

READ THE OBJECT: A Dataset and Task for Text-Grounded Object Understanding in Real-World Images

Arpita Chowdhury^{1,†} , Hyojin Park², Hoang Le² , Wei-Lun Chao³ ,
Munawar Hayat² , Fatih Porikli² , and Shweta Mahajan^{4,5,†} 

¹ The Ohio State University, USA

² Qualcomm AI Research, USA *

³ Boston University, USA

⁴ York University, Canada

⁵ Vector Institute for AI, Canada

chowdhury.150@osu.edu, shwemaha@yorku.ca

Abstract. Objects in the real world are often distinguished not only by their visual appearance, but also by the text printed on them. Locating players by their jersey numbers in sports, navigating highways with traffic signs, or identifying medicine vials based on the dosage, all require linking textual cues with the correct object instance for robust perception. Recent progress in vision and language considers scene-text recognition and object recognition as two independent tasks and overlooks a critical capability of object-text binding. To address this gap, we introduce READ THE OBJECT, a carefully curated benchmark for evaluating instance-level associations between text and objects in natural, real-world images. The proposed benchmark explicitly and rigorously annotates each object with its corresponding text; including visual attributes and spatial cues to disambiguate multi-instance scenes. Within this unified framework, we define six complementary tasks for evaluating text-grounded perception and reasoning. Experiments on state-of-the-art models reveal a clear binding gap: despite strong performance on isolated text recognition and global reasoning, they struggle to associate text with the correct object instance in complex scenes. READ THE OBJECT establishes a new standard for bridging reading, perception, and reasoning in vision-language research.

Keywords: Text-Object Association · Referring Expression Comprehension · Benchmark Dataset · Scene Text Understanding

1 Introduction

Multimodal understanding, where agents interpret and reason over sensory inputs such as vision and language, is central to building systems that interact

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

[†]Core contribution.



Fig. 1: Scene-text based referring expression task. (left:) Our findings reveal that state-of-the-art visual understanding and object localization models fail to localize visual objects based on associated scene-text due to a lack of joint scene-text and object localization and understanding benchmarks. (Right:) We introduce READ THE OBJECT with comprehensive annotations of objects and associated scene-text.

effectively with the real world. Recent advances in vision-and-language research have led to remarkable progress in tasks including image captioning [36, 39], referring expression [7], scene-text recognition [10, 38], and visual question answering [37]. In practical settings, conversational agents are frequently asked to reason about diverse visual information. Some queries require identifying objects (e.g., “Which animal is this?”), while others involve reading scene text (e.g., “What is the dosage strength indicated on the vial?”). The former relies on object detection and classification, whereas the latter depends on optical character recognition (OCR) to extract textual content from images.

Many real-world scenarios require associating scene text with the specific object instance on which it appears. For example, a navigation system must identify the traffic sign labeled “Exit 18” at a busy intersection rather than nearby vehicles that may also contain numbers, and a robot may be asked to grasp the coffee mug labeled “11 oz” among visually identical cups. Similar challenges arise in retail settings where products are distinguished by labels or expiration dates, and in sports analytics where jersey numbers identify players. In such environments, detecting text or localizing objects alone is insufficient; systems must correctly associate each text instance with its corresponding object. This instance-level association between scene text and objects is essential for reliable perception and robust visual reasoning in real-world applications, including robotic manipulation, autonomous navigation, and intelligent assistants.



Fig. 2: Illustration of the proposed READ THE OBJECT benchmark. Evaluation of text-grounded object understanding through six task types—(1) Referring Expression Comprehension, (2) Text-Object Grounding, (3) OCR, (4) Image Captioning, (5) Text-Region Grounding, and (6) Visual Question Answering.

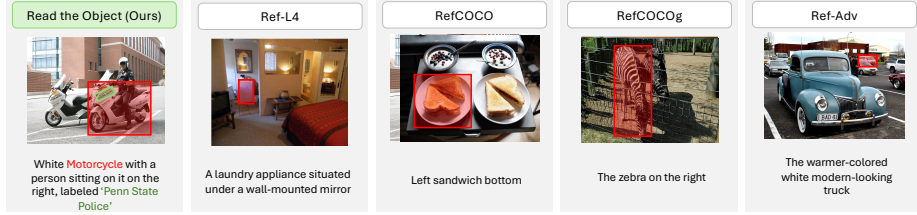


Fig. 3: Comparison with existing referring expression benchmarks. Examples from RefCOCO and its variants (Ref-L4, RefCOCOg, Ref-Adv) and READ THE OBJECT. While prior datasets rely on visual attributes and spatial relations to ground objects, Ours additionally leverages scene text on objects as a grounding signal, thereby enabling evaluation of text-object associations in complex scenes.

The success of object localization [7, 11] and OCR detection [30, 43] is driven by the availability of a wide range of annotated datasets. For example, a stand-alone OCR module and certain multimodal models [25, 26, 41, 43] can output text present in an image, whereas object detection models [15, 30] can accurately localize objects. Classic referring expression datasets such as RefCOCO [46], RefCOCO+ [8], RefCOCOg [18], and Ref-Adv [11] primarily rely on visual attributes and spatial cues to identify objects, and rarely incorporate scene text as a distinguishing attribute (Fig. 3). Conversely, OCR benchmarks focus on detecting and recognizing text regions but do not link those text instances to the objects that contain them. This separation leaves a fundamental gap in evaluating instance-level text-object association, a capability that will be essential for assistive robots and similar real-world systems, such as in autonomous driving.

To bridge this gap, we introduce **READ THE OBJECT**, a novel task and benchmark for evaluating *joint scene-text and object understanding*—the ability to associate textual content with the correct object instance in an image. Our

dataset explicitly links object instances with their inscribed text and augments them with fine-grained attributes and positional cues to resolve ambiguity in multi-object scenes. The benchmark captures both single-instance scenarios and more challenging multi-instance settings where multiple objects share identical or similar text, requiring models to jointly reason about visual appearance, spatial relations, and textual content. Our densely annotated benchmark enables six complementary tasks that probe different aspects of text-object reasoning, including scene-text-object grounding, referring expression comprehension with scene text, scene-text-based visual question answering, text-region grounding, OCR in complex visual contexts, and multi-text image captioning (Fig. 2). On these tasks, we evaluate a diverse set of models spanning object detection, OCR systems, and multimodal vision-language models.

Our experimental analysis uncover a fundamental gap in current multimodal systems—the lack of instance-level compositional grounding between textual and visual information. Our key contributions are: (i) We introduce READ THE OBJECT, a novel benchmark for evaluating *text-object binding*, bridging object grounding and scene-text understanding. (ii) We construct a densely annotated dataset that explicitly links object instances with their inscribed text, enabling evaluation under complex multi-instance and ambiguous settings. (iii) We define **six complementary tasks** that evaluate text-grounded perception and reasoning, and provide the first systematic analysis showing that current vision-language models struggle with text-object association.

2 Related Work

In this section, we review the datasets related to reading text from scene images, the associated tasks, and models for reading text from natural images.

Datasets for scene-text or object detection. To advance the capabilities of AI models to detect text from images, datasets such as COCO-text [39], SynthText [14], ICDAR2013 [16], HierText [33], TextOCR [38], SynthText150k [32], LAION-OCR [35] have been proposed. These datasets provide text annotations in the entire visual context for each image sample for transcription purposes. Datasets such as OpenImageV5 Text [22] include text-spotting providing regions corresponding to particular text in the image. This is analogous to object detection where bounding-boxes or masks of particular visual objects in images are provided. To reason about the scene-text, TextCaps [36] provides an image-captioning dataset which includes the scene-text in the image description. Similarly, TextVQA [37] and ST-VQA [42] contain question-answer pairs that require reading text from images without making explicit association between objects and their associated text. Recent work such as MM-VET [47] and CC-OCR [44] include question-answer pairs that require OCR understanding in images for document parsing or information extraction. These datasets do not include object-level associations of text, and therefore, cannot be utilized for fine-grained understanding of object instances based on the related scene-text.

Object detection and grounding datasets including COCO captions [9] or Visual Genome [21] contain object bounding-box annotations but do not have

annotations for scene-text and its corresponding object. Similarly, widely dataset for referring expression RefCOCO [46] annotates objects based on the presence of object name or class in the expression but cannot detect objects when provided with the scene-text description of the object.

Vision tasks for scene-text understanding. Early work on scene-text recognition focused primarily on OCR tasks involving detection and extraction of text [5, 24, 31]. Classical tasks such as Textspotting [17] detect all the text instances in the image followed by recognition of the text in those instances. These tasks, however, do not cater to the semantic meanings of the text in images. To address this, scene-text based captioning [36] combined the task of caption generation with scene-text to enable the inclusion of image text in visual description. Scene-text understanding tasks further include visual-question answering [37, 47] requires the models answer questions about the image text. Unlike these tasks, which focus on scene-text in a global visual context, our proposed tasks require establishing relations between objects and the inscribed text.

Methods for scene-text understanding. Text parsers including CRAFT [3], TROCR [25], Masktextspotter [27], MonkeyOCR [26], or PaddleOCR [10] read text from image documents such as pdfs or images dense in text. OmniParser [40] can read text and labels, for example, from graphical user interfaces. The models do not have object detection capabilities and cannot localize targeted text or objects. Vision-language models such as LLaVA [29] and Qwen [4], trained on large-scale datasets with dense annotations, have been benchmarked on different tasks involving OCR in knowledge-based reasoning and visual question answering. Recent Deepseek OCR [41] demonstrates remarkable capabilities across different OCR tasks. The reliance of VLMs on general-purpose vision encoders and language models means that they often lack the fine-grained detection and recognition capabilities, particularly for text reading and object association.

