/PaddleOCR

[3] https://gutenberg.org

Table 1: Quantitative comparison with previous multimodal large language models on Visual Question Answering (VQA) and Key Information Extraction (KIE) datasets. TGDoc has two versions depending on the resolution of CLIP (Radford et al., 2021): 224×224 and 336×336. The best and the second results (accuracy %) are highlighted in **bold** and underlined, respectively.

Method	Resolution	LLM	VQA						KIE		
			STVQA	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	FUNSD	SROIE	POIE
BLIP-2 $OPT_{6.7b}$ (Li et al., 2023)	224	OPT-6.7B	13.36	10.58	21.18	0.82	8.82	7.44	0.00	0.00	0.02
BLIP-2 $FlanT5_{XXL}$ (Li et al., 2023)	224	FlanT5-11B	21.70	30.74	32.18	4.86	10.17	7.20	1.19	0.20	2.52
OpenFlamingo (Awadalla et al., 2023)	224	LLaMA-7B	19.32	27.82	29.08	5.05	14.99	9.12	0.85	0.12	2.12
LLaVA (Liu et al., 2024b)	224	Vicuna-7B	22.08	11.36	28.86	4.49	13.78	7.28	1.02	0.12	2.09
MiniGPT-4 (Zhu et al., 2023)	224	Vicuna-7B	14.02	11.52	18.72	2.97	13.32	4.32	1.19	0.04	1.31
mPLUG-Owl (Ye et al., 2023c)	224	LLaMA-7B	29.26	28.62	40.28	6.88	**16.46**	9.52	1.02	0.64	3.26
UniDoc (Feng et al., 2023b)	224	Vicuna-7B	30.78	34.50	40.72	6.47	13.75	10.48	1.19	1.40	3.92
TGDoc	224	Vicuna-7B	35.10	36.08	46.86	9.55	13.78	13.20	5.32	0.85	11.53
UniDoc (Feng et al., 2023b)	336	Vicuna-7B	35.16	36.80	46.24	7.70	14.74	10.88	1.02	**2.92**	5.10
LLaVAR (Zhang et al., 2023b)	336	Vicuna-7B	30.36	29.38	39.40	6.73	12.25	8.00	1.02	1.36	6.48
TGDoc	336	Vicuna-7B	**44.56**	**40.84**	**56.74**	**18.64**	14.92	**17.84**	**14.42**	2.21	**22.65**

Table 2: Analysis of our data with LLaVA (Liu et al., 2024b) and LLaVAR (Zhang et al., 2023b). "Que len" and "Ans len" represent the average number of tokens after the LLaMA (Touvron et al., 2023) tokenization.

Data	Instruction	Conv	Que len	Ans len
$LLaVA_{pretrain}$	CC3M	595K	15.9	15.4
$LLaVAR_{pretrain}$	PaddleOCR	422K	17.2	48.8
$Ours_{pretrain}$	PaddleOCR(box)	99K	42.5	80.7
$LLaVA_{finetune}$	GPT-4	158K	15.9	32.4
$LLaVAR_{finetune}$	GPT-4	16K	15.1	93.1
$Ours_{finetune}$	GPT-4(box)	12K	31.2	41.6

rationale behind the responses. For example, a query could be "Who is the author of this volume? Please justify your response with the bounding box $[x_{min}, y_{min}, x_{max}, y_{max}]$." The model's response would include the author's name along with the bounding box coordinates, like: "The work is authored by <text>$[x_{min}, y_{min}, x_{max}, y_{max}]$."

Text-rich scene images. We curate a corpus of 1,000 high-resolution (minimum 1024×1024 pixels), text-rich images from LAION-400M dataset (Schuhmann et al., 2021). Then we employ two OCR tools, PaddleOCR [2] and EasyOCR[4] to process each image. In addition, we utilize BLIP-2 (Li et al., 2023) to generate three descriptions for every image. Subsequently, as shown in Figure 2, these elements, two OCR results along with the generated captions, are prompted to instruct GPT-4 in generating multi-turn conversations that concentrate on the textual content of image. However, we observe some inconsistencies in the generated data, such as the unnecessary phrases like "based on the paddleocr". We manually review the conversation of each image to guarantee the quality, resulting in our final instruction fine-tuning dataset.

Finally, we provide an analysis of our data with that of LLaVA (Liu et al., 2024b) and LLaVAR (Zhang et al., 2023b) in Table 2. Refer to Appendix D for examples of our pre-training and fine-tuning datasets.

