Improving Scene Text Recognition in Multimodal Large Language Models using Visual Text Grounding

Shashank Krishna Vempati
Yardi School of Artificial Intelligence
Indian Institute of Technology, Delhi
New Delhi, Delhi, India
aiy227509@iitd.ac.in

Chetan Arora
Yardi School of Artificial Intelligence
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi
New Delhi, Delhi, India
chetan@cse.iitd.ac.in

Abstract

Recent advances in multimodal large language models (MLLMs) have enabled strong performance on vision–language tasks, yet they remain limited in spatial scene text understanding due to inadequate spatial grounding of text. In this work we propose Visual Grounding for Scene Text (VGST) to instruction-tune MLLMs for improved fine-grained text localization and recognition in complex, cluttered scenes. Specifically we introduce three tasks/objectives for reverse localization of text as an instruction-tuning mechanism, where the model is guided to extract textual content based on spatial localization cues, thereby enhancing its spatial grounding ability. To further enhance spatial text awareness, we curate a reasoning-centric dataset containing over 27,000 question–answer pairs spanning diverse real-world scenarios. We evaluate our model (VGST) on three benchmarks covering sparse to dense text distributions: SVT, Occluded RoadText, and HierText, where it consistently outperforms strong MLLM baselines. Specifically, VGST achieves relative improvements of 8.28%, 8.18%, and 27.3% in Character Recognition Rate (CRR) for the text reverse localization task; 5.48%, 5.2%, and 5.13% in recall for text localization; and 8.7%, 3.21%, and 2.45% in F1 scores for end-to-end text recognition, respectively. Prompt sensitivity analysis shows that instruction tuning on a specific task using varied prompt formulations leads to robust performance on that task, even when the prompts at inference differ from those seen during training. These results establish VGST as a reliable and effective solution for spatially aware scene text understanding in unconstrained real-world images. Our code and dataset are available here: https://anonymous.4open.science/r/VGST

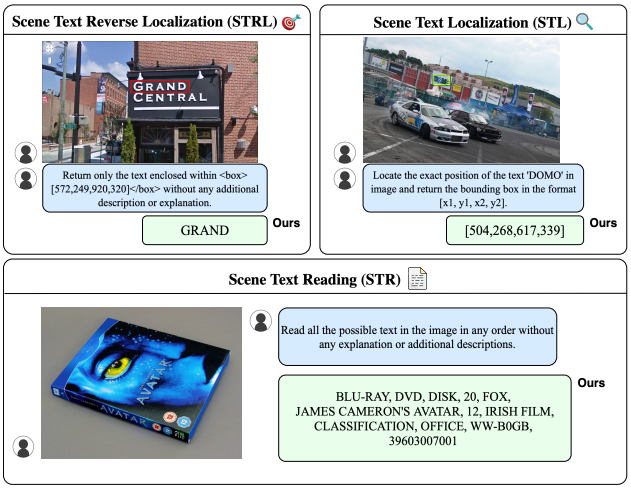

Figure 1. We demonstrate our model's capability to handle all three key scene text tasks: reverse localization (top-left, spatial understanding), precise localization (top-right), and comprehensive scene text recognition (bottom), using our proposed instruction-tuning framework based on reverse localization.

1. Introduction

Background and Motivation. MLLMs have emerged as powerful general-purpose vision–language systems capable of interpreting images via natural language prompts. Their versatility supports diverse applications, including Visual Question Answering (VQA), Document Analysis, and Key Information Extraction. Central to these tasks is the ability to extract and understand textual content from images, a capability that remains essential yet challenging in unconstrained, real-world settings.

Our Focus. In this work, we consider three complementary tasks for scene text understanding: Scene Text Reverse Localization (STRL), where the model identifies text con-

tent given a bounding box; Scene Text Localization (STL), which requires predicting the bounding box for a specified text phrase; and Scene Text Reading (STR), involving transcription of all visible text in an image. Together, these tasks evaluate fine-grained spatial reasoning and comprehensive text extraction in complex scenes.

Challenges. Despite recent advances, state-of-the-art MLLMs [3, 10, 32, 49, 54] continue to face challenges in scene text understanding. While these models perform well on structured document tasks and OCR-centric benchmarks such as OCRBench [35] and OCRBench v2 [17], they often struggle generalizing to unstructured and visually diverse settings in the real-world scenes. Unlike documents with consistent layouts, scene images contain text embedded in cluttered, variable backgrounds, with diverse fonts, orientations, occlusions, with little contextual cues. These factors make STR and STL substantially more complex, requiring precise spatial reasoning and robust recognition under unpredictable conditions.

Visual-Linguistic Spatial Alignment. An underexplored dimension is how well MLLMs leverage visual–linguistic spatial alignment in such settings. Despite strong language modeling capabilities, current systems exhibit limited spatial grounding, often producing coarse or inaccurate bounding boxes. Prompting MLLMs for bounding-box-level extraction frequently results in localization errors, highlighting a persistent gap from human-level spatial understanding. Although recent advances in object-level grounding [43, 55, 61, 63] have improved general visual reasoning, they remain inadequate for the fragmented and weakly correlated text regions common in scene images. Additionally, current MLLMs frequently fail on seemingly simple text extraction tasks such as reading from a given box, identifying text phrases, or recognizing multiple scattered text instances. These failures expose limitations in region-level text grounding and instance-aware recognition, underscoring the need for targeted improvements in spatially grounded scene text understanding.

