Detect Anything via Next Point Prediction

Qing Jiang[1,2‡], Junan Huo[1,2], Xinyu Chen[2,3,4], Yuda Xiong[1,2], Zhaoyang Zeng[1,2]
Yihao Chen[1,2], Tianhe Ren[1,2], Junzhi Yu[3], Lei Zhang[1,2†]
[1] South China University of Technology
[2] International Digital Economy Academy (IDEA)
[3] Peking University [4] Zhongguancun Academy
https://rex-omni.github.io/

Abstract

Object detection has long been dominated by coordinate regression-based models. Although recent efforts have attempted to leverage MLLMs to tackle this task, they face challenges like duplicate prediction and coordinate misalignment. We bridge this gap and propose ***Rex-Omni****, a 3B-scale MLLM that achieves SOTA object perception performance. On benchmarks like COCO, Rex-Omni attains performance comparable to or exceeding regression-based models (e.g. DINO, Grounding DINO). This is enabled by three key designs: 1)* ***Task Formulation:*** *We use learnable quantized coordinate tokens to represent predicted coordinates, simplifying learning and enhancing token efficiency.* ***2) Data Engines:*** *We construct specialized data engines to generate over 13 million high-quality, semantically rich annotations for grounding, referring, pointing and OCR tasks;* ***3) Training Pipelines****: we employ a two-stage training process: large-scale SFT on 22 million data, followed by GRPO-based RL post-training, which uses geometry-aware rewards to both mitigate SFT-induced behavioral deficiencies and improve coordinate accuracy. Beyond standard detection, Rex-Omni's inherent language understanding enables a wide range of versatile capabilities, including object referring, pointing, GUI grounding, OCR, and more. We believe that Rex-Omni can pave the way for more versatile and language-aware visual perception systems. Code is available at* https://github.com/IDEA-Research/Rex-Omni.

1. Introduction

Object detection [4, 15, 40, 58, 60, 66, 69, 81, 83] has long been a foundational task in Computer Vision for its broad applications. The field has progressed from early CNN-based architectures, such as YOLO [59] and Faster R-CNN [60], to Transformer-based models like DETR [4] and DINO [81], while the task itself has evolved from traditional closed-set detection to open-set detection [9, 20, 32, 39, 48, 48] to better handle emerging real-world challenges.

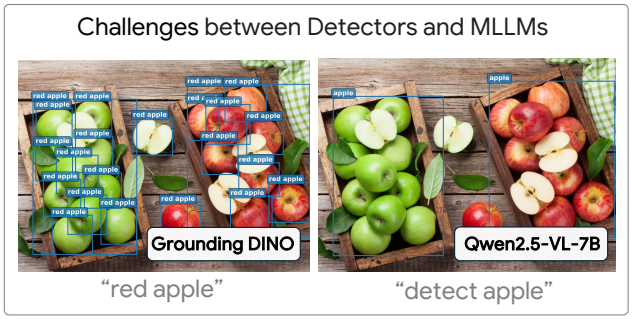


Figure 1. Detectors excel in localization but are weak in language understanding, while MLLMs are strong in language but struggle with precise localization.

A key objective in object detection is to identify any object. A prevalent approach is open-vocabulary detection, where models like Grounding DINO [39] leverage text encoders (e.g. BERT [26], CLIP [56]) to perform category-level open-set detection. However, these methods are constrained by a relatively shallow language understanding, which limits their ability to handle complex semantics. As shown in Figure 1, Grounding DINO fails to distinguish "red apple" from all apples, highlighting an inherent limitation in their language comprehension. In contrast, MLLM [1, 7, 12, 30, 44, 49, 68, 70, 72] benefit from the strong language understanding capabilities of LLMs, presenting a promising path for integrating advanced language comprehension into object detection. A common approach [2, 5, 16, 21, 24, 46, 71, 77, 79, 82, 85] is to represent coordinates as discrete tokens [6] and predict coordinates through next-token prediction. While conceptually elegant, existing methods have rarely matched the performance of traditional regression-based detectors. As shown in Figure

This work was done during Qing, Junan's internship and Xinyu's academic visit at IDEA. † Corresponding author. ‡Project Lead.