[4]https://github.com/JaidedAI/EasyOCR

4 Experiments

4.1 Implementation Details

The implementation of TGDoc was executed on the Linux platform with eight A100 GPUs. For vision encoder, we selected CLIP (Radford et al., 2021) and conducted experiments with image sizes of 224×224 and 336×336, respectively. It has been observed that text-related tasks generally benefit from higher resolution (Feng et al., 2023b; Ye et al., 2023a; Zhang et al., 2023b). We selected Vicuna (Chiang et al., 2023), which is optimized for multi-tasks and an evolution of LLaMA (Touvron et al., 2023) as our large language model. For pre-training, only the linear projection layer was trained. We applied the learning rate of 2e-3 and a batch size of 128. While for fine-tuning, we adjusted both the linear projection layer and the large language model, reducing the learning rate to 2e-5 with a batch size of 32. The AdamW optimizer (Loshchilov and Hutter, 2017) was utilized for parameter updates, with each stage undergoing one epoch of training. Moreover, We employed a cosine annealing scheduler (Loshchilov, Ilya and Hutter, Frank, 2016) and set the maximum sequence length limited to 2048.

4.2 Datasets and Evaluation Metrics

To validate the effectiveness of text-grounding in enhancing document understanding, we evaluate our method on six text-based Visual Question Answering (VQA) datasets, including STVQA (Biten et al., 2019), OCRVQA (Mishra et al., 2019),

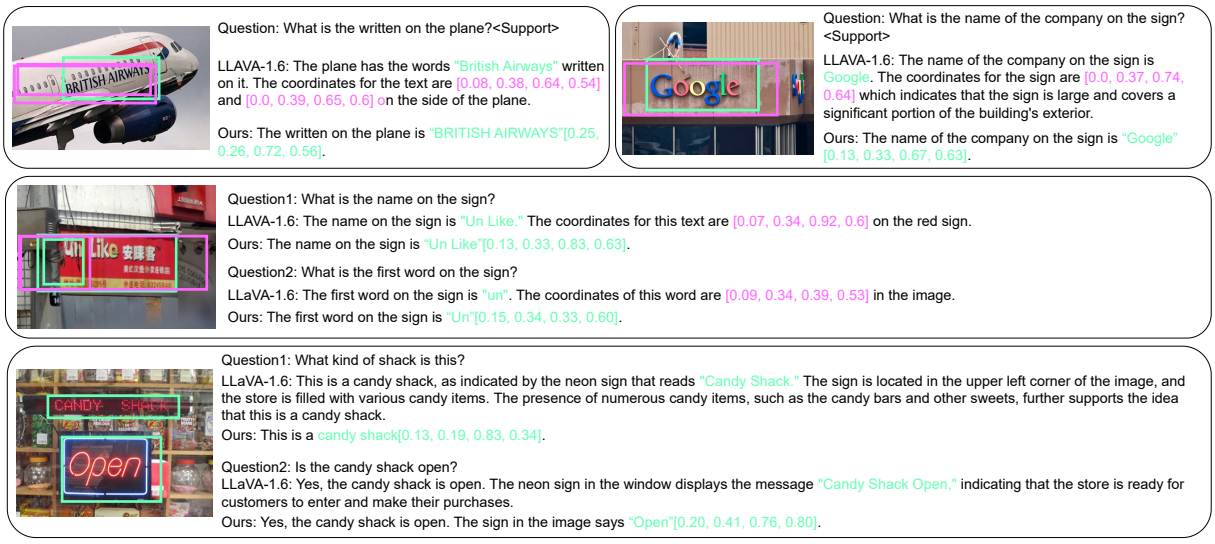

Figure 4: Visualization results of TGDoc. The bounding boxes generated by the model are visualized and integrated into the model's input. Results indicate that training with text-grounding data yields more accurate bounding boxes than those produced by the high-resolution model LLaVA-1.6 (Liu et al., 2024a). Additionally, results show that the model can accurately focus on regions pertinent to the answers, which enhances the model's interpretability.

TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022). In addition, we also employ three key information extraction (KIE) datasets, including FUNSD (Jaume et al., 2019), SROIE (Huang et al., 2019), and POIE (Kuang et al., 2023). For each question, we append the instruction "Support your reasoning with the coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$" at the end of the question to prompt the model to output the bounding box where the answer is located. Following the evaluation protocol proposed by Liu *et al.* (Liu et al., 2023b), we use accuracy metric to determine if the model-generated content correctly contains the ground truth answer.