3 READ THE OBJECT Dataset

Motivation for text–object grounding. Objects in real-world environments are often distinguished by the text written on them rather than by visual appearance alone. Beyond object or text recognition, successfully resolving ambiguities in such scenarios requires determining which text belongs to which physical object. This instance-level association between scene text and objects—what we refer to as *text–object grounding*—is critical for reliable perception and reasoning. For example, in autonomous driving, a perception system must associate the text “Exit 22” with the correct highway sign rather than nearby vehicles that may also contain text such as advertisements or license plates. In retail or warehouse settings, visually identical product packages may only differ by printed information such as size, flavor, or expiration date, requiring systems to identify the correct item based on its label (Fig. 1). Similarly, a robot may be instructed to retrieve “the box labeled fragile” or “the container marked sterile,” where multiple objects share similar shapes but differ only in textual markings.

To enable joint scene-text and object understanding in vision-language and text-based image grounding models, we introduce READ THE OBJECT dataset and benchmark. Our dataset provides rich and dense annotations of scene-texts, the associated objects in images, and a referring expression linking these objects with the scene-texts.

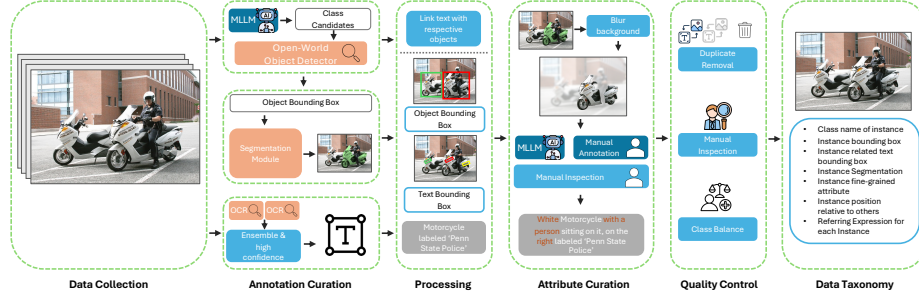


Fig. 4: Data curation pipeline for READ THE OBJECT. We build the dataset through a multi-stage pipeline: Images are first processed with open-vocabulary detectors, segmentation models, and OCR systems to extract candidate objects and text. Then, text instances are linked to their corresponding objects, followed by manual verification to ensure accurate supervision.

3.1 Benchmark Construction Process

Source datasets and preprocessing. We collect images containing objects and the scene text inscribed on them, establishing explicit associations between each object and its corresponding text. Surprisingly, there exists no dataset that directly connects text regions with their related objects within the same image. To construct such a dataset (*cf.* Fig. 4), we utilize a diverse collection of natural image sources—ranging from images in the wild to structured road scenes with traffic signs and vehicles. By combining multiple datasets with different visual domains and object distributions, our benchmark captures a wide range of object–text interactions.

- (i) **OpenImages V5.** We curate natural images with instance-level bounding boxes and segmentation masks. For objects annotated only with bounding boxes, we generate segmentation masks using SAM [20]. Existing text annotations from prior work [22] cover only a subset of the images, so we annotate remaining images using multiple OCR models and retain only the text regions where both models agree with high confidence (see suppl.).
- (ii) **MS COCO.** We use MS COCO [28] images with their instance segmentation masks and bounding boxes. Scene text is obtained from COCO-Text [39], which provides OCR annotations aligned with COCO images. We use these

annotations to establish associations between scene-text regions and the corresponding visual objects in each image.

- (iii) **NuScenes.** To capture autodrive settings where text-based grounding is essential (e.g., identifying cars, road signs), we incorporate the NuScenes [6] dataset. Beyond the provided object categories, we expand the “road-scene vocabulary” using GPT-4o (details in the suppl.). We then extract object bounding boxes using Grounding DINO [30] and obtain segmentation masks and scene-text regions following the same procedure described above.

To prepare a clean dataset with objects and its associated scene-text, we retain *legible*, *English*, and *non-crowd* text annotations. For reliable grounding, we keep text regions whose bounding boxes lie entirely within the object’s segmentation mask, discarding cases where text overlaps only with bounding boxes. Bounding boxes are often coarse and may include text not actually on the object. Then we apply *Perceptual hashing* to detect and remove duplicate images.

Scene-text based referring caption. To enable models to accurately localize objects based on the text written on them, we introduce a new set of text-to-object annotations. Each annotation provides a natural-language description that refers to the object(s) through their scene-text (e.g., “a mug labeled 12oz”), ensuring the referenced text region lies fully within the corresponding object mask (Fig. 5(c)). We consider four scenarios: (i) *Single-text* $\rightarrow$ *Single-object*, where a single unique text instance is linked to exactly one object; (ii) *Single-text* $\rightarrow$ *Multi-object*, where the same text appears on multiple instances of an object category (e.g., multiple mugs labeled “12 oz”), or across different object categories (e.g., a mug and a jug both labeled “11 oz”); (iii) *Multi-text* $\rightarrow$ *Single-object*, where a single object contains multiple distinct text regions, for which we provide both separate bounding boxes (as in PreSTU [19]) and a merged bounding box covering all referenced text regions; and (iv) *Multi-instance* $\leftrightarrow$ *Multi-text*, where multiple objects each contain multiple or repeated text instances, such as retail products on a shelf with their own labels or price tags.

Attribute generation and disambiguation. In images containing multiple instances of the same object, scene-text alone is insufficient for distinguishing between them. To achieve unambiguous grounding, we augment both the scene-text and object samples with fine-grained visual attributes (e.g., color, material, shape) and positional cues (e.g., left, right, top, center).

To extract these attributes, we first crop each object instance and blur the background to ensure the model focuses solely on the object. Following this, multiple multimodal LLMs are prompted to generate candidate attributes. When two instances share identical attributes, human annotators provide a distinguishing attribute that is unique to each instance. A second round of verification ensures consistency and correctness (details in the suppl.), yielding discriminative referring expressions for multi-instance scenarios. To obtain high-quality samples, we manually verified attribute consistency.

Dataset consolidation and taxonomy. We provide a rich set of annotations for the curated dataset. To ensure high quality dataset, we manually validated

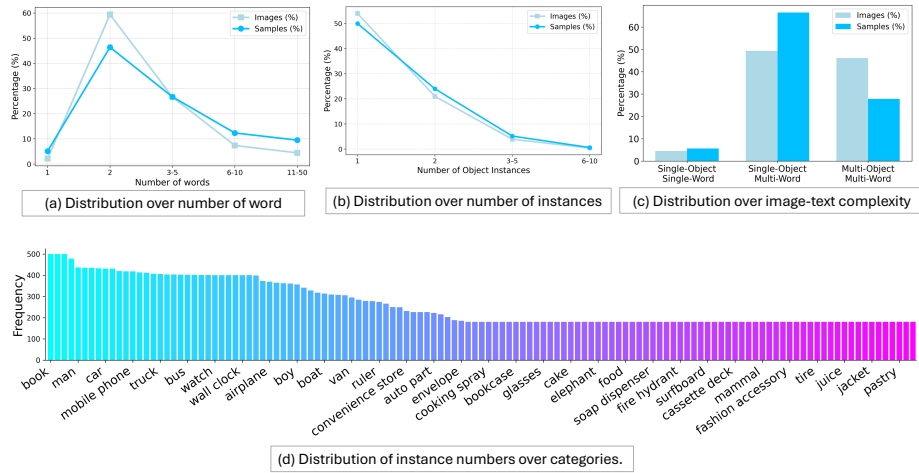


Fig. 5: Dataset statistics. READ THE OBJECT comprises of high quality and challenging dataset with dense annotations.

random subsets from each split to ensure grounding and diversity. For each image sample, we include (i) the class names for the object instances in the image, (ii) the bounding box annotations for the object instances, (iii) the scene text associated with each instance, *i.e.*, the text occurring within the object’s bounding box and its corresponding text region, (iv) the instance segmentation mask and fine-grained attributes of the object (color, shape, size), (v) the relative position of the instance with respect to other objects in the image, and (vi) a referring expression with scene-text and object instance pairs.

The dense and extensive annotations in our dataset make it applicable to a wide range of tasks including referring expression comprehension, text-grounded VQA, captioning, text-object grounding, and text-region grounding. Details of each task can be found in Tab. 4, Fig. 2. (More in Appendix B of suppl.)

3.2 Dataset Statistics and Analysis

READ THE OBJECT constructs 25,208 samples with over 34,085 object instances for the referring expression task from over 6,000 images. Our dataset exhibits high scene-text density, averaging more than five text regions per image. As shown in Fig. 5(a), over 90% of samples contain more than one scene-text instance, with a long tail extending to 11–50 text regions per image, highlighting the richness and complexity of the scenes. In addition to dense scene-text annotations, the dataset also contains diverse object instances, with an average of 1.35 objects per image. The distribution of samples over the number of object instances is shown in Fig. 5(b) and Tab. 1. Overall, the dataset is well distributed across 97 object classes, as illustrated in Fig. 5(d). For each object category, the imbalance factor remains ≤ 4 , with samples ranging between 100 and 400.

Table 1: Object- and sample-level statistics of READ THE OBJECT for the referring expression task. Average composition and distribution of object and text instances per sample.