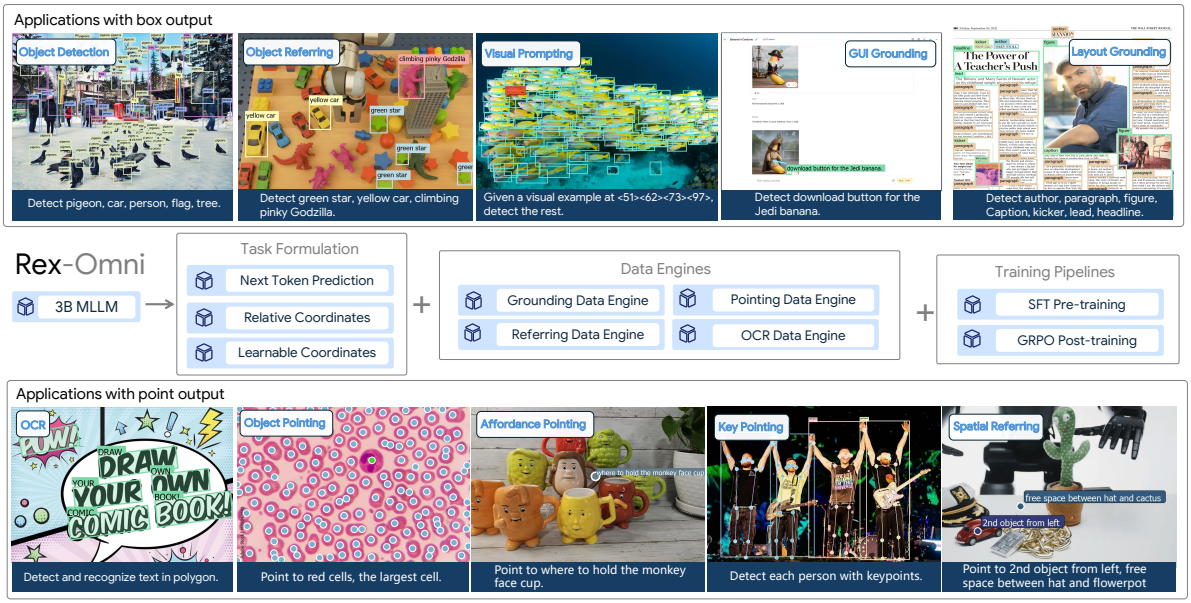


Figure 2. We introduce Rex-Omni, a 3B-parameter MLLM with strong visual perception capabilities.

1, even advanced MLLMs like Qwen2.5-VL [2] struggles with precise object localization.

We argue this performance disparity arises from two core challenges. Firstly, MLLMs treat coordinate prediction as a discrete classification task with cross entropy loss as supervision. This contrasts with traditional regression-based models that use geometry-aware losses (e.g., L1, GIoU) sensitive to small geometric offsets. This classification approach creates a significant learning difficulty, where even minor pixel misalignments can lead to disproportionately large losses, hindering precise localization. This necessitates strategies to ease coordinate learning complexity and demands extensive data to learn the token-to-pixel mapping.

Secondly, MLLMs commonly employ Supervised Fine-tuning (SFT) for teacher-guided training [54]. While efficient, this creates a mismatch between training and inference, as the model is never exposed to its own, potentially imperfect predictions. This lack of exposure to autonomous generation scenarios prevents the model from developing robust behavioral awareness. Consequently, during inference without direct guidance, it often struggles to regulate its output structure, leading to anomalous behaviors like duplicate predictions or object omissions. Addressing these two intertwined challenges is crucial for advancing MLLM-based object detection.