4.3 Comparison with Other MLLMs

Quantitative analysis. We compare our method with several multimodal large language models on Visual Question Answering (VQA) and Key Information Extraction (KIE) benchmarks. Table 1 summarizes the quantitative comparison results. We conduct experiments using CLIP (Radford et al., 2021) at two resolutions: 224×224 and 336×336. We observe that increasing the CLIP's resolution leads to improved experimental results. This finding is consistent with previous research (Feng et al., 2023b; Ye et al., 2023a; Zhang et al., 2023b), as higher resolution contributes to a richer representation of the low-level details in textual content, aiding in image comprehension. Compared to other methods, our approach consistently achieves the best results on several benchmarks under two distinct resolution settings. More precisely, under the 336×336 setting, our method achieves state-of-the-art performance on all benchmarks with only 10% additional training data than LLaVAR (Zhang et al., 2023b). Furthermore, our approach markedly surpasses LLaVAR's performance on specific datasets, evidenced by improvements of 14.2% on STVQA, 11.46% on OCRVQA, 17.34% on TextVQA, and 11.91% on DocVQA datasets.

Although at an image resolution of 336×336, our approach fails to obtain SOTA results on the InfographicVQA dataset (Mathew et al., 2022). We analyze that this benchmark comprises various complex graphs that require fundamental reasoning and arithmetic skills, as answers cannot be directly extracted from the images. Consequently, our text-grounding method, which relies on directly extracting answers from images and text, proves insufficient. This deficiency is particularly noticeable in cases where the answer cannot be directly extracted from the image and necessitates further computations. Moreover, we quantitatively analyze that the InfographicVQA (Mathew et al., 2022) dataset possesses a high resolution and an extreme aspect ratio compared to other scene images. We average the dimensions across the dataset, resulting in $(w, h) = (1140, 2928)$. In the paper, we resize the image to accommodate the vision encoder of 224×224 or 336×336. However, such resizing

Table 3: Ablation studies on our collected data. The best and the second results (accuracy %) are highlighted in **bold** and underlined, respectively. $\mathcal{P}$ denotes the data gathered for the pre-training stage, while $\mathcal{F}$ denotes the data for the fine-tuning stage.

	Config		Benchmark			
	$\mathcal{P}$	$\mathcal{F}$	textVQA	DocVQA	FUNSD	POIE
Base			39.40	6.73	1.02	6.48
Base$_{+\text{pretrain}}$	✓		45.08	11.09	6.35	11.93
Base$_{+\text{finetune}}$		✓	51.23	13.93	13.94	15.52
Base$_{+\text{both}}$	✓	✓	53.11	15.56	14.12	18.74
Base$_{+\text{both+box}}$	✓	✓	**56.74**	**18.64**	**14.42**	**22.65**

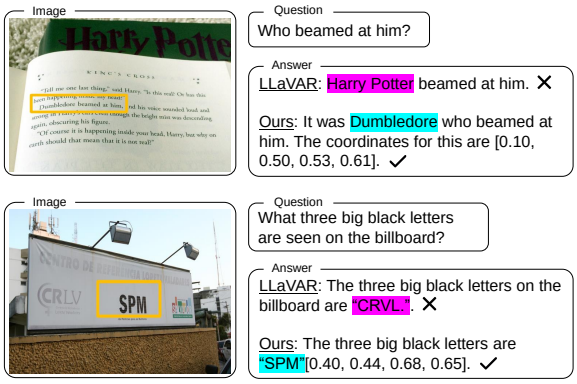

Figure 5: Examples of incorporating text-grounding capability. Compared to LLaVAR (Zhang et al., 2023b), our method delivers correct answers with enhanced accuracy.

approach may distort text in images, generating a discrepancy with the training data. As a result, the model struggles to detect the text in images and impedes text-grounding ability, which may increase the inference load on the model, and adversely affect its performance. Under these conditions, the utility of bounding boxes is restricted, and their requirement may potentially compound its reasoning workload. For future research, we aim to investigate a strategy that combines chain-of-thought reasoning with text-grounding to improve performance under these challenging conditions.