Statistic	Value
Total samples	25,208
Total object instances	34,085
Average objects per sample	1.35
Average texts per sample	5.44
Avg. objects per sample (text>1)	1.37
Avg. objects per sample (object>1)	2.41

Table 2: Text-level statistics of READ THE OBJECT. Counts and lexical properties of all annotated text instances across 6,000 images, illustrating the high density and diversity of scene text.

Statistic	Value
Total text instances	137,246
Alphanumeric texts	117,327
Alphanumeric & ≥ 3 characters	68,406
Non-alphanumeric texts	4,057
Texts with ≤ 3 characters	53,586
Rotated text ($> 20^\circ$)	8,493

Table 3: Comparison of referring expression datasets. Prior REC datasets rely on visual attributes (appearance and spatial cues) to localize objects. READ THE OBJECT additionally annotates scene-text regions and explicitly links them to object instances, enabling evaluation of text-object association in referring expressions, based on both detailed visual attributes and textual content.

Dataset	Image Count	Sample Count	Avg. Expression Length	Scene-text BBox	Object-Text Association	Detailed Visual Attr.
RefCOCO	3,000	7,596	3.6	✗	✗	✗
RefCOCO+	3,000	7,578	3.6	✗	✗	✗
RefCOCOg	3,900	7,596	8.4	✗	✗	✗
Ref-Adv	2,833	5,000	11.5	✗	✗	✓
Ours	6,000	25,208	17.6	✓	✓	✓

To characterize the difficulty of joint object-text understanding, Fig. 5(c) reports the proportions of samples across three configurations: single-object-single-text, single-object-multi-text, and multi-object-multi-text. Fewer than 5% of samples fall into the simplest configuration, while roughly $\sim 30\%$ contain multiple objects and text regions, underscoring the complexity of the benchmark.

The dataset contains 137,246 scene-text instances with substantial variation in type, length, and orientation, including text rotated more than 20 degrees. Tab. 2 summarizes these properties, showing a mix of alphanumeric and non-alphanumeric text with an average word length of approximately four characters.

Finally, Fig. 6(c) provides statistics of the READ THE OBJECT benchmark across six tasks. A single image may generate multiple samples depending on the task. Fig. 6(a) shows that the ground-truth captions in our image captioning task are relatively descriptive, with an average length of around 20 words, making the task more challenging for models reasoning over scene text. Similarly, Fig. 6(b) illustrates the balanced distribution of question types in our VQA benchmark. Samples for each task is Fig. 7. Additional details on the benchmark tasks are provided in Appendix B.

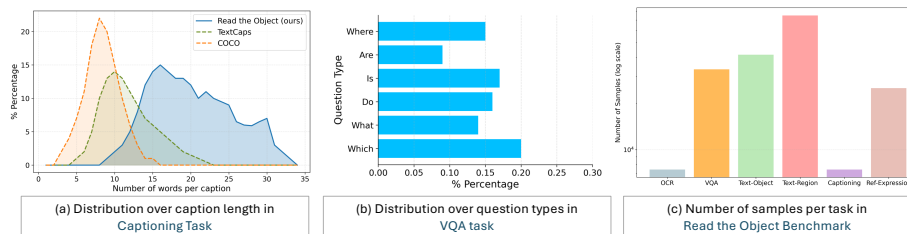


Fig. 6: Task-specific statistics. We show the benchmark statistics of READ THE OBJECT for different tasks (a) distribution of caption length for captioning; (b) question types for VQA, and (c) Number of annotated samples for each task.

Table 4: Overview of tasks supported by READ THE OBJECT. The benchmark evaluates text-object association through six tasks spanning grounding, referring expressions, OCR, VQA, and captioning.

Task	Definition	Input	Output
Text-Object Grounding	Given a text string, find all objects in the image that contain or are associated with that text.	Image, Target Text	Object name(s) + Object bounding box(es)
Referring Expression Comprehension	Given a natural-language expression that includes both text and visual attributes, locate the described object and its text region.	Image, Referring Expression	Object bbox + Text bbox
OCR	Detect and recognize all visible text regions in the image.	Image	Text regions + Transcriptions
Text-Region Grounding	Given a query text string, localize the corresponding text region(s) within the image.	Image, Query Text	Text bounding box(es)
Visual Question Answering	Answer questions that require reading text and understanding its relation to the visual context.	Image, Question	Answer text
Multi-Text Captioning	Generate a caption that jointly describes objects and the visible text on them.	Image	Caption text

3.3 Comparison with Existing Benchmarks

Tab. 3 compares READ THE OBJECT with widely used referring expression comprehension (REC) benchmarks, including RefCOCO [46], RefCOCO+ [8], RefCOCOg [18], and Ref-Adv [11]. These datasets are primarily constructed from the MS COCO image collection [28], which provide object annotations and natural images. While they have been instrumental in advancing referring expression grounding, they rely mainly on visual attributes—such as color, appearance, and spatial relations—to identify objects. Importantly, these benchmarks do not annotate scene-text regions or establish explicit associations between text and the object instances on which the text appears. As a result, they cannot evaluate whether models can identify objects based on the text written on them.

In contrast, READ THE OBJECT is constructed from multiple complementary sources, including the latest OpenImages version, MS COCO [28], and NuScenes [6]. Combining these datasets allows us to capture a broader range of

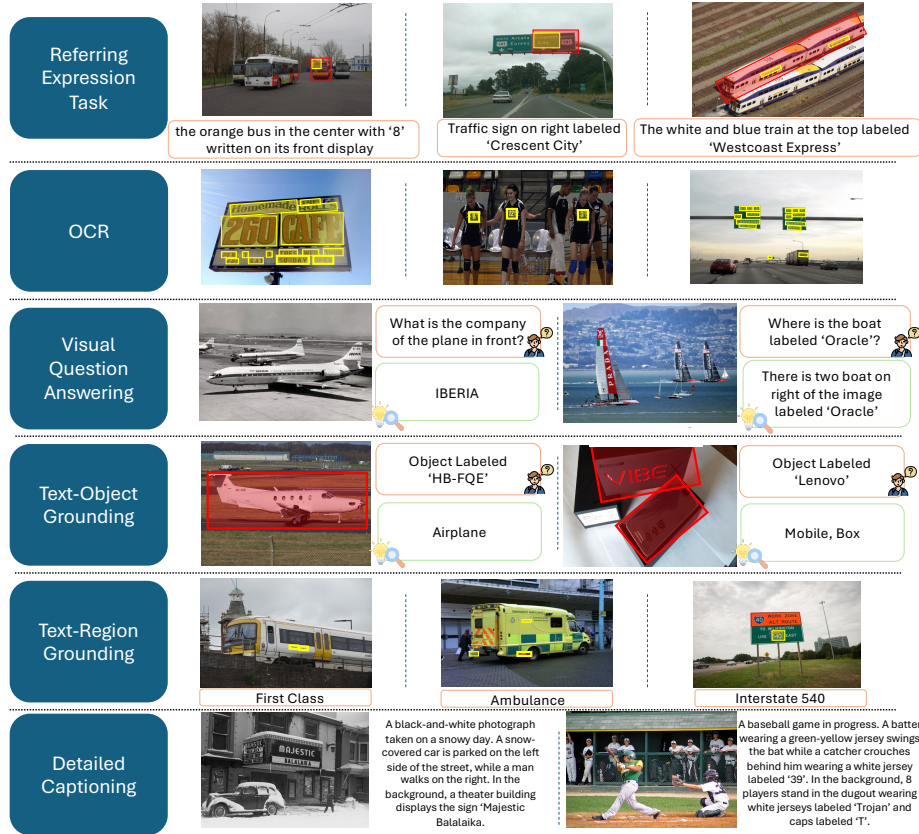


Fig. 7: Representative samples from the six tasks supported by READ THE OBJECT dataset. Yellow boxes denote scene-text bounding boxes, while red boxes denote object bounding boxes, highlighting the challenge of associating textual content with the correct object instances in complex real-world scenes.

real-world scenes, from everyday object-centric environments to road scenes containing traffic signs and vehicles. Unlike prior REC datasets, READ THE OBJECT explicitly annotates scene-text regions and links them to the corresponding object instances, enabling evaluation of text-object association. This design allows referring expressions to depend jointly on visual attributes and textual content, which is necessary for studying text-grounded perception and reasoning.

4 Experiments

Given the diversity and density of the annotations in READ THE OBJECT benchmark, we evaluate our model across tasks as outlined in Tab. 4. Our tasks integrate scene understanding with object localization, attribute selection, scene-text recognition, and visual reasoning. In the following, we validate and analyze the performance of different models on our benchmark for scene-text-based referring expression comprehension, OCR, and VQA.

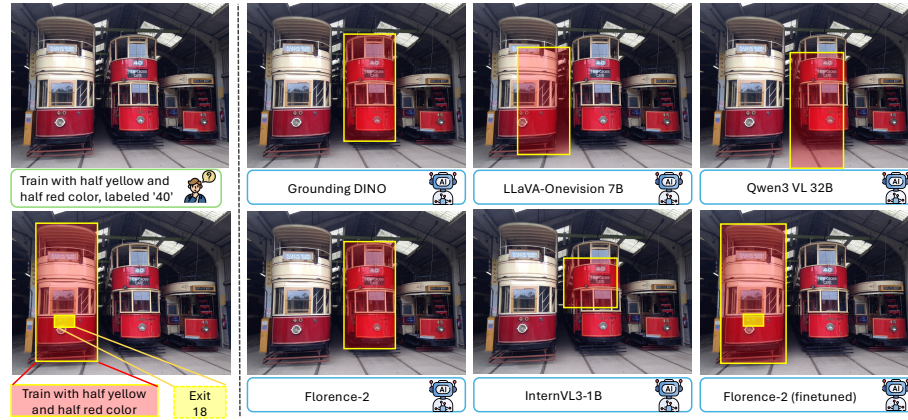


Fig. 8: Qualitative comparison of different models on READ THE OBJECT benchmark for referring expression task. Left: Ground truth. Right: Output of different models.