To overcome these limitations, we propose Rex-Omni, a 3B-scale MLLM that achieves competitive performance with traditional detectors while excelling in language understanding. We address the aforementioned challenges through three core design principles:

- **Task Formulation**: We unify visual perception tasks under a coordinate prediction framework by generating coordinate sequences. We adopt a quantized coordinate representation, mapping each coordinate value to one of 1,000 learnable discrete tokens.
- **Data Engines**: To facilitate learning the token-to-pixel mapping and foster robust language comprehension, we design specialized data engines for grounding, referring, pointing and OCR tasks. These engines generate semantically rich data essential for coordinate prediction.
- **Training Pipelines**: We adopt a two-stage training paradigm. Firstly, we perform SFT on 22 million data to teach basic coordinate prediction skills. Secondly, we apply GRPO-based [64] RL post-training with geometry-aware rewards to enhances coordinate precision and mitigates undesirable behaviors (e.g., duplicate predictions) that arise from the teacher-guided nature of SFT.

After this two-stage training, Rex-Omni achieves superior performance across a diverse range of perception tasks, all through direct coordinate prediction (see Figure 2). To quantitatively assess its performance, we first evaluate Rex-Omni on the COCO benchmark [34]. In a zero-shot setting, Rex-Omni surpasses traditional regression-based models (e.g., DINO [81], Grounding DINO [39]) and other MLLMs (e.g., SEED1.5-VL [16]) in F1-score. We then evaluate it across eight different perception tasks and this strong performance also extends to other benchmarks.

Rex-Omni consistently outperforms both traditional detectors and other MLLMs, establishing a unified framework that effectively combines precise localization with robust language understanding. We demonstrate that MLLMs have the profound potential to define the next generation of object detection models, offering unprecedented versatility and a truly language-aware approach to visual perception.

2. Related Work

Regresson-based Object Detection Methods. Object detection has long been dominated by regression-based methods, which predict bounding box properties (e.g. center coordinates and dimensions). This field has evolved significantly, from early anchor-based CNN models like YOLO [59] and Faster R-CNN [60], to anchor-free approaches such as FCOS [69]. A major paradigm shift occurred with Transformer-based detectors like DETR [4], which reframed detection as a direct set prediction problem, with subsequent models like DINO [81] enhancing performance. The continuous improvement of these detectors has also been fueled by crucial innovations in architecture (e.g., FPN [35]), loss functions (e.g., Focal Loss [36]).

Open-set Object Detection Methods. The goal of open-set object detection is to identify arbitrary object categories without task-specific fine-tuning. The prevalent approach is text-prompted open-vocabulary detection [9, 14, 25, 32, 39, 61], which leverages vision-language models like CLIP [55] to align text descriptions with visual regions, achieving impressive zero-shot recognition. To handle rare or hard-to-describe objects, visual prompts [19, 23, 27] have also been introduced, using visual examples (e.g., boxes or points) for recognition. Recent models like T-Rex2 [23] also combine both text and visual prompts to enhance performance. However, as discussed previously, these open-set detectors are fundamentally constrained by the relatively shallow language understanding, which limits their ability to interpret complex, context-rich descriptions.

MLLM-based Object Detection Methods. To overcome the shallow language understanding of traditional open-set detectors, a promising direction is to leverage Multimodal Large Language Models (MLLMs) for object-level perception. The core idea is to reframe object detection as a language modeling task by representing bounding box coordinates as a sequence of discrete, quantized tokens [2, 5, 6, 16, 51, 71, 77–79]. While this elegantly unifies detection with the native capabilities of LLMs, existing methods often struggle with fine-grained spatial precision, suffering from low recall, coordinate drift, and duplicate predictions.

3. Task Formulation

In this section, we present Rex-Omni's task formulation design, covering its coordinate representation, the specific output formats for different tasks, and its model architecture.

3.1. Coordinate Representation

MLLM-based approaches for coordinate prediction can be categorized into three paradigms: **1) Direct Coordinate Prediction**, where coordinates are treated as discrete tokens and directly predicted [5, 71, 77, 79, 82]; **2) Retrieval-based Methods**, where the LLM predicts the index of a predefined proposal [21, 22, 24, 46]; and **3) External Decoders**, where the LLM's outputs are passed to an external module for final coordinate generation [29, 42, 73, 80]. We adopt the direct coordinate prediction strategy for Rex-Omni, motivated by its simplicity and end-to-end nature.