Qualitative analysis. In Figure 4, we present the qualitative results of our approach. We notice that methods such as LLaVA (Zhang et al., 2023a), which lack bounding box data training, are unable to generate reasonable bounding boxes even when the prompt is adjusted. As a result, we compare our approach with the more advanced, higher-resolution model LLaVA-1.6 (Liu et al., 2024a). Our analysis reveals that our method not only generates answers but also visually emphasizes the regions that support its responses. In comparison to LLaVA-1.6, the bounding box generated by our method is more precise and reasonable. The bounding box suggests that the model needs to identify the specific regions pertinent to the query, emulating the human cognitive process of locating pertinent information to answer questions. The results show that the model can ascertain the approximate positions of answers in connection with the questions, thus enabling accurate extraction of text from images for model's response. Our method effectively combines image content and textual data, improving its overall comprehension and precision in diverse situations.

4.4 Ablation Studies

In this section, we conduct ablation studies on our curated pre-training and fine-tuning datasets, encompassing four distinct experiments. The first experiment (labeled as "Base") excludes the data we collected, *i.e.*, the same setting as LLaVAR (Zhang et al., 2023b). The second (labeled as "Base$_{+\text{pretrain}}$") focuses on the impact of our pre-training data. The third (labeled as "Base$_{+\text{finetune}}$") investigates the influence of our fine-tuning data. To rigorously assess the impact of text-grounding on the model's comprehension ability, we undertake the fourth experiment (labeled as "Base$_{+\text{both}}$"), which removes all bounding boxes in data. All experiments are uniformly performed with an input image size of 336×336.

We present the results in Table 3. From the results, we can obtain following conclusions. First, alignment is crucial. The data collected during pre-training aids in the precise alignment of the vision and language models, leading to a substantial improvement in performance, particularly in text-rich scenarios. Second, although the volume of fine-tuning data is smaller than that of pre-training data, the performance based on fine-tuning data surpasses that based on pre-training. This suggests that the incorporation of specific grounded instruction tuning data during the fine-tuning stage enables the model to more accurately interpret user instructions. This leads to enhanced performance in various tasks. Additionally, this method facilitates the model's rapid acquisition of text-grounding skills, allowing it to focus on image regions directly pertinent to the answers. Third, text-grounding is effective. compared to the results without text-grounding, the introduction of text-grounding further improves the performance. As shown in Figure 5, compared to LLaVAR (Zhang et al., 2023b), our method provides more accurate answers.

Table 4: Performance of different model size.

LLM Model	TextVQA	OCRVQA	STVQA	InfoVQA	ChartQA	POIE
Vicuna-7B	56.74	40.84	44.56	14.92	17.84	22.65
Vicuna-13B	57.36	43.32	46.64	17.10	19.20	31.43

Table 5: Performance of whether freeze CLIP encoder.

Config	TextVQA	OCRVQA	STVQA	InfoVQA	ChartQA	POIE
freeze	57.36	43.32	46.64	17.10	19.20	31.43
unfreeze	55.84	42.04	44.62	16.53	18.4	29.03

4.5 Discussion

The experimental results significantly substantiate the efficacy of the text-grounding capability. Additionally, the experiment yields further insights, which will be discussed in the following.

Impact of the LLM's size. Since increasing the resolution of the visual module can improve performance, what would happen if we increase the size of the language model? We fix the input of vision encoder at 336×336 and increase the LLM's size to 13B. Table 4 shows the comparative results. We can infer that the overall performance of our method is contingent upon the capabilities of the LLM, *i.e.*, the larger the size, the stronger the performance of our method.

Impact of the vision encoder. CLIP (Radford et al., 2021) is trained on billions of image-text pairs and is not optimized for text, resulting in a domain gap that can affect the quality of the extracted visual features. To address this issue, we fine-tune all model parameters during the fine-tuning stage. The results are shown in Table 5. Interestingly, there is a slight decrease in performance, which is counterintuitive. We hypothesize that this decline is due to the fine-tuning data being much smaller than CLIP's original training data—less than one percent—which may disrupt the original structure of CLIP before it learns the textual features, leading to a loss in performance.