Our Approach. Rather than framing weak spatial grounding as a general shortcoming, we approach it as a design opportunity: by instruction tuning on carefully constructed, location-aware tasks, we aim to guide the model toward more spatially grounded behavior. This motivates our reverse localization objectives, which are designed to strengthen the alignment between visual input and textual output. In doing so, our approach also implicitly probes whether the LLM component can effectively leverage visual features when explicitly supervised for spatial reasoning. Hence, we propose an instruction-tuning framework that enhances spatial grounding and recognition using the `Qwen2.5VL-7B` model as the running baseline. We design specialized question–answer pairs that define reverse localization tasks while keeping the vision encoder frozen. This setup enables us to improve visual–text alignment across multiple OCR-related tasks without retraining the vision backbone. Though, not explicitly experimented with, due to low compute availability, we believe our approach will generalize to other MLLMs as well.

Contributions. **(1)** We introduce three instruction-tuning objectives for reverse localization that explicitly infuse spatial text awareness into an MLLM model. **(2)** We present **ReaLoc27K-QA**, a dataset of over 27,000 question–answer pairs curated to advance spatial reasoning through reverse localization tasks. **(3)** We demonstrate consistent improvements in localization, reverse localization, and text recognition across complex scene images. **(4)** We show that instruction tuning alleviates weak visual-linguistic spatial alignment, particularly for spatial text extraction, by enabling more effective utilization of visual features.

2. Related Work

Scene Text Understanding. Scene text understanding has long been a fundamental problem in computer vision, with research split across three sub-tasks: scene text detection [30, 31, 57, 60], recognition [5, 24, 65], and end-to-end reading [21, 22]. These methods typically rely on supervised training over large annotated datasets and are often tailored for either sparse or dense text distributions. However, most approaches require task-specific architectures and struggle with generalization to real-world variability, such as occlusion, low contrast, and non-standard layouts. Recently, unified frameworks like Platypus [52] have attempted to bridge the gap, yet they still fall short in leveraging natural language interfaces for achieving robust spatial grounding.

MLLMs for OCR and Scene Text Tasks. Recent models like [3, 10, 15, 16] demonstrate emergent OCR capabilities but primarily focus on holistic vision–language tasks rather than spatially grounded text understanding. In contrast, OCR-focused models such as [8, 58, 59, 62] are tailored for document-based VQA and reading, often leveraging layout-aware architectures and pretraining. However, these efforts largely operate on structured documents and offer limited support for scene-level text. Recent evaluation studies [47] reveal that even strong MLLMs under-perform on localization and spatially precise reading tasks. Our work targets this gap and extends the scope of current MLLMs.

Spatial Grounding in Vision–Language Models. Spatial reasoning remains a key challenge for MLLMs. Prior work on grounding [34, 39, 43] has primarily focused on object-level understanding using either caption-based localization or referring expression grounding. While these models show promise for generic object detection, their performance does not translate well to scene text due to the fragmented, low-saliency, and often contextually disjoint

nature of textual regions in the wild. Attempts to incorporate OCR modules into MLLMs [26, 35] have demonstrated limited success, often relying on tightly coupled pipelines.

Instruction Tuning for Spatial Reasoning. Instruction tuning has emerged as an effective strategy to adapt MLLMs [46] to diverse downstream tasks including spatial reasoning [7, 27, 41, 53]. Works like [12, 42] have demonstrated improved generalization through prompt-based fine-tuning. However, few efforts have explored task-specific tuning aimed at improving spatial text understanding. Our work builds upon this gap by introducing reverse localization as an instruction-tuning strategy, where spatial cues are encoded as part of the prompt, thereby guiding the model to associate localized visual features with semantic content. This approach enables better alignment for text grounding, setting it apart from prior work that primarily emphasizes global or document-level semantics.

3. Proposed Approach

Existing MLLMs [3, 10, 49, 54, 59] often struggle with extracting text from scene text images due to the dominance of object-centric visual features [19]. Specifically, high-level image representations passed through the projection module tend to suppress fine-grained textual cues, leading the LLM decoder to under-attend to text regions in favor of more salient object features. Building on insights from prior works [6, 18, 56], we hypothesize that such misalignment results in the LLM component under-utilizing visual information in text understanding tasks. To mitigate this, we propose a reverse localization-based instruction-tuning framework aimed at enhancing visual-linguistic synergy, as detailed in the following subsections.

Architecture and Training. Our model, **VGST**, is built on Qwen2.5VL-7B [4] architecture, selected for its strong baseline performance on generic text reading tasks [17] and its computational efficiency. Figure 2 illustrates our overall framework, which comprises a visual encoder, a projection module, and a large language model (LLM) decoder. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we follow the dynamic resolution strategy proposed in [4], resizing it to the closest multiple of 28 along each dimension to preserve fine details. The visual encoder, based on a Vision Transformer (ViT) with window-based attention, processes I to produce a variable-length set of feature tokens $F = \{f_1, \ldots, f_n\}$. These features are then linearly projected into the LLM's embedding space via proposed projection module. In parallel, the input textual instruction prompt is tokenized into embeddings compatible with the decoder. Our instruction-tuning protocol follows the strategy described in [4], while enabling updates to both the projection module and the LLM during fine-tuning. This design ensures that spatial cues from the visual domain are preserved and meaningfully aligned with language representations, supporting our hypothesis.