Table 5: Evaluation on READ THE OBJECT benchmark for referring expression and object bounding box.

Category	Model	Avg mAP	mAP@0.50	mAP@0.75	mAR@1
Object Grounding Models	MM Grounding DINO [47]	27.19	35.41	29.61	35.37
	Florence-2 [43]	15.71	23.50	14.13	22.30
	Grounding DINO [30]	17.78	20.33	19.75	21.44
	YOLO-world [45]	14.74	21.92	16.18	21.02
Multimodal LLMs	LLaVA-NeXT 1.6 7B [23]	21.27	38.44	20.99	41.62
	InternVL3 8B [48]	2.41	11.37	0.89	11.37
	Qwen3-VL 2B Instruct [34]	5.25	17.69	1.86	16.28
	Qwen3-VL 8B Instruct [34]	5.66	16.34	2.99	19.18
	Qwen3-VL 30B Instruct [34]	4.87	17.52	0.97	16.14
	LLaVA-OneVision 7B [23]	12.72	31.66	7.92	29.54
	DeepSeek-OCR [41]	39.21	45.33	34.63	48.51
Fine-tuned Model	Florence-2 (Object)	53.62	55.19	51.97	52.19
	Florence-2 [43] (Text)	35.96	39.21	31.09	39.21

4.1 Referring expression with READ THE OBJECT

Referring expression comprehension is used as a standard task for evaluating the object detection and localization capabilities of object grounding and vision-language models. In this, given an input image and a referring text, the models are expected to identify *i.e.* provide bounding boxes of the objects based on the description. In our benchmark, we consider *the referring expression descriptions with scene-text*. To evaluate and analyze the performance of state-of-the-art models on READ THE OBJECT benchmark, we consider object grounding models – MM-grounding DINO [47], Florence-2 [43], Grounding DINO [30], and YOLO-world [45]. We extensively evaluate the advanced multimodal large language models, including LLaVA-Next 1.6 7B [23], InternVL3-8B, InternVL3-1B [48], Qwen3-VL-2B-Instruct [34], Qwen3-VL-8B-Instruct [34], Qwen3-VL-30B-Instruct [34], LLaVA-OneVision-7B (SI) [23], and the recent DeepSeek-OCR

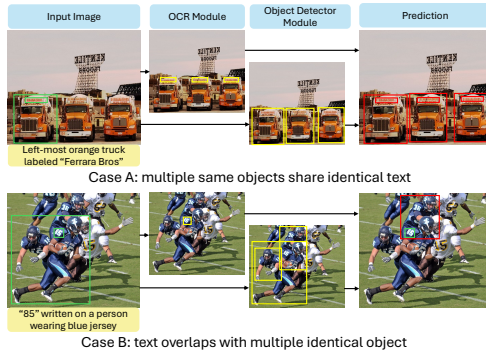


Fig. 9: Failure modes of simple baseline.
Two stages (OCR + object detection).

Table 7: Performance comparison in the OCR benchmark.

Model	ANLS
CRNN [12]	0.57
TrOCR [25]	0.77
Mask TextSpotter v3 [27]	0.82
OmniParser [40]	0.86
Florence-2 [43]	0.89
DeepSeekOCR [41]	0.92

Table 6: Avg. mAP for modular baseline analysis.

	Unique Text + Single Object	Unique Text + Multi Object	Multi Text Instance + Multi Object
DS+Grounding-DINO	79.02	36.31	17.93
DS + Florence-2	81.43	34.11	15.09

Table 8: Performance comparison on the VQA benchmark.

Model	ANLS	ROUGE-L
LLaVA-OneVision 7B [23]	0.39	0.41
Qwen3-VL 8B Instruct [34]	0.36	0.58
LLaVA-NeXT 1.6 7B [23]	0.38	0.44
InternVL 3.5 8B [48]	0.34	0.42
Qwen3-VL 30B Instruct [34]	0.38	0.55

and DeepSeek-vl-1.3b-chat [41]. To provide a simple reference point, we also construct a two-stage modular baseline that independently detects text and objects. Then we post-process them to get the bounding box for both text and related objects for the referring expression task.

Evaluation metrics and results. We evaluate the mean average precision (mAP) and the mAP across different intersection over union (IoU) thresholds *i.e.* mAP@0.50 and mAP@0.75. To compute the mAP, we first compute the IoU, the overlap between the predicted bounding box and the ground truth bounding box. We then compute the area under the precision-recall curve for each of the instances, providing the average precision (AP) and mAP is the mean over all the classes. Additionally, we report the mean average recall (mAR@1) for the predictions with the highest confidence.

The quantitative results in Tab. 5 show that existing text-conditioned object grounding models perform poorly on READ THE OBJECT, highlighting the difficulty of jointly reasoning about scene text and visual objects. Even recent multimodal large language models trained on large-scale image-text data exhibit a substantial performance gap on our benchmark. For example, LLaVA-OneVision-7B achieves less than 13 mAP. To analyze the effect of model scale, we evaluate the Qwen family of models with 2B, 8B, and 30B parameters. Despite the increase in model capacity, performance improves only marginally, suggesting that scaling does not resolve the challenge.

To further examine whether training on the scene-text annotations enhances model capability, we fine-tune Florence-2, a strong referring expression grounding model, on the training split we curated for this task (in Appendix C). Fine-tuning yields significant improvements of 37.92% for object grounding and 20.26% for text grounding, indicating that the benchmark provides useful supervision for learning text-object associations. However, the overall performance remains far from saturated, suggesting substantial room for improvement and highlighting the difficulty of the task on our proposed benchmark. Additionally, qualitative examples in Fig. 8 demonstrate that both object grounding and multimodal approaches frequently fail to localize the correct object based on scene-text. Given an image and the referring expression “train with half yellow and half red color, labeled 40,” the model must jointly reason about object class, visual attributes, and scene text. Florence-2, fine-tuned on our dataset, can correctly localize both the object bounding box and the associated text region.

Modular baseline analysis. To understand whether the task can be solved using existing pipelines, we construct a simple modular baseline that independently detects scene text and objects and then associates them. Specifically, we combine DeepSeek-OCR with strong object detectors such as Florence-2 and Grounding-DINO. As shown in Tab. 6, this baseline performs reasonably well in simple cases containing a single object-text pair. However, its performance drops sharply in cluttered scenes where multiple objects share similar text Fig. 9(a) or where text regions overlap with multiple objects Fig. 9(b). These results indicate that independently detecting text and objects is insufficient for reliable scene-text grounding, highlighting the need for models that explicitly learn instance-level text-object associations.

4.2 OCR detection with READ THE OBJECT

We evaluate across a wide range of state-of-the-art models for OCR detection, specifically, CRNN [12], TrOCR [25], Mask TextSpotter v3 [27], Omniparser [40], Florence-2 [43] and DeepSeekOCR [41].

Evaluation metrics and results. Following prior work, we evaluate for the Average Normalized Levenshtein Similarity (ANLS). It measures the number of edits in the predicted text required to reach the ground truth. Additional metrics are included in the supplemental. The evaluation of different OCR models on our READ THE OBJECT benchmark in Tab. 7 reveals the high-quality of our annotations and it aligns with the expected OCR performance across different models. Notably, Florence-2 performs on par with DeepSeekOCR. However, it struggles on the task of scene-text-based referring expression reinforcing our observation on the complexity of scene-text-based object grounding.

4.3 VQA on READ THE OBJECT

In our benchmark, VQA requires an understanding of the scene-text to reason about the visual content. We evaluate for the VQA task LLaVA-Next 1.6 7B [23],

InternVL3 8B [48], Qwen3-VL 8B Instruct [34], LLaVA-OneVision 7B [23], and Qwen3-VL 30B Instruct [34], shown in Tab. 8.

Evaluation metrics and results. We quantify the VQA performance on standard metrics: similar to OCR, we report the ANLS score and Rouge-L, which measures the degree of overlap between the generated text and the ground truth based on the length of the longest common subsequence. The multimodal models struggle to perform on the scene-text-based visual question answering tasks with scores being around 40% across all the models. This again shows the inability to jointly reason on scene-text and objects.

5 Conclusion

We present the importance of a joint understanding of scene-text and objects for real-world visual reasoning. We introduced READ THE OBJECT benchmark with dense scene-text and object annotations to establish relationships between objects and their associated scene-texts. Our comprehensive analysis, evaluating diverse vision, OCR, and multimodal models across six complementary tasks, reveals that current object grounding and vision-language models struggle significantly with text-object association, especially in cluttered environments. We believe this benchmark will steer the research on multimodal models that can navigate and understand the text rich complexities of the real world.