Impact of the input resolution. Typically, text-rich images are high-resolution, and compressing them to meet CLIP's (Radford et al., 2021) input sizes often results in a loss of textual detail. Increasing CLIP's input resolution can improve the results across all benchmarks. In this paper, besides Infographic VQA dataset, performance on the FUNSD and SROIE datasets remains relatively low even after increasing the resolution. In contrast, the POIE dataset exhibits a relatively high improvement. We find that the average dimensions of the FUNSD, SROIE, and POIE datasets are $(762, 1000)$, $(1303, 2331)$, and $(489, 566)$, respectively. Notably, POIE has a lower resolution. Our method achieves the best performance on POIE at a 224×224, highlighting the effectiveness of text-grounding in KIE tasks. When the resolution is increased to 336×336, the advantage of our method is further amplified.

Figure 6: Some examples of imprecise bounding boxes produced by our method.

Precision of bounding boxes. In Figure 6, we present examples of inaccurate bounding boxes generated by our methods. Compared to Figure 4, these boxes are either near the text or cover a fragment of it. We speculate that this issue may stem from the capabilities of the visual encoder CLIP (Radford et al., 2021), which is based on ViT (Dosovitskiy et al., 2020) and trained primarily on images of natural objects. This training emphasizes global features of images rather than optimizing for fine-grained local details like textual boundaries, leading to challenges in text detection. However, we argue that the bounding boxes, despite not always being exact, are vital for localization in VQA tasks. They can narrow down the model's retrieval scope, benefiting the overall understanding process.

5 Conclusion

Our research investigates text-grounding in document understanding, highlighting its significant role in improving model comprehension of text-rich scenes. We create 99,000 PowerPoint slides for pre-training, focusing on detection, recognition, and spotting tasks for image-text alignment. For fine-tuning, we compile a dataset of 12,000 high-quality conversations with bounding box annotations to specify text locations. Experimental results with the collected grounded instruction-following dataset reveal that text-grounding improves the model's interpretability and performance by identifying answer-related areas.

Limitations

Quantitative results of text-grounding. In this paper, we do not quantitatively assess the model's text-grounding ability as that in visual grounding. There are two main reasons. First, our study focuses on exploring how MLLMs improve in text-rich images through text-grounding. Second, similar to humans searching for answers in knowledge-based questions, the answer may be distributed across multiple areas of the image, complicating the development of an effective metric for text-grounding. Should each section be marked individually, or should a single, largest bounding box be used? If the latter is applied and the answer spans both the upper left and lower right corners, the bounding box would cover the entire image, making text-grounding meaningless.

Limitations in comprehending complex graphs. As shown in Table 1, text-grounding is more effective in text-rich scenarios than in complex graph contexts. This is because VQA for complex graphs often requires multi-step reasoning, which necessitates the model's ability to perform chain-of-thought (COT) processing. However, during data construction, we rarely developed such datasets due to the inherent complexity of creating COT datasets. Integrating text-grounding capabilities introduces additional challenges, such as determining whether to apply text-grounding at every step of COT or only at key steps to optimize model performance. This issue warrants further investigation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Proceedings of the Advances in Neural Information Processing Systems*, 35:23716–23736.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. ICDAR 2019 competition on scene text visual question answering. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1563–1570.

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

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *URL https://vicuna. lmsys. org*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. 2024. InternLM-XComposer2-4KHD: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*.

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

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Hao Feng, Qi Liu, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2023a. DocPedia: Unleashing the power of large multimodal model in the frequency

domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*.

Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. 2023b. UniDoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-PaperOwl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32th ACM International Conference on Multimedia*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1516–1520.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *Proceedings of the International Conference on Document Analysis and Recognition Workshops*, pages 1–6.

Hyung Il Koo. 2016. Text-line detection in camera-captured document images using the state estimation of connected components. *IEEE Transactions on Image Processing*, 25(11):5358–5368.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: practical dataset and end-to-end solution. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 36–53.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742. PMLR.

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

Xiangtai Li, Xia Li, Ansheng You, Li Zhang, Guangliang Cheng, Kuiyuan Yang, Yunhai Tong, and Zhouchen Lin. 2021. Towards efficient scene understanding via squeeze reasoning. *IEEE Transactions on Image Processing*, 30:7050–7063.

Zechao Li and Jinhui Tang. 2016. Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing*, 26(1):276–288.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26763–26773.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Loshchilov, Ilya and Hutter, Frank. 2016. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

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

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual question answering by reading text in images. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 947–952.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Proceedings of the Advances in Neural Information Processing Systems*, 35:27730–27744.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. KOSMOS-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR.