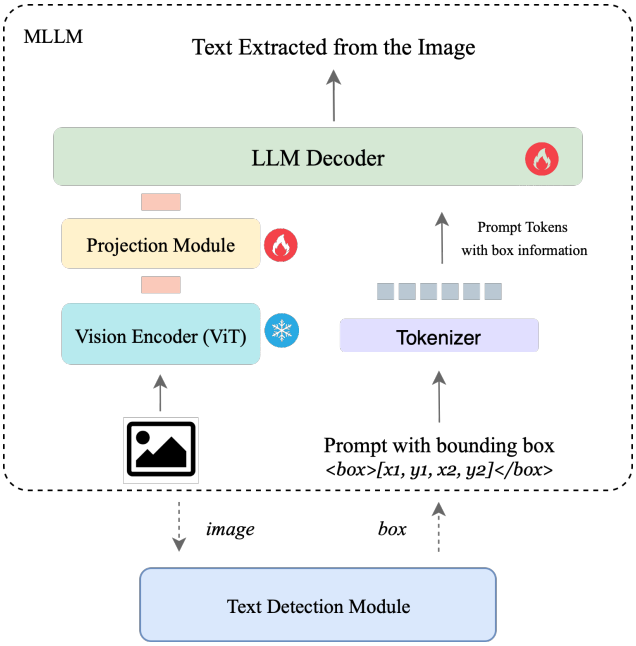

Figure 2. Architecture diagram inspired by [4], comprising a Vision Encoder, Projection Module, Prompt Tokenizer, and Qwen LLM Decoder. The Vision Encoder is frozen, while the Projection Module and LLM are unfrozen for training with reverse localization objectives. For bounding box-based interaction, an external text detection module is used to provide localized prompts, enabling more effective extraction of text.

Spatial Coordinate Encoding. Spatial coordinates provide a textual mechanism to specify regions of interest within an image. While MLLMs have historically struggled with interpreting such structured spatial inputs, recent work has shown that targeted fine-tuning can improve spatial reasoning and grounding [29, 38, 43, 45, 53]. Motivated by these insights, we hypothesize that explicitly training the model with coordinate-based prompts, where the instruction specifies a bounding box and requests the text within it, can enhance spatial understanding. To encode bounding boxes in the prompt, we experiment with three representations:

1. Absolute pixel coordinates (ABS),
2. Normalized to $[0, 1000]$ (NORM1000), and
3. Normalized to $[0, 1]$ (NORM01).

Each format expresses the coordinates of the bounding box as $[x_{min}, y_{min}, x_{max}, y_{max}]$, where (x_{min}, y_{min}) and (x_{max}, y_{max}) indicate the top-left and bottom-right corners, respectively. Following [4], we enclose these coordinates within special `<box>` and `</box>` tokens in the prompt (e.g., `<box>`$[x_{min}, y_{min}, x_{max}, y_{max}]$`</box>`) to clearly delimit the region specification. As shown in Table 1, prompts using absolute coordinates yield the best performance across three datasets with our base model, sug-

SCR	SVT	OccRT	HierText
ABS	**90.76**	**67.80**	**45.14**
NORM1000	34.31	29.56	17.79
NORM01	25.75	18.39	14.32

Table 1. **Spatial Coordinate Representation (SCR).** CRR comparison across different coordinate systems used for prompting the base Qwen2.5-VL-7B model with the prompt *"Read the text in the box: <box> without any additional description or explanation."*

gesting that preserving original scale information in textual instructions aligns well with image regions. The table also indicates that accuracy declines as the text density in the image increases.

Reverse Localization Objectives. We introduce a suite of instruction-tuning objectives aimed at enhancing the model's ability to spatially ground text within images, a task we refer to as Scene Text Reverse Localization (STRL). This builds on our coordinate-based prompting strategy, where the prompt includes a bounding box input and the model is expected to return the text contained within that region. This setup requires a tight association between visual regions and their textual content, effectively reversing the typical OCR task. The three instruction-tuning objectives are:

1. **Txt-Pred (Text Prediction):** Given an image and a prompt specifying a bounding box, the model is trained to predict the text contained within that region.
2. **Neg-Pred (Negative Text Prediction):** To improve robustness and mitigate overfitting, this objective introduces bounding boxes that do not contain any text. The model must learn to indicate the absence of text.
3. **Res-Loc (Reason-Based Reverse Localization):** This objective extends STRL by including an additional reasoning cue in the prompt, leveraging the model's language understanding to guide text extraction from the specified region.

The structured prompt templates used for *Txt-Pred* and *Neg-Pred* are detailed in Table 2. Having a diverse set of prompt formats helps reduce the model's sensitivity to prompt phrasing; during training, one template is randomly selected for each QA pair. VQA-style examples of these instruction formats are shown in Table 3.