Acknowledgment. Shweta Mahajan is funded in part, from the Canada Research Chairs Program.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI, J.: EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including Latin, Chinese, Arabic, Devanagari, Cyrillic, etc. <https://github.com/JaidedAI/EasyOCR> (2020)
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9365–9374 (2019)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 71–79 (2018)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: Nusenes: A multimodal dataset for autonomous driving. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 11618–11628 (2020)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)

8. Chen, J., Wei, F., Zhao, J., Song, S., Wu, B., Peng, Z., Chan, S.H.G., Zhang, H.: Revisiting referring expression comprehension evaluation in the era of large multimodal models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 513–524 (2025)
9. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
10. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
11. Dong, Q., Yang, K., Ju, L., Zhao, H., Zhang, Y., Wang, Y., Zeng, H., Lu, J., Fu, Y.: Ref-adv: Exploring mllm visual reasoning in referring expression tasks. arXiv preprint arXiv: 2602.23898 (2026)
12. Fu, X., Ch’ng, E., Aickelin, U., See, S.: Crnn: a joint neural network for redundancy detection. In: IEEE International Conference on Smart Computing. pp. 1–8 (2017)
13. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2917–2927 (2021)
14. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2315–2324 (2016)
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
16. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: International Conference on Document Analysis and Recognition. pp. 1516–1520 (2019)
17. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: European Conference on Computer Vision. pp. 512–528 (2014)
18. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 787–798 (Oct 2014)
19. Kil, J., Changpinyo, S., Chen, X., Hu, H., Goodman, S., Chao, W.L., Soricut, R.: Prestu: Pre-training for scene-text understanding. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15224–15234 (2023)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
21. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* **123**(1), 32–73 (2017)
22. Krylov, I., Nosov, S., Sovrasov, V.: Open images v5 text annotation and yet another mask text spotter. In: Asian Conference on Machine Learning. pp. 379–389 (2021)
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

24. Li, H., Wang, P., Shen, C.: Towards end-to-end text spotting with convolutional recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5238–5246 (2017)
25. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13094–13102 (2023)
26. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)
27. Liao, M., Pang, G., Huang, J., Hassner, T., Bai, X.: Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In: European Conference on Computer Vision. pp. 706–722 (2020)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
31. Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J.: Fots: Fast oriented text spotting with a unified network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5676–5685 (2018)
32. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9809–9818 (2020)
33. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1049–1059 (2022)
34. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), <https://arxiv.org/abs/2412.15115>
35. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
36. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European Conference on Computer Vision. pp. 742–758 (2020)
37. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)

38. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8802–8812 (2021)
39. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
40. Wan, J., Song, S., Yu, W., Liu, Y., Cheng, W., Huang, F., Bai, X., Yao, C., Yang, Z.: Omniparser: A unified framework for text spotting key information extraction and table recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15641–15653 (2024)
41. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
42. Xia, H., Lan, R., Li, H., Song, S.: St-vqa: shrinkage transformer with accurate alignment for visual question answering. *Applied Intelligence* **53**(18), 20967–20978 (2023)
43. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4818–4829. IEEE (2024)
44. Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., et al.: Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21744–21754 (2025)
45. Yaseen, M.: What is yolov9: An in-depth exploration of the internal features of the next-generation object detector. arXiv preprint arXiv:2409.07813 (2024)
46. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision. pp. 69–85 (2016)
47. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2023)
48. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)

READ THE OBJECT: A Dataset and Task for Text-Grounded Object Understanding in Real-World Images –Supplementary Material–

The supplementary material is organized as follows:

- section A details the full benchmark curation pipeline (corresponding to section 3 in the main paper).
- section B provides task-specific curation details for all benchmark tasks (corresponding to subsection 3.2 in the main paper).
- section C describes the training data curation process (corresponding to Figure 4 in the main paper) and lists all training hyperparameters.
- section D provides more experimental results.

A Benchmark Curation Process

This section provides detailed explanations of our benchmark generation process and the attribute-quality verification steps, which were omitted from the main paper for brevity.

A.1 Source Datasets and Preprocessing

NuScenes Class Expansion The NuScenes dataset provides bounding boxes for a small set of autonomous-driving related categories such as *car*, *truck*, *van*, *pedestrian*, and *cyclist*. However, for text-grounded object understanding, many additional road-scene entities contain visually meaningful text (for example, *billboards*, *traffic sign*, *storefronts*).

To expand this “road-scene vocabulary,” we prompt GPT-4o with diverse NuScenes images to identify additional object categories containing text. Specifically, we sample 50 visually distinct frames from 10 selected NuScenes driving sequences (front, front-left, and front-right cameras) and provide each image along with the prompt shown in Figure 11b. GPT-4o returns a set of candidate object categories for each image.

We then merge and normalize these category lists using GPT-4o to remove synonyms and duplicates, producing an expanded vocabulary of road-scene objects likely to contain scene text. Finally, we use this expanded class list as input to Grounding DINO [30] to obtain bounding boxes for these newly introduced object categories across the full NuScenes dataset.

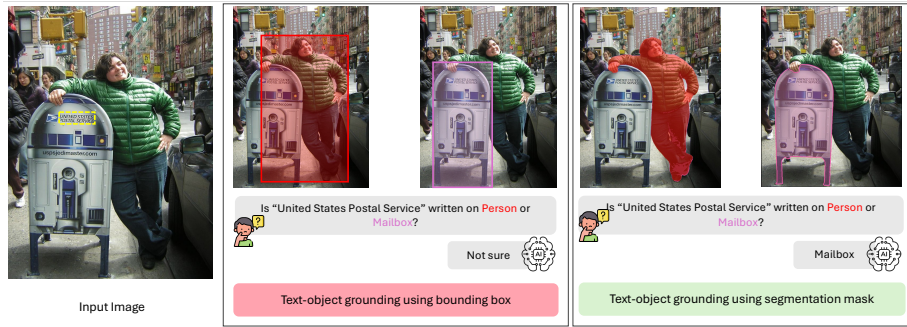


Fig. 10: Segmentations masks for text-object grounding. Coarse object bounding boxes lead to ambiguous text-object associations, while segmentation masks provide precise spatial boundaries for accurate correspondence.

OCR Aggregation and Text Annotation Pipeline The OpenImages V5 training set includes partial OCR supervision from prior work [22], covering only subsets 1,2,5,f, which account for roughly a quarter of the total images. These annotations are sparse and primarily contain single-object and single-text cases. To obtain comprehensive OCR coverage across the remaining subsets (0,3,4,6 to 9,a to e), we run two independent OCR systems: EasyOCR [2] and PaddleOCR [10] on every image.

For each detected text region, we record (i) the recognized text string, (ii) the confidence score from each OCR model, and (iii) the predicted bounding box. We retain a text bounding box only if:

- the recognized text content from both OCR models matches exactly,
- both confidence scores exceed 0.7, and
- the IoU between the two predicted text boxes exceeds 0.7.

This yields high-quality, model-consistent OCR annotations for all OpenImages V5 images. We use the same approach to get the OCR bounding box for the Nuscense dataset as well.

Segmentation-based Text-Object Association Given the large number of images and their object bounding boxes, we construct text-object association candidates by checking which object bounding boxes intersect each text bounding box. However, as illustrated in Figure 10, object bounding boxes in images can be coarse and overly wide, causing ambiguous associations. For example, a text bounding box containing the word “United States Postal Service” may lie inside both the bounding box of a person and a nearby mailbox, even though the text physically appears only on the mailbox.

To resolve these ambiguities, we generate a segmentation mask for each object using SAM [20], using the provided object bounding boxes as prompts. We then determine the correct text-object correspondence by checking whether the text bounding box lies entirely within the object’s segmentation mask. As shown



(b)

A.2 Attribute Generation and Disambiguation

To construct discriminative attributes, we first ask GPT-4o to generate a list of semantically meaningful attribute types for each object category (e.g., for person: hair color, clothing color, body pose, accessories; for vehicle: body color, make, visible branding). The prompt used for this task is shown in Figure 12, first row. These attribute types define the visual cues that should be probed from multimodal models.

For each object instance in the image, we crop the object and blur the remaining scene to ensure that models focus solely on the target instance. We then prompt two multimodal LLMs: LLaVA-OneVision [23] and MiniGPT-4 [1], to describe the object according to the attribute types proposed by GPT-4o. For every attribute type (e.g., hair color, action), we query both models separately and record their candidate answers. To improve correctness, we perform

Model	Task Description	Prompts
GPT-4o	Attribute-Type Generator Prompt (per object category)	SYSTEM PROMPT: You are an expert in fine-grained visual recognition. Your task is to produce a compact list of attribute types that are useful for distinguishing multiple instances of the same object category within the same image. • Output only generic attribute types, not specific values. • Attributes should be visually observable and instance-specific (e.g., for a person: hair color, clothing color, pose, accessories; for cars: body color, brand logo, roof shape). • Do not include text unless the object itself contains printed text. • Do not describe any particular object. • Output format: ["attribute_type_1", "attribute_type_2", ...]
LLaVA and MiniGPT-4	Attribute Extract Per Object Instance	SYSTEM PROMPT: You are an expert vision-language model. You will be given a cropped image of a single object instance. The background may be blurred; focus only on the visible object. • You must answer strictly about the specific attribute type being asked (e.g., "hair color," "shirt color," "pattern," "pose"). • Your answer must be: short, factual, and directly grounded in the visible crop. • Do not describe irrelevant regions. • Do not speculate if an attribute is not visible—reply "UNSURE". USER PROMPT: Attribute type: hair color Question: "What is the person's hair color?" Image
LLaVA and MiniGPT-4	Cross-checking Attribute Generation	SYSTEM PROMPT: You will be shown an image and a candidate answer describing one attribute of the object. Your task is to verify whether the answer matches what is visible in the image. Output "CORRECT" or "INCORRECT". Do not explain. Do not infer beyond the visible region.
GPT-4o	Attribute-Set Difference Checker Prompt	SYSTEM PROMPT: You will be given two sets of attributes describing two object instances from the same image. Your task is to determine whether the two instances differ on any attribute. If differences exist, list only the attribute types where they differ (e.g., "hair color", "accessory", "shirt color"). If there is no meaningful difference, output: "NO_DIFFERENCE". Do not invent attributes. Be concise and precise.