Ronald Rivest. 1992. The md5 message-digest algorithm. Technical report.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

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

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-Play VQA: Zero-shot VQA by conjoining large pre-trained models with zero training. *arXiv preprint arXiv:2210.08773*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lin Wang and Kuk-Jin Yoon. 2021. PSAT-GAN: Efficient adversarial attacks against holistic scene understanding. *IEEE Transactions on Image Processing*, 30:7541–7553.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-Instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. 2024. LongVILA: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. 2020. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023a. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023c. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. StrucTexTv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. 2023a. LLaVA-Grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

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

A Model Prompts

During the pre-training process, We undertake three tasks: detection, recognition, and spotting. Table 6 provides a selection of instructions for each task. We utilize GPT-4 to generate additional instructions, enabling the model to manage a diverse range of instruction formats. This approach enhances the model's ability to follow user commands and demonstrates robust generalization performance.

B Conversation Format of Our Model

System Message. "A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions."
Training template of model. We present one training template of the model as follows: *<System Message>USER: <image><Image Embedding></image>What is the book's title based on the image? Please provide the supporting text and its bounding box. ASSISTANT: <Model's Output>*

C Data Processing Workflow

To ensure data diversity, we first indexed various domains from the internet using query terms and collected 1.2 million PowerPoint (PPT) data samples. We then employed a hashing method to remove duplicates. Utilizing the multilingual capabilities of PaddleOCR, we detected and extracted text from each image. Images with a non-English character ratio exceeding 10% were removed. Additionally, we ensured that the largest bounding box in the PaddleOCR results occupied more than 5% of the entire image, accommodating the low-resolution CLIP's encoding efficiency for dense text. After data processing, 99,000 data samples were retained. The final data distribution is shown in Figure 7.

D Training Data Presentation

In Figures 8 and 9, we present the data we have collected. Figure 8 displays the PowerPoint data, while Figure 9 showcases the high-resolution, text-rich data gathered from CC3M.

E Additional Qualitative Visualization Results

Figures 10 and 10 provide additional visualization results for the STVQA and TextVQA datasets, illustrating the capabilities and effectiveness of text-grounding.

Instruction template 1	
task	Detection
examples	Please supply the coordinates <box> for the text <text> within the image. I would like to know the coordinates <box> for the text <text> in the image, please. Please identify the bounding box coordinates <box> for the text <text> within the image. Can you please give me the coordinates <box> for the text <text> in the image? Kindly provide the bounding box coordinates <box> for the text <text> within the image.

Instruction template 2	
task	Recognition
examples	Please recognize and supply the text enclosed within the given bounding box <box>. Can you identify and provide the text that falls within the specified coordinates <box>? I'd like you to detect and furnish the text contained within the provided bounding box <box>. Could you please extract and share the text within the defined coordinates <box>? Please locate and provide the text that is encompassed by the given bounding box <box>.

Instruction template 3	
task	Spotting
examples	Could you locate the text in the image and furnish the coordinates <box> for each text block? Please recognize all the text within the image and supply the coordinates <box> for each text element. Kindly identify and extract text from the image, and supply the coordinates <box> for each text portion. I'd appreciate it if you could identify and provide the coordinates <box> for all text found in the image. Kindly pinpoint the text in the image and provide the coordinates <box> for each text block.

Legend:

<box>	Values represent $[x_{min}, y_{min}, x_{max}, y_{max}]$
<text>	Values indicate the text in the bounding box

Table 6: Examples of instruction template for the pre-training stage.