Training Data. We prepare training data for these objectives by curating samples from several OCR and VQA benchmarks. Specifically, the COCO-Text[50], Uber-Text[64], and HierText[37] datasets are used to construct training examples for the *Txt-Pred* and *Neg-Pred* objectives. For the *Res-Loc* objective, we derive supervision from the TextVQA[48] dataset due to its rich question-text grounding structure. To further support the *Res-Loc* objective, we introduce a new dataset: ReaLoc27K-QA, consisting of approximately 27,000 reason-based localization question–answer pairs. This resource helps the model learn to attend to both visual and linguistic cues when extracting text from specified regions. The effectiveness of the *Res-Loc* objective and our dataset is quantitatively demonstrated in Table 8 in the Experiments section, where it shows significant contributions to model accuracy.

Prompt	Prompt Text
Prompt-1	<image>\n Extract the text content located at: <box> <<BOUNDARY_CONDITION>>
Prompt-2	<image>\n OCR this specific region: <box> <<BOUNDARY_CONDITION>>"
Prompt-3	<image>\n Output the text present in <box> <<BOUNDARY_CONDITION>>"
Prompt-4	<image>\n Identify the text present at this location <box> <<BOUNDARY_CONDITION>>"
Prompt-5	<image>\n Read the text in the box: <box> <<BOUNDARY_CONDITION>>"

Table 2. Different prompt templates used for text extraction in our Reverse Localization Objectives. Each prompt includes the <image> tag followed by a distinct instruction. The tag <<BOUNDARY_CONDITION>> refers to the common instruction appended to all prompts: "without any additional description or explanation. Output should not include text from other parts of the image apart from the bounding box provided."

Objective	Prompt	Target
Text-Pred	What's the text in <box>?	Actual text
Neg-Pred	What's the text in <box>?	No text
Res-Loc	Contextual Clue + <box>?	Actual Text

Table 3. An overview of the three instruction-tuning Reverse Localization objectives proposed in this paper.

Loss Formulation. We adopt a standard next-token prediction framework with cross-entropy loss to train our model, focusing on textual outputs. Given an input image I, its associated prompt P, and the target answer sequence $T = \{t_1, t_2, \ldots, t_N\}$, the objective is to maximize the conditional likelihood of each token t_i given the image, prompt, and all previously generated tokens. Formally, the training loss is defined as:

$$\mathcal{L} = -\sum_{n=1}^{N} \log p_\phi(t_n \mid I, P, t_{<n}), \tag{1}$$

where ϕ represents the model parameters. The term $p_\phi(t_n \mid I, P, t_{<n})$ denotes the probability of generating the n-th token t_n given the image I, the prompt P, and all previously generated tokens $t_{<n}$.

Figure 3. Samples from our proposed **ReaLoc27K-QA** dataset. The dataset is designed to provide models with both bounding-box information and accompanying reasoning to support text localization through language understanding. For visualization, the bounding boxes from the prompts are overlaid on the images. Please zoom in better visualization.

4. Contributed Dataset: ReaLoc27K-QA

We contribute a new dataset, *ReaLoc27K-QA*, designed to explicitly exploit spatial reasoning within VQA. *ReaLoc27K-QA* consists of $27,049$ image–question–answer (IQA) pairs, each with bounding boxes and detailed reasoning annotations. The dataset encourages models to condition on both the provided reason and the localized region to answer questions more accurately. To further advance spatial reasoning, we designed this dataset to leverage the capability of MLLMs to interpret richly structured textual prompts describing spatial relationships, moving beyond simple region annotations toward more nuanced, language-grounded spatial understanding.

Dataset Construction. We build our dataset based on the TextVQA [48] dataset, which offers image–question–answer (IQA) pairs. To obtain the bounding boxes and corresponding text regions, we first run PaddleOCR [13] on all images to extract OCR-detected text along with their bounding box coordinates. We then map the answer in each IQA pair to the appropriate OCR-extracted bounding box by aligning the answer text with the recognized OCR output, thereby identifying the relevant localized region. To generate high-quality reasoning annotations, we use Gemini-2.5 Pro[1] with few-shot prompting. The model is given the OCR text, bounding box coordinates, question, and answer, and asked to produce a new question that explicitly incorporates both the reason and the bounding box context, transforming it into a reason-based localization task. All generated annotations were subsequently subjected to thorough manual and automated verification to ensure correctness, consistency, and alignment with the intended objective. Figure 3 shows some representative examples from the dataset.

5. Experiments and Results

Implementation Details. We implemented **VGST** on a Linux system using PyTorch, initializing our model from the Qwen2.5-VL-7B-Instruct weights. Instruction tuning was performed in two stages: first for 2 epochs using only the Rev-loc objective, and then for an additional 2 epochs combining all three objectives. Training was conducted on 8 NVIDIA A100 GPUs (40GB each) using `bfloat16` precision with a learning rate of 2×10^{-7}, employing a cosine scheduler and a warmup ratio of 0.03. We used a per-device batch size of 2 and gradient accumulation steps of 4. This setup yielded the best results, with any deviation leading to a consistent but moderate drop in performance, highlight-

[1] https://aistudio.google.com/

ing sensitivity to hyperparameters. The model was trained with a maximum context length of $32,000$ tokens, enabling it to process long-text prompts effectively. We tuned only the projection module and the LLM component while keeping the vision encoder frozen, following our initial hypothesis. Image input resolutions ranged widely, from 784 up to 7.84 million pixels, supporting diverse real-world use cases. For all our experiments, we set the temperature of VGST to 10^{-6} and tested on single A100 GPU.