Within the direct prediction paradigm, variations include: 1) Relative coordinates with special tokens (e.g. Pix2Seq [6]), where each quantized coordinate (0-999) is a single special token; 2) Relative coordinates without special tokens (e.g. SEED1.5-VL [16]), where each quantized coordinate is represented by multiple digit tokens; and 3) Absolute coordinates (e.g. Qwen2.5-VL [2]), where coordinates are directly tokenized into individual digits without quantilization. We choose the first approach for two primary reasons: 1) Relative coordinates reduce learning complexity by confining the task to a 1,000-class classification problem. 2) Special tokens are highly token-efficient, representing a bounding box with only four tokens versus 15+ tokens in other schemes.

3.2. Input Format

Rex-Omni adopts a unified text-based interface for all visual perception tasks. Each task is expressed as a natural language query that specifies the target objects to be identified in the image.

Text Prompts. For most tasks, the model receives an image paired with a text prompt. The prompt can describe single or multiple targets, with multiple targets concatenated by commas. For instance, a multi-object detection query can be formatted as: *"Please detect pigeon, person, truck, snow in this image. Return the output in box format."* We design distinct query templates to guide the model for different perception tasks.

Visual Prompts. While text prompts are highly generalizable, they struggle with objects that are rare or difficult to describe linguistically [20]. To address this, Rex-Omni also supports visual prompting, allowing users to provide bounding boxes as input. Unlike existing methods [19, 23, 61] that treat visual prompting as a feature-matching problem, Rex-Omni maintains its unified text-based interface. A visual prompt is first converted into quantized coordinate tokens. The model is then guided by natural language instructions to identify all objects of the same category. For example: *"Given visual prompt {object1: [<12><412><339><568>]}, detect the rest."*

3.3. Output Format for Each Task

The output for each visual task is uniformly represented as a structured token sequence, organized as:

`<|object_ref_start|>PHRASE<|object_ref_end|><|box_start|>COORDS<|box_end|>`

Here, PHRASE is the object category or description, and COORDS is the coordinate sequence. We retain the original special

tokens from our base model, Qwen2.5-VL-3B, for this demarcation. The COORDS format varies by task. For tasks involving boxes (e.g., object detection), COORDS is a sequence of [x0, y0, x1, y1] coordinates, such as <12><42><512><612>. For tasks involving points (e.g., object pointing), it is a sequence of [x0, y0] pairs, like <100><150>. For polygons (e.g., OCR), it is a sequence of vertices, e.g., <10><20>.... For multi-phrase detection, outputs are concatenated with commas.

3.4. Model Architecture

Rex-Omni is built upon the Qwen2.5-VL-3B [2] model with minimal architectural modifications. Specifically, we repurpose the final 1,000 vocabulary tokens in Qwen2.5-VL to serve as dedicated special tokens, each representing a quantized coordinate value.

4. Training Data

To equip Rex-Omni with both precise coordinate prediction capabilities and strong language understanding, we utilize two sources of training data: publicly available datasets and automatically annotated data generated by our custom-designed data engines.

4.1. Public Datasets

We leverage a diverse range of publicly available datasets to train Rex-Omni across numerous subtasks. For each task, we defined a set of question templates to construct corresponding question-answer (QA) pairs. In total, approximately 8.9 million public data samples were utilized. Details are available in the Appendix.

4.2. Data Engines

Training Rex-Omni to learn a fine-grained mapping between its 1,000 quantized coordinate tokens and the continuous pixel space requires a massive volume of high-quality data, exceeding what is available in public datasets. Furthermore, existing datasets often lack the instance-level semantic richness (e.g., referring expressions) needed for robust language grounding. To address these limitations, we developed a dedicated suite of data engines to generate large-scale, high-quality training data tailored for fine-grained spatial reasoning and language grounding.