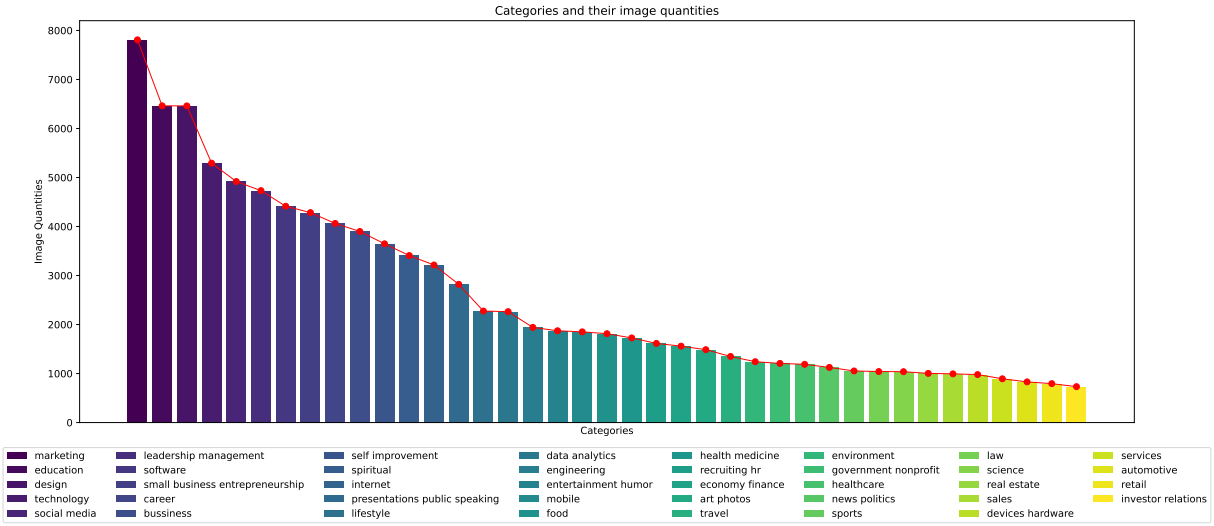

Figure 7: The distribution of data categories of the PowerPoint slides.

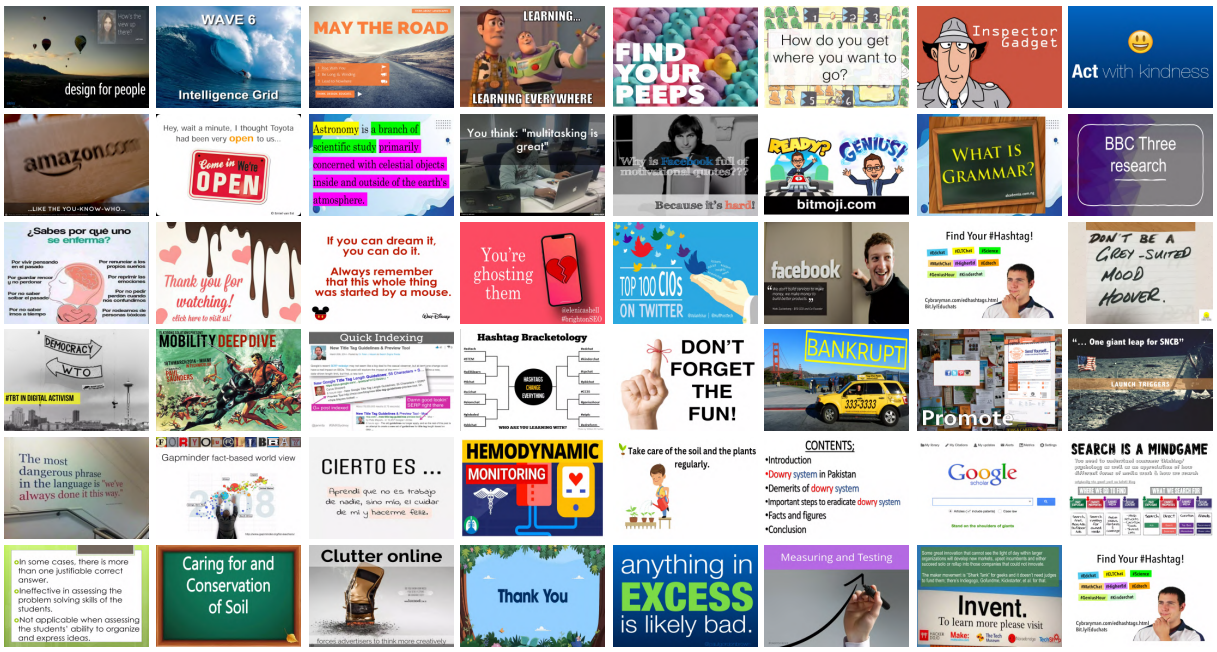

Figure 8: Examples of our collected PowerPoint data for pre-training.