Datasets Used. We use six publicly available scene-text datasets: COCO-Text [50], HierText [37], Uber-Text [64], Street View Text (SVT) [51] Occluded RoadText (OccRT) [1] and IC15 [25]. For training, we use $11,786$ images from COCO-Text, $8,281$ images from HierText, and $16,927$ images from Uber-Text. These datasets support the *Txt-Pred* and *Neg-Pred* objectives in our framework. To generate IQA pairs, we randomly sample a set of text instances (with their bounding boxes) from each image and associate them with prompts drawn at random from the set shown in Table 2. This procedure yields 285,449 QA pairs. Combined with the ReaLoc27K-QA dataset, our total training corpus comprises 312,498 QA pairs. For evaluation, we use all 249 images from the SVT dataset, 202 validation images from the OccRT dataset as test, 817 images randomly sampled from the HierText test split and complete 500 test set images from IC15 dataset. These four datasets are selected to enable comprehensive evaluation across a broad spectrum of scene-text scenarios. SVT is characterized by sparse, low-density text. OccRT contains medium-density text with frequent occlusions and small-scale text regions. HierText represents high-density text scenes with numerous small instances. This diversity ensures that our evaluation rigorously tests the model's ability to localize and recognize text in varied and challenging real-world settings. The consistent accuracy improvements observed across these datasets underscore the model's enhanced spatial understanding and robust performance across multiple tasks.

Evaluation Metrics. STRL performance is measured using the Word Accuracy Ignoring Case and Symbols (WAICS) [24]-based Character Recognition Rate (CRR), which normalizes for case and symbol variations. STL is evaluated using Intersection over Union (IoU) based recall to assess localization accuracy. STR performance is quantified with precision, recall, and F1-score using unique word-level matching. Additionally, we analyze the model's tendency to hallucinate text by reading instances that are not present in the scene, providing further insight into its reliability and robustness across diverse visual contexts.

5.1. Quantitative Evaluation

We compare our proposed **VGST** with several existing MLLMs, as well as with the base model on which our approach is built. Although our model is trained exclusively on the STRL task, we observe consistent improvements not only in STRL performance but also in the model's ability on related tasks such as STL and STR.

Model	ICL	SVT	OccRT	HierText
GOT[54]	✗	87.29	38.16	52.01
InternVL-2.5-7B[11]	✗	85.72	59.88	38.74
Qwen2.5-VL-7B[4]	✗	90.93	67.80	44.61
VGST (ours)	✗	**97.51**	**76.12**	**67.32**
Qwen2.5-VL-7B	✓	92.38	69.34	67.29
VGST (ours)	✓	**96.97**	**78.95**	**82.72**

Table 4. **Scene Text Reverse Localization.** CRR metric comparison across benchmark datasets using model-specific prompts. The results highlight performance with and without In-Context Learning (ICL), where the model is provided with two illustrative examples before being prompted with the target question.

Model	SVT	OccRT	HierText
GOT[54]	42.84	28.75	26.19
InternVL-2.5-7B[11]	61.77	39.53	42.12
Qwen2.5-VL[4]	78.86	44.17	48.69
VGST (ours)	**84.34**	**49.37**	**54.72**

Table 5. **Scene Text Localization.** Average recall across images on benchmark datasets using model-specific prompts.

Scene Text Reverse Localization. Our proposed training objectives are specifically designed with the STRL task in mind. This task also serves as a controlled setup for investigating our hypothesis that the LLM component in MLLMs can under-utilize encoded visual features, a limitation discussed in the broader Visual Language Models (VLM) literature. To examine this hypothesis in a concrete setting, we focus on applying joint tuning to both the LLM and the visual projection module responsible for selecting relevant features from visual component. This targeted approach yields substantial performance improvements in the STRL task across three standard benchmark datasets, as summarized in Table 4. These results provide empirical support for our hypothesis within the scope of text-centric visual tasks and lay the groundwork for future investigations across other architectures.

Scene Text Localization. In STL task, we prompt the model to predict the bounding box location of a specific text instance within an image. To reduce potential sources of ambiguity, we select, for each image, a set of instances that are unique (i.e., not repeating) and have less than 90% textual overlap with other instances. Although our fine-tuning procedure is focused on the STRL task, we observe a significant improvement in localization ability with our proposed training objectives. Specifically, VGST demonstrates a substantial performance gain over its base model. Table 5 reports the quantitative results, showing consistent improvements across datasets with varying text densities. For all