Fig. 12: Task-specific prompt templates for attribute disambiguation. Overview of the tasks and corresponding prompt templates used for GPT-4o, LLaVA, and MiniGPT-4 to generate, verify, and differentiate fine-grained attributes for multi-instance disambiguation.

cross-verification: answers generated by MiniGPT-4 are validated by LLaVA-OneVision on the same crop (and vice versa). Details of LLaVA and MiniGPT prompts is described in Figure 12. Inconsistent or uncertain outputs are flagged for human inspection.

After gathering attributes for each instance, we ask GPT-4o (without images) to compare the attribute sets of two instances from the same image and identify whether any attribute differs (e.g., "Instance A has dark hair; Instance B has light hair"). If a distinguishing attribute exists, we use it as the disambiguating cue in the referring expression.

If GPT-4o finds no meaningful difference, the pair is escalated to human annotators. Annotators view the two cropped instances (from the same image) and determine whether any fine-grained distinguishing cue exists (e.g., presence of glasses, stripe pattern, backpack). All proposed differences are cross-validated by three additional annotators. If no appearance-based difference can be found, annotators fall back to relative positional cues (e.g., "left clock," "right clock," "closer to the mailbox").

Through this multi-step automatic + human pipeline, we obtain reliable, fine-grained, instance-level distinguishing attributes for all multi-instance cases in the dataset. The final data distribution is shown in Figure 13.

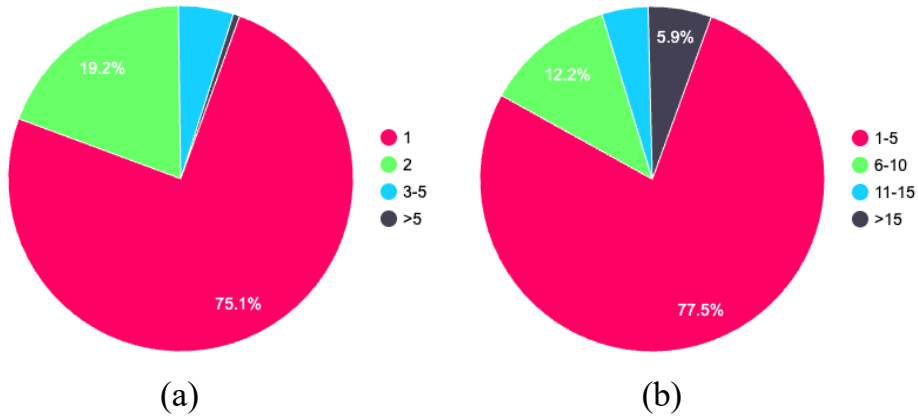


Fig. 13: Instances in READ THE OBJECT benchmark. (a) Percentage of number of object instances, (b) Percentage of number of text instances in READ THE OBJECT benchmark

B Task Curation Details

In this section, we describe each of the six tasks in our READ THE OBJECT benchmark: OCR, VQA, Text-Object Grounding, Text-Region Grounding, Captioning, and Referring Expression Comprehension. We also describe their generation procedures, quality-control steps, and sample images illustrated in the accompanying figures.

B.1 Referring Expression Comprehension

We described the full curation pipeline for our referring-expression comprehension dataset in the main paper and in section A. Here, we provide a concise summary.

We begin with three source datasets and annotate text regions, segmentation masks, and object instances, where it was needed. We then establish precise text-object associations using our mask-guided linking procedure and generate fine-grained attributes for cases where multiple object instances share the same scene text. These attributes include visual cues (e.g., color, material, shape) and spatial cues (e.g., left, right, top), enabling unambiguous identification of the correct instance.

Each sample is finally converted into a natural referring expression by combining the object category, scene text, attributes, and spatial context (e.g., ‘toyota’ on the white car on the left side of the road). In total, we collected a total of 25,208 samples. This results in one of the most comprehensive and fine-grained referring-expression grounding datasets, addressing both text grounding and multi-instance disambiguation.

Example referring-expression annotations are shown in Figure 14.

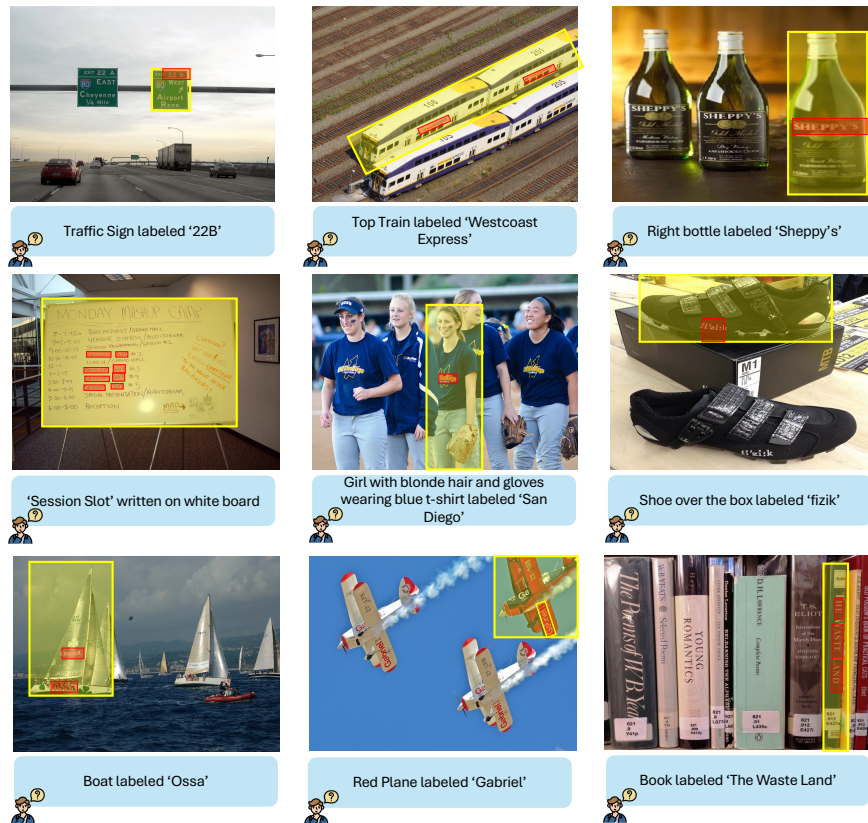


Fig. 14: Additional qualitative examples. Example samples for **Referring Expression Comprehension** task in **READ THE OBJECT** benchmark.

B.2 Scene-Text-Aware Caption

For the Scene-Text-Aware Captioning task in READ THE OBJECT, our goal is to produce high-quality descriptive captions that explicitly incorporate both object attributes and the scene text present in the image.

Caption Generation. We first generate candidate captions for each image using four multimodal LLMs: DeepSeek-OCR, LLaVA-OneVision, Qwen2.5-VL-72B, and Florence-2. We use a simple captioning prompt instructing each model to produce a concise (1–2 sentence) description of the scene. To construct a reliable ground-truth caption, we then provide GPT-4o with:

- the four LLM-generated captions,
- the text and its bounding boxes, and
- the object category and bounding boxes along with attributes.

We instruct GPT-4o to treat the object and text annotations as ground-truth and to synthesize a single, accurate, 1–2 sentence caption that faithfully describes the image while incorporating the relevant scene-text. We also supply one example demonstration per image to guide GPT-4o’s formatting and style.

Human Verification. Finally, annotators manually review all 6000 captioned images. They edit any caption where GPT-4o’s output is incomplete, misleading, or inconsistent with the image itself. This produces a high-quality, text-grounded caption for every image, explicitly linking visual objects and the text written on them.

Example captions are provided in Figure 20.

B.3 OCR-Grounded VQA

For the OCR-Grounded VQA task, we generate question-answer pairs that require both text recognition and object understanding. Each VQA sample is derived from the Referring Expression Comprehension subset of READ THE OBJECT. Instead of using the image directly, we provide GPT-4o with all ground-truth metadata for the sample, including: object categories and attributes, object bounding boxes, text content and text bounding boxes, and the verified caption for the image. GPT-4o is then prompted to produce diverse, open-ended VQA questions requiring visual-text reasoning, using only this structured ground truth (no image).

We use the prompt provided in Figure 15, which instructs GPT-4o to generate 3–5 diverse questions that cannot be answered without the scene-text and object grounding information. The questions span text-object association, spatial reasoning, attribute-based reasoning, and existence/count queries. Each answer is a short, factual explanation that integrates scene-text and object attributes.

The distribution of question types in our VQA benchmark is shown in the main paper. As illustrated, the dataset is balanced across major interrogative forms such as *Which*, *What*, *Do*, *Is*, *Are*, and *Where*, ensuring diversity in reasoning patterns. Representative VQA examples are provided in Figure 16.

SYSTEM PROMPT:

You are generating detailed, image-grounded question-answer pairs for a TextVQA benchmark. Each question must require the model to read the text and understand the visual objects or spatial layout. It should be impossible to answer correctly without seeing the image.