4.2.1. Grounding Data Engine

Our Grounding Data Engine follows a common strategy [9, 20, 52, 61, 62] of generating image captions, extracting noun phrases, and using a grounding model to assign bounding boxes. However, distinguishing from prior works, we introduce a crucial Phrase Filtering stage to enhance annotation quality. Our four-stage pipeline is as follows: **1) Image Captioning:** Generate descriptive captions using Qwen2.5-VL-7B given an input image. **2) Phrase Extraction:** Extract noun phrases (e.g. lemon, green lemon) using NLP tool SpaCy. **3) Phrase Filtering:** To minimize ambiguity, we remove phrases with descriptive attributes, retaining only base class names (e.g. lemon is kept, green lemon is discarded). This is motivated by the limited language understanding of current grounding models. For instance, when prompted with "green lemon", these models often detect all lemons in the image, regardless of their color, which introduces significant labeling errors. **4) Phrase Grounding:** Use an open-vocabulary detector DINO-X [61] to assign bounding boxes to the filtered phrases. This engine processes images from COYO [3] and SA-1B [27] after rigorous preprocessing, yielding a curated dataset of approximately 3 million images with high-quality grounding annotations.

4.2.2. Referring Data Engine

To generate semantically rich referring data (e.g., "a man in a yellow shirt") at scale, we designed a fully automated Referring Data Engine, overcoming the scalability limitations of manual annotation methods like HumanRef [24]. Our four-stage pipeline is as follows: **1) Expression Generation:** We prompt Qwen2.5-VL-7B with images and category labels to generate natural, human-like referring expressions. **2) Pointing:** For each expression, Molmo [12] is used to predict a corresponding spatial point. **3) Mask Generation:** SAM [27] generates a mask for each ground-truth bounding box. **4) Point-to-Box Association:** Each predicted point is aligned with a SAM-generated mask: if a point lies within a mask, the corresponding bounding box is linked to the referring expression. This engine processes images from O365 [63], OpenImages [28], COYO and SA-1B, yielding approximately 3 million images with automatically generated referring annotations.

4.2.3. Other Data Engines

Beyond grounding and referring data, we developed two additional lightweight data engines: **1) Pointing Data Engine:** To generate point-level supervision, we adopt a geometry-aware strategy that converts existing bounding box annotations into precise point annotations. Given a box, we first obtain its mask via SAM [27], then compute a candidate point from the minimum-area enclosing rotated rectangle of the mask. This process yields approximately 5 million point-level samples. **2) OCR Data Engine:** We utilize PaddleOCR to annotate images from the COYO dataset [3]. This engine extracts both polygonal boundaries and text transcriptions, and also computes corresponding bounding boxes, resulting in approximately 2 million OCR-annotated samples

In total, combining publicly available datasets with data generated from our data engines, we amassed a training set of 22 million high-quality annotated images.

5. Training Pipelines

We employ a two-stage training strategy. In the first stage, Supervised Fine-Tuning is performed on 22 million annotated samples using a teacher-guided training approach, enabling the model to acquire fundamental coordinate prediction capabilities. In the second stage, we apply reinforcement learning based on the GRPO framework, which further refines the model's performance by combining geometry-aware rewards with behavior-aware optimization, thus addressing the limitations of the SFT stage and enhancing overall prediction quality.

5.1. Stage1: Supervised Fine-Tuning

Since the model predicts coordinates in the form of quantized tokens ranging from 0 to 999, it must first learn how to accurately map these discrete values back to continuous pixel locations within the image. This corresponds to a 1,000 way classification problem over spatial positions, which requires substantial supervision to achieve reliable performance. Therefore, we begin training with a teacher-guided supervised fine-tuning stage on large-scale anno-

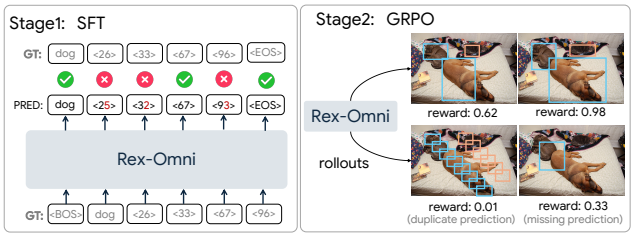


Figure 3. Our two-stage training pipeline: A SFT stage teaches basic coordinate prediction, followed by a GRPO-based RL stage that corrects behavioral issues using geometry-aware rewards.

tated data, enabling the model to acquire the fundamental ability to interpret and predict spatial coordinates.