Model	Size	SVT			OccRT			HierText			IC15		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Closed-source Models													
Gemini-1.5[15]	-	52.79	78.52	63.13	59.45	57.23	58.32	-	-	-	71.26	72.75	71.99
ChatGPT-4V[47]	-	54.88	84.70	66.60	61.59	44.57	51.72	68.67	43.01	52.89	72.57	68.54	70.50
Open-source Models													
GOT[54]	580M	55.63	66.43	60.55	36.34	19.45	25.34	55.24	34.52	42.49	60.95	44.02	51.12
InternVL-2[9]	7B	20.09	81.71	32.25	33.98	48.11	39.83	47.09	51.32	49.11	35.58	75.63	48.39
InternVL-2.5[11]	7B	50.92	**86.55**	63.87	46.13	51.54	48.68	65.82	59.37	62.43	66.07	**78.97**	71.94
Qwen2.5-VL[4]	7B	45.66	88.37	60.21	42.83	**63.88**	51.05	63.28	**67.79**	65.46	57.62	78.49	66.45
VGST (ours)	7B	**57.58**	86.00	**68.97**	**65.95**	60.54	**63.13**	**79.43**	61.71	**69.46**	**77.97**	72.69	**75.24**

Table 6. **Scene Text Reading.** Comparison of different vision-language models on text detection benchmarks. P, R and F1 scores are reported for four datasets. All computations are performed after post-processing to remove extraneous text generated by models. **Prompt:** *"Read all visible text from the image without any additional explanation or description."*

Method	Size	Edit Dist.↓	F1-score↑	BLEU↑	METEOR↑
UReader [58]	7B	0.568	0.661	0.258	0.488
LLaVA-NeXT [33]	34B	0.499	0.558	0.379	0.678
TextMonkey [36]	7B	0.331	0.743	0.521	0.728
DocOwl1.5 [20]	7B	0.334	0.788	0.525	0.708
InternVL-ChatV1.5 [10]	26B	0.267	0.834	0.587	0.744
Qwen-VL-Max [3]	>72B	0.182	0.881	0.586	0.848
GOT [54]	580M	0.112	0.926	0.676	0.896
VGST (Ours)	7B	**0.075**	**0.942**	**0.819**	**0.931**

Table 7. Comparison of various models on English OCR task. Results of all the models are taken from [54]

evaluations, we use a standard IoU threshold of 0.5 to compute metrics.

Scene Text Reading. STR is a highly anticipated capability of MLLMs due to its importance in a variety of downstream tasks, including visual question answering and scene understanding. Prior works [8, 58] have explored overall text reading but has typically focused on document images, where textual content is better formatted and exhibits higher inter-instance correlation. In contrast, scene text presents unique challenges such as occlusion, object dominance, and low to no correlation between different text instances within the same image. While our VGST model did not achieve an improvement in overall recall for STR, we observed a notable reduction in hallucination compared to other MLLMs, suggesting that our training objectives may improve prediction reliability even when overall coverage remains similar. Table 6 shows the quantitative results for this experiment.

English Scene Text OCR. GOT [54] introduced a benchmark of 200 English scene text images and evaluated several foundation models on OCR tasks. We use the same benchmark and observe that our model outperforms all prior methods across all metrics. Table 7 shows the result.

Model	SVT	OccRT	HierText
Base Model	89.45	67.80	44.61
+ Instruction Tuning (Without Res-Loc)	93.52	62.77	48.75
+ Instruction Tuning (With Res-Loc)	**97.73**	**75.98**	**71.90**

Table 8. Ablation study showing performance of **VGST** on STRL task. Instruction tuning with the *Text-Pred* and *Neg-Pred* objectives improves performance over the base. Adding the *Res-Loc* objective leads to the best performance across all datasets, highlighting its critical role in enabling effective spatial grounding.

5.2. Qualitative Results

Figure 4 shows qualitative results for the three (STRL, STL and STR) tasks, highlighting VGST's spatial understanding and scene text reading ability.

6. Ablation Study

Recent works [2, 23, 40] have shown that careful hyperparameter selection during instruction-tuning can significantly influence model performance and may even compromise existing capabilities. While we report the best-performing settings found within our computational constraints, more extensive hyperparameter searches could potentially yield higher accuracy, though at a considerable computational cost, as noted in prior studies. We have conducted an ablation study to evaluate the effectiveness of introducing our new dataset. Specifically, we first instruction-tuned the model on only the first two tasks, which are inherently complementary and cannot be meaningfully separated. In addition, we experimented with an initial fine-tuning stage using our ReaLoc27K dataset, followed by further tuning on all three tasks. This approach yielded improved results, suggesting that our dataset effectively introduces reasoning cues into the prompts and enhances the model's spatial text understanding. Table 8 presents this ablation quantitatively.