Context:

In this image, there is: {obj_descriptions}

Four captioning models produced the following captions for extra context:

{caption_1} {caption_2} {caption_3} {caption_4}

Generate 3–5 diverse question-answer pairs in valid JSON format, following these principles:

Goals

- Every question must require *visual-text reasoning*, not language-only guessing.
- Answers must combine text recognition, object grounding, or spatial/attribute reasoning.
- Each answer should be a short explanatory sentence (1–2 lines), not just a single word.

Types of questions to include

- Text–Object Association:
“Which object has the text ‘X’ on it?”
- Object–Text Recognition:
“What text is written on the [object with specific attribute]?”
- Spatial Relationship:
“Is the object with text ‘X’ to the left or right of [another object]?”
- Existence / Count:
“Is there any object with the text ‘X’ visible?”
- Attribute Comparison:
“Which object with ‘X’ has a different color or shape?”
- Yes/No Reasoning:
“Do both objects show the same word?”
- Compositional Reasoning:
“Which object has both [attribute] and text ‘X’?”

Rules

- Use only the provided objects, text, and bounding boxes.
- Do not hallucinate new text or new objects.
- Avoid trivial, repetitive, or language-guessable questions.
- Each answer must be grounded in the provided attributes and text.

Fig. 15: VQA prompt template. Gpt-4o prompt for generating samples for OCR-Grounded VQA task.

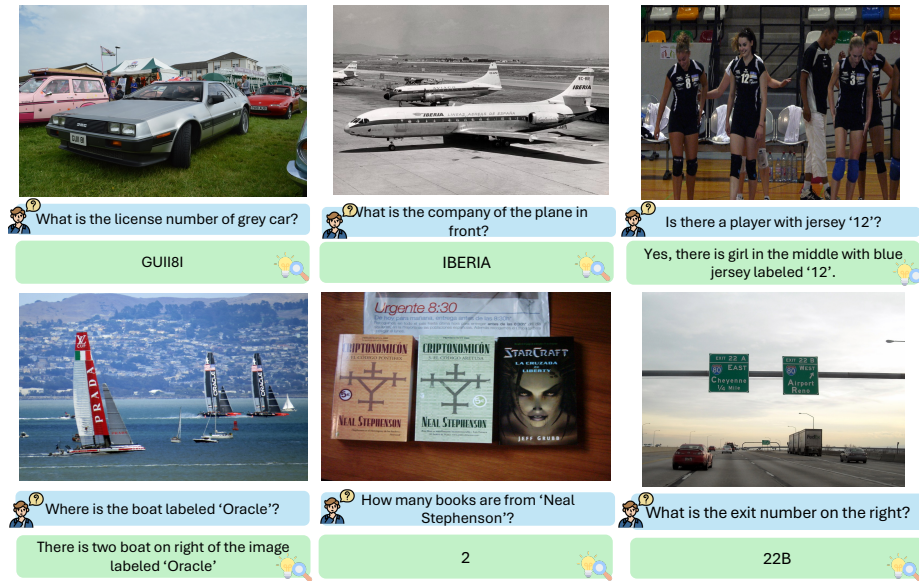


Fig. 16: Qualitative examples on OCR-VQA. Example samples for **OCR-Grounded VQA** task in **READ THE OBJECT** benchmark

B.4 OCR

For the OCR task, we leverage the text annotations already produced during our referring-expression curation pipeline. Specifically, we aggregate text regions from multiple OCR systems, retaining only high-confidence text bounding boxes based on cross-model agreement (see subsection A.1). In addition, we incorporate existing annotated text regions from COCO-Text [39] and OpenImagesV5 OCR annotations [22].

Across all sources, our OCR subset contains **137,246** text bounding boxes spanning approximately 6,000 images. In Figure 11a, the word cloud visualizes the relative occurrence of object categories, where larger words correspond to more frequent classes. Representative examples of our OCR annotations are shown in Figure 17.

B.5 Text-Object Grounding

This task evaluates whether a model can localize all object instances associated with a given scene-text. For example, if the query text is “New York,” the model must identify every object in the image (e.g., a billboard, a storefront sign) on which this text appears. Unlike referring-expression comprehension, this task does not require fine-grained attribute disambiguation. In multi-instance scenarios, the correct output is therefore a set of object category names along with bounding boxes for all instances bearing the specified text.



Fig.17: Qualitative results on OCR. Example samples for **OCR** task in READ THE OBJECT benchmark

The need for this task arises from real-world settings where text appears on multiple objects, and applications—such as visual search, inventory management, or robotic selection—require retrieving **all** objects containing a specific text string. Although this task can be derived from our referring-expression annotations, we explicitly define it as a standalone benchmark to measure a model’s ability to map text regions back to all relevant visual objects.

Our final Text-Object Grounding task contains **34,085** object instances. Representative examples are shown in Figure 18.

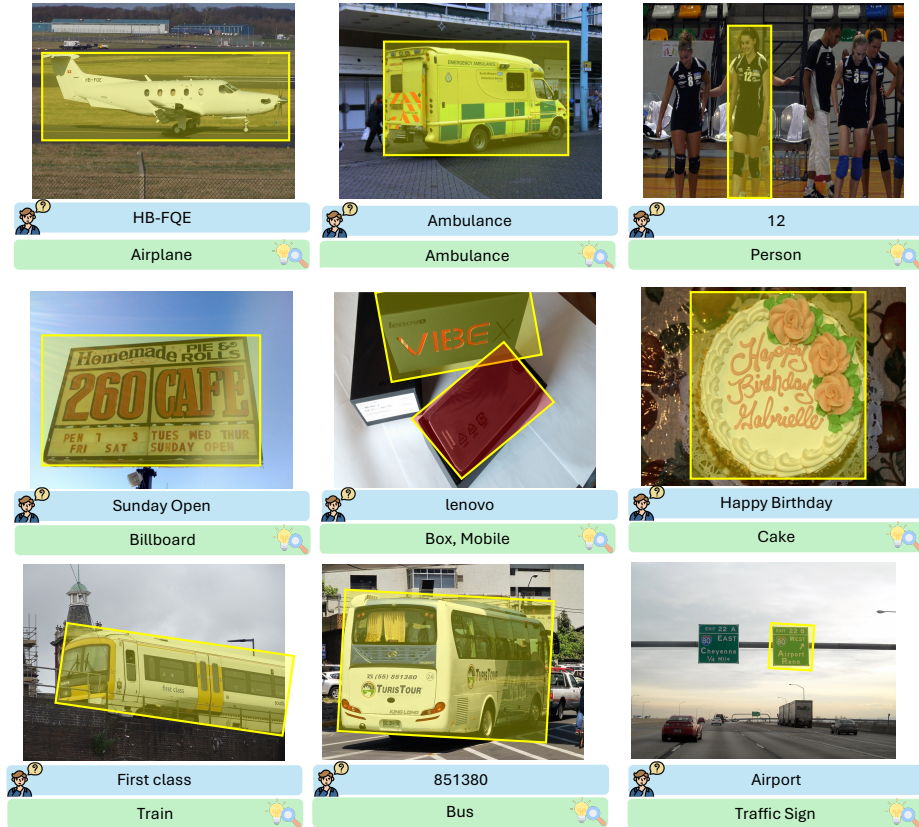


Fig. 18: Example samples for **Text-Object Grounding** task in READ THE OBJECT benchmark

B.6 Text-Region Grounding

The Text-Region Grounding task measures a model’s ability to localize every spatial occurrence of a given text string within an image. Given a query such as

“New York,” the model must output all text-region bounding boxes where the phrase appears, regardless of the associated object category. Unlike Text–Object Grounding, which maps text to the objects bearing it, this task focuses purely on detecting the spatial positions of the text itself.

This task is essential for applications requiring fine-grained scene-text understanding, such as document-like scene parsing, signboard reading, or extracting repeated text patterns. Since each text region is explicitly annotated via OCR, these grounding labels can be obtained directly from our aggregated OCR annotations.

The final Text–Region Grounding subset contains **137,246** text instances across the dataset. Examples of our text-region annotations are shown in Figure 19.

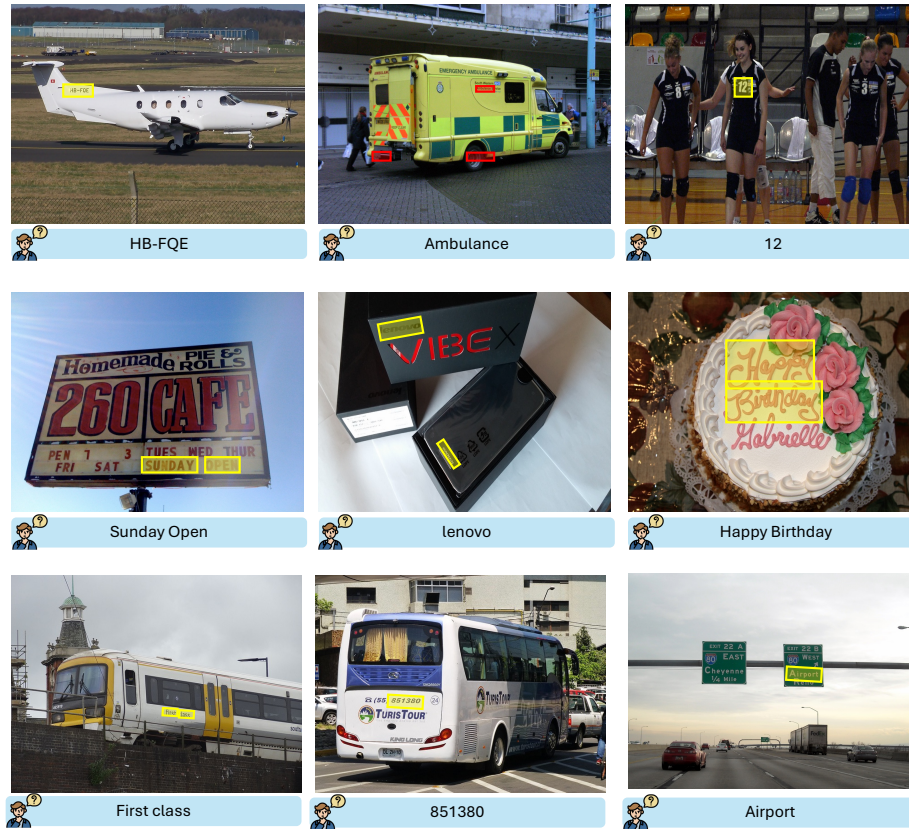


Fig.19: Qualitative results on text-region grounding.Example samples for Text-Region Grounding task in READ THE OBJECT benchmark

Table 9: Statistics of training datasets used for Florence-2 LoRA fine-tuning.