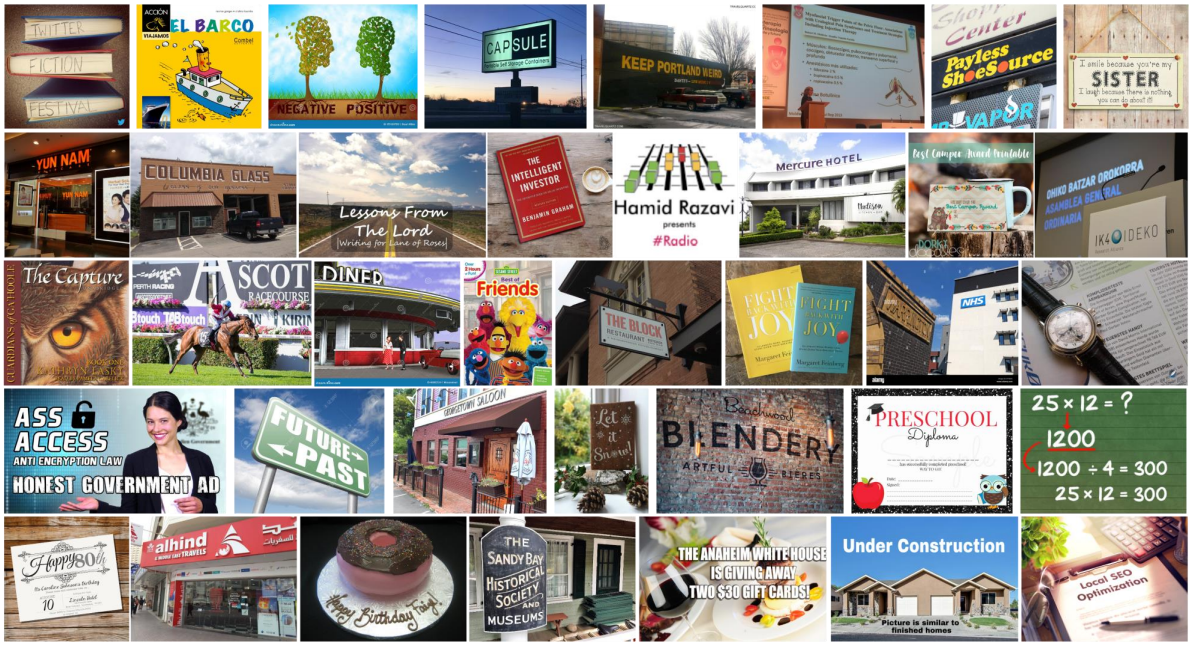

Figure 9: Examples of our collected book data and CC3M data for fine-tuning.

Figure 10: More visualization results of our method on STVQA (Biten et al., 2019) and TextVQA (Singh et al., 2019) dataset.

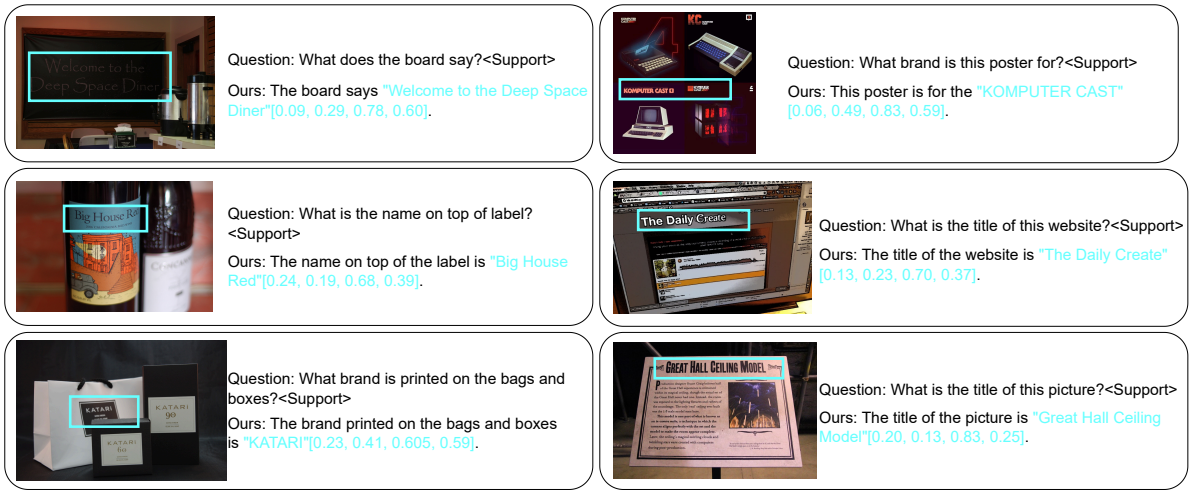

Figure 11: More visualization results of our method on STVQA (Biten et al., 2019) and TextVQA (Singh et al., 2019) dataset.