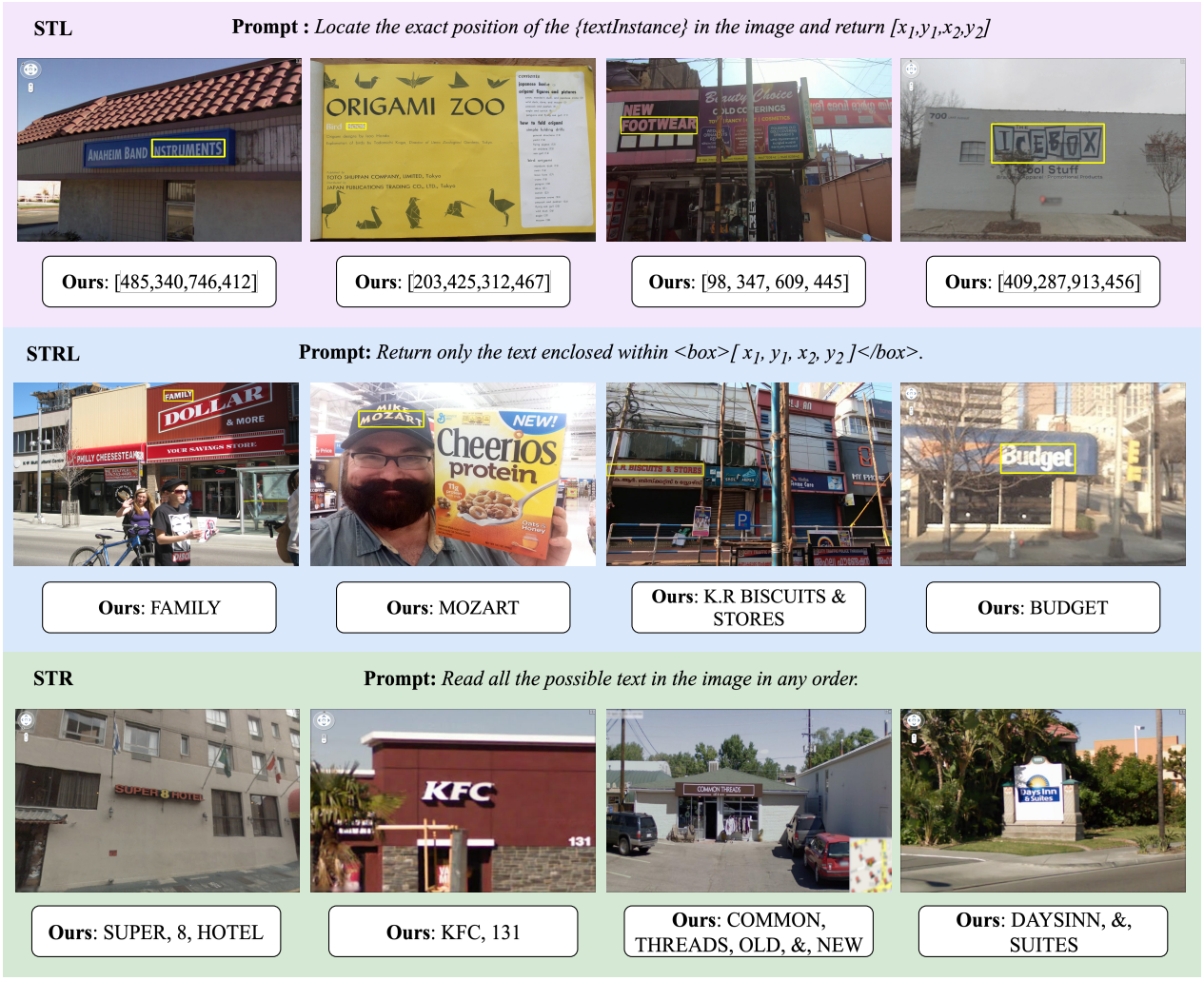

Figure 4. Visualization results of the proposed **VGST** model across all three addressed tasks. For the STL task, predicted bounding boxes generated by the model are overlaid on the input images to illustrate localization performance. For the STRL task, the ground truth bounding boxes are shown to facilitate visual comparison and interpretation. Please zoon in for better visualization.

7. Failure Cases and Future Directions

In analyzing the model outputs, we observe that certain failure modes can be attributed to the nature of training data and model limitations. First, the distribution shift between generic scene image datasets used during pretraining and the more structured, text-centric nature of STRL tasks likely contributes to the under-utilization of fine-grained visual cues. Second, inconsistencies in localization performance where small text instances are sometimes correctly identified but others near visually complex regions are missed highlights potential gaps in the model's spatial reasoning and attention capabilities. These findings suggest that current MLLMs may struggle to effectively prioritize and reason over relevant visual features in scenarios requiring fine spatial discrimination.

We observed that the model exhibits sensitivity to prompt variations, consistent with findings in prior studies [14, 28, 44]. However, even simple fine-tuning improved performance on related but differently phrased prompts at evaluation. This suggests that mitigating prompt sensitivity remains an important area for further investigation.

8. Conclusion

In this work, we demonstrated that instruction tuning with reverse localization objectives can meaningfully enhance spatial text understanding in MLLMs. We introduced VGST, a model tailored for scene text tasks, and showed that structured reverse localization prompts help unlock latent capabilities in localization and reading. Additionally, we contributed ReaLoc27K-QA, a dataset explicitly de-

signed to inject spatial reasoning into the training process. Our results show significant improvements across reading, localization, and reverse localization benchmarks. These findings indicate that current MLLMs often overlook critical spatial cues, particularly in cluttered and text-dense environments. This highlights reverse localization as not just a promising direction, but a necessary component in the future training of MLLMs. It should be treated as a core mechanism for enabling fine-grained visual-textual understanding, rather than as a peripheral objective.