Dataset Name	Number of Samples	Number of Images
MsCOCO	51,241	43,186
OpenImagesV5	1,049,309	215,004
SynthText	932,856	78,085
Total	2,033,406	336,275

C Training Details

To assess whether simple fine-tuning on a strong vision-language model can improve performance on our benchmark, we apply LoRA fine-tuning to Florence-2. Despite extensive training on our curated dataset, the fine-tuned model still does not perform well on our benchmark (see main paper), demonstrating the difficulty of text-grounded object understanding.

C.1 Training Dataset Curation

For training, we use images from COCO, OpenImages V5, and SynthText. After constructing the benchmark splits and balancing categories, we use all remaining images from these datasets as our training pool. This results in a large-scale training set designed to expose the model to diverse text-object interactions.

SynthText provides synthetic text over natural scene images. To align it with our task, we retain only the samples where the synthetic text lies fully inside an object bounding box. Additionally, we generate new synthetic instances by pasting the synthesized text directly onto the object region, increasing the variety of text placements. These machine-generated samples aim to improve a model’s ability to read text in natural contexts. After filtering and preprocessing, the full training set contains: **2,033,406** total training samples, and **336,275** unique images. Details in Table 9.

Following the same procedure as for the benchmark (see subsection A.2), we generate fine-grained attributes for objects when needed. Due to the size of the dataset, manual verification is infeasible; instead, we rely on agreement across multiple multimodal LLMs to validate attributes automatically.

Balancing Multi-Instance Cases. In the raw training data, fewer than 10% of samples contain multiple object instances sharing the same text. Because our benchmark includes balanced multi-object scenarios, we augment training data using a crop-and-paste strategy [13]. We duplicate object instances and place them at different spatial locations within the image, assigning positional cues such as “left person” and “right person” to support multi-instance grounding.

Overall, this large-scale curated dataset enables robust fine-tuning; however, as shown in the main paper, even a strong model like Florence-2 struggles on our benchmark, highlighting the challenge of fine-grained, text-grounded object understanding.

Table 10: Training hyperparameters used for fine-tuning.

Category	Hyperparameter	Value
Training	Epochs	10
	Batch Size	32
	Max Sequence Length	1024
Optimization	Optimizer	AdamW
	Learning Rate	5×10^{-6}
	LR Scheduler	Linear
LoRA Configuration	Rank (r)	16
	Alpha	16
	Dropout	0.05

C.2 Training Hyperparameters

This section details the hyperparameters employed for Florence-2 LoRA fine-tuning in Table 10.

D More Results

D.1 Measuring Text Localization Under Referring Expressions

Although none of these models were actually trained for the kind of fine-grained text grounding we probe, we still evaluated their behavior by giving each model a referring expression and prompting it to output a bounding box for a specific piece of text—for example, asking for the location of the substring “a” in “boat labeled ‘a’.” The results are summarized in Table 11. Surprisingly, even models like Florence-2, which were trained for OCR and grounding, struggle to localize the correct text region reliably. In contrast, DeepSeek-OCR consistently finds the text more accurately, but only when the prompt is written in a model-specific, very careful way; moreover, it tends to crop all visible text rather than the particular instance referenced in the query. This behavior highlights how brittle current MLLMs are when asked to perform targeted text grounding, even when they contain OCR components.

D.2 Qualitative Results on Caption

In Figure 20, it can be seen our scene-text-aware captions consistently provide accurate, text-grounded descriptions that correctly associate each piece of text with the specific object on which it appears. Unlike existing multimodal LLMs (DeepSeek-OCR, Florence-2, InternVL3-78B), which frequently misread text, omit key labels, or assign them to the wrong object instance, our annotations explicitly fuse object attributes, text attributes, and segmentation-based grounding to ensure correctness. This yields captions that faithfully capture all visible text

Table 11: Text-level detection performance across models.

Model	Avg mAP	mAP@0.50	mAP@0.75	mAR@1
MM-Grounding DINO	1.0e-4	7.0e-4	7.0e-4	5.7e-3
Florence-2	2.0e-3	6.5e-3	4.0e-3	2.0e-4
llava-next 1.6 7B	1.0e-3	9.9e-3	9.9e-3	2.0e-4
InternVL3-1B	1.0e-3	1.0e-3	0	1.0e-4
Qwen3-VL-2B-Instruct	5.9e-3	9.9e-3	9.9e-3	1.8e-3
Qwen/Qwen3-VL-8B-Instruct	6.93e-3	9.9e-3	9.9e-3	2.5e-3
Qwen3-VL-30B-Instruct	3.96e-3	9.9e-3	0	5.0e-4
grounding DINO	0	0	0	0
LLaVA-OneVision-7B (SI)	9.9e-4	9.9e-4	0	1.0e-4
DeepSeek-OCR	4.4e-1	5.7e-1	4.3e-1	2.9e-1

(e.g., “Fizik,” “M1,” “Gabriel,” “ZS-LPK”) and precisely link them to the correct shoe or airplane, avoiding the common misalignment and confusion exhibited by current models. Our captions further excel by providing fine-grained, instance-level distinctions—such as color, spatial position, and relative layout—that disambiguate between multiple similar objects sharing identical or nearly identical text. Baseline MLLMs often hallucinate narrative details, invent contextual stories, or add attributes not present in the image, whereas our captions remain concise, factual, and grounded strictly in the visual evidence. Through multi-LLM agreement, mask-guided text-object linking, and subsequent human verification, our annotations capture both the semantic content of scene-text and the subtle visual differences across object instances, producing the most reliable and text-faithful captions for real-world images.



Two black cycling shoes labeled "Fizik" are shown in the image. The top shoe rests on its original box, which sits on a wooden table and displays additional labels, including "M1" and "MTB." The second shoe appears lower in the frame, placed directly on the table. Behind the shoes, several papers dated 2012 are partially visible, adding context to the scene and suggesting the items may be part of a product display or documentation set.

Ground Truth



Three planes are flying in the sky. The two aircraft on the left are white, and the one on the right is red. All three display the word "Gabriel." The leftmost plane also shows the model label "ZS-LPK" on its wings, while the middle white plane carries the label "ZS-LRW." The red plane on the right is marked "ZS-00B." A smoke trail follows behind them.

Ground Truth

This image displays a pair of black Nike Air Max 90 shoes with black and grey laces and a black sole. The shoes are sitting on a white box with a black Nike logo on the side. The box is sitting on a wooden table. There is a yellow and black box behind the shoes on the table. There is a white wall in the background. The image is a photograph.

DeepSeek-OCR

Four biplanes flying closely together against a clear blue sky backdrop. The planes are painted predominantly white with red accents on both wings and tail fins; each plane has "Ga" written prominently near the front of the fuselage. They emit trails of smoke from behind them as they fly side by side at an angle that suggests formation flight or aerobatic maneuvering. Each aircraft is positioned slightly above another to create depth within the composition. Their design includes two main wings stacked vertically atop short empennage structures adorned with additional markings such as numbers ("12," "13") indicating possibly racing identification codes. This arrangement creates dynamic lines leading towards the horizon where more planes can faintly be seen continuing into the distance.

DeepSeek-OCR

The image shows a pair of black cycling shoes on top of a black cardboard box. The shoes have a velcro strap on the top and a buckle on the side. The box has the brand name "Fizik" written on it in white letters. The background appears to be a store with other items on display.

Florence-2

The image shows three airplanes flying in formation in the sky. The airplanes are red and white in color with the word "Cadillac" written on the side. They are flying in a V-formation, with the first airplane on the left, the second one in the middle, and the third one on the right. The sky is clear and blue, and there is a trail of white smoke trailing behind the airplanes. The aircrafts appear to be in mid-flight, with their wings spread wide and their engines roaring. The image appears to be taken from a low angle, looking up at the airplanes as they soar through the air.

Florence-2

A pair of black Fizik M1 mountain bike shoes displayed on top of their box, featuring reflective straps and a ratchet buckle.

InternVL3-78B

The image features three biplanes flying in formation against a clear blue sky, leaving trails of white smoke behind them. The planes are painted in a striking color scheme with red and white accents, and have the word "Gabriel" prominently displayed on their wings and fuselage. The planes are flying close together, suggesting a coordinated aerial display or performance. The image captures the dynamic motion and precision of the flight, showcasing the skill of the pilots.

InternVL3-78B

Fig. 20: Qualitative results on captioning. Example samples for **Scene-Text-Aware Caption** task in **READ THE OBJECT** benchmark.