References

[1] ICDAR 2024 Occluded RoadText Competition, 2024. 6

[2] Jacob Adkins, Michael Bowling, and Adam White. A method for evaluating hyperparameter sensitivity in reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2, 3, 7

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 3, 6, 7

[5] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer, 2022. 2

[6] DI Campbell, S Rane, T Giallanza, N De Sabbata, K Ghods, A Joshi, A Ku, SM Frankland, TL Griffiths, JD Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. neurips, 2024. 3

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 3

[8] Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, et al. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*, 2025. 2, 7

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 7

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 2, 3, 7

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Ex-

panding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 6, 7

[12] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 3

[13] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. 5

[14] Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024. 8

[15] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 2, 7

[16] OpenAI et al. Gpt-4 technical report, 2024. 2

[17] Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2025. 2, 3

[18] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. 3

[19] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13861–13871, 2024. 3

[20] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 7

[21] Mingxin Huang, Jiaxin Zhang, Dezhi Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, and Lianwen Jin. Estextspotter: Towards better scene text spotting with explicit synergy in transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19495–19505, 2023. 2

[22] Mingxin Huang, Hongliang Li, Yuliang Liu, Xiang Bai, and Lianwen Jin. Bridging the gap between end-to-end and two-step text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15608–15618, 2024. 2

[23] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language models fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. 7

[24] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20543–20554, 2023. 2, 6

[25] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 6

[26] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2, 2021. 3

[27] Dohwan Ko, Sihyeon Kim, Yumin Suh, Minseo Yoon, Manmohan Chandraker, Hyunwoo J Kim, et al. St-vlm: Kinematic instruction tuning for spatio-temporal reasoning in vision-language models. *arXiv preprint arXiv:2503.19355*, 2025. 3

[28] Ao Li, Zongfang Liu, Xinhua Li, Jinghui Zhang, Pengwei Wang, and Hu Wang. Modeling variants of prompts for vision-language models. *arXiv preprint arXiv:2503.08229*, 2025. 8

[29] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 3

[30] Min Liang, Jia-Wei Ma, Xiaobin Zhu, Jingyan Qin, and Xu-Cheng Yin. Layoutformer: Hierarchical text detection towards scene text understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15665–15674, 2024. 2

[31] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):919–931, 2023. 2

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 7

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 2

[35] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 2, 3

[36] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 7

[37] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1059, 2022. 4, 6

[38] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640, 2024. 3

[39] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 2

[40] Krishna Kanth Nakka, Ahmed Frikha, Ricardo Mendis, Xue Jiang, and Xuebing Zhou. Federated hyperparameter optimization through reward-based strategies: Challenges and insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4236–4244, 2024. 7

[41] Michael Ogezi and Freda Shi. Spare: Enhancing spatial reasoning in vision-language models with synthetic data. *arXiv preprint arXiv:2504.20648*, 2025. 3

[42] Tanawan Premsri and Parisa Kordjamshidi. Tuning language models with spatial logic for complex reasoning. In *The 4th International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*. 3

[43] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024. 2, 3

[44] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pages 303–313. Springer, 2025. 8

[45] Karun Sharma and Vidushee Vats. Think to ground: Improving spatial reasoning in llms for better visual grounding. In *Workshop on Reasoning and Planning for Large Language Models*. 3

[46] Zhang Shengyu, Dong Linfeng, Li Xiaoya, Zhang Sen, Sun Xiaofei, Wang Shuhe, Li Jiwei, Runyi Hu, Zhang Tianwei, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. 3

[47] Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*, 2023. 2, 7

[48] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 4, 5

[49] 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. 2, 3

[50] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 4, 6

[51] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464, 2011. 6

[52] Peng Wang, Zhaohai Li, Jun Tang, Humen Zhong, Fei Huang, Zhibo Yang, and Cong Yao. Platypus: A generalized specialist model for reading text in various forms. In *European Conference on Computer Vision*, pages 165–183. Springer, 2024. 2

[53] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023. 3

[54] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. General ocr theory: Towards ocr-2.0 via a unified end-to-end model, 2024. 2, 3, 6, 7

[55] Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pages 164–182. Springer, 2024. 2

[56] Yuxuan Xie, Tianhua Li, Wenqi Shao, and Kaipeng Zhang. Tp-eval: Tap multimodal llms' potential in evaluation by customizing prompts. *arXiv preprint arXiv:2410.18071*, 2024. 3

[57] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. Textfusenet: Scene text detection with richer fused features. In *IJCAI*, pages 516–522, 2020. 2

[58] 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, 7

[59] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 2, 3

[60] Maoyuan Ye, Jing Zhang, Juhua Liu, Chenyu Liu, Baocai Yin, Cong Liu, Bo Du, and Dacheng Tao. Hi-sam: Marrying segment anything model for hierarchical text segmentation.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. 2

[61] Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. Rod-mllm: Towards more reliable object detection in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14358–14368, 2025. 2

[62] Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*, 2024. 2

[63] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843, 2025. 2

[64] Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang, Keith Seifert, and Ben Kadlec. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop - CVPR 2017*, Hawaii, U.S.A., 2017. 4, 6

[65] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Can Huang, Hao Liu, Xin Tan, Zhizhong Zhang, and Yuan Xie. Multi-modal in-context learning makes an ego-evolving scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15567–15576, 2024. 2