5.2. Stage2: Reinforcement Post-Training

5.2.1. Limitations of SFT

While SFT allows the model to quickly acquire basic coordinate prediction capabilities by leveraging massive amounts of labeled data, it presents two key limitations:

Geometric Discretization Issue. Using CE loss for coordinate prediction introduces a discretization problem. Coordinates are represented as vocabulary tokens (from <0> to <999>), and the model is trained to classify each token exactly. However, this formulation is misaligned with the continuous nature of geometry in spatial tasks. For example, if the ground-truth token is <33> but the model predicts <32>, the difference in pixel space may be negligible, yet the CE loss penalizes it as a completely incorrect prediction. Conversely, if the ground truth is <0><0><100><100> but the model predicts <0><0><100><1000>, only one token is misclassified. In this case, the CE loss remains relatively small, even though the resulting bounding box is severely misaligned.

Behavioral Regulation Deficiency. In the SFT stage, teacher-forced training relies on full ground-truth sequences for efficient parallel learning. This setup fixes the number of predicted boxes to the ground-truth count, preventing the model from autonomously learning how many objects to predict. Consequently, during inference the model often fails to regulate output quantity, leading to two typical errors: *1)* predicting fewer boxes than required (missed detections), or *2)* predicting more boxes than necessary (repetitive detections with identical or slightly shifted coordinates). These behaviors reflect the model's lack of effective output regulation.

5.2.2. GRPO-based Post-Training

To address the geometric and behavioral limitations of SFT, we adopt a reinforcement post-training strategy based on GRPO [64]. GRPO enables the model to explore its output space and improve through reward-guided optimization. Given an input image and prompt (I, x), the model samples G complete responses $\{o_1, \ldots, o_G\}$ from its policy π_θ. For each output o_i, we compute a scalar reward r_i and normalize it to obtain the relative advantage:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)}. \tag{1}$$

Through this formulation, we can mitigate SFT's limitations by using geometry-aware rewards for spatial alignment and allowing variable-length outputs to penalize repetition.

Geometry-aware Rewards. To provide informative feedback, we design three geometry-aware reward functions.

Box IoU Reward. For tasks requiring bounding boxes (e.g., object detection), this reward encourages both accurate localization and correct category alignment. Given predicted boxes $\hat{B}$ and ground-truth boxes B^*, we perform a ground-truth-guided matching where for each ground-truth box b_j^*, we find its highest IoU-matched predicted box $\hat{b}_i$. If the category labels match, a reward r_j equals to their IoU is assigned; otherwise, $r_j = 0$. The final reward is an F1-style score computed from the recall and precision:

$$r^{\text{IoU}} = \frac{2 \cdot \left(\frac{\sum_{j=1}^{n} r_j}{|\hat{B}|}\right) \cdot \left(\frac{\sum_{j=1}^{n} r_j}{|B^*|}\right)}{\left(\frac{\sum_{j=1}^{n} r_j}{|\hat{B}|}\right) + \left(\frac{\sum_{j=1}^{n} r_j}{|B^*|}\right)} \tag{2}$$

Point-in-Mask Reward. For tasks like object pointing, this reward evaluates if a predicted point lies within the target object's mask. For each ground-truth box, we first generate a segmentation mask using SAM [27]. A reward of 1 is assigned if a predicted point with the correct category falls within this mask; otherwise, the reward is 0. The final reward is also an F1-style score computed from these binary rewards.

Point-in-Box Reward. For GUI grounding, this is a simple binary reward. A reward of 1 is assigned if the predicted point falls within the target UI element's ground-truth bounding box; otherwise, the reward is 0. This encourages precise interaction.

We sample 66K data from the SFT stage to serve as training data for the GRPO stage. More details are listed in Appendix.

6. Experiments

Evaluation Benchmarks. We evaluate Rex-Omni across a diverse range of visual perception tasks to demonstrate its versatility. Our evaluation begins with object detection tasks, including common object detection on COCO [33], long-tailed object detection on LVIS [17], and dense and tiny object detection on VisDrone [13] and our manually collected Dense200 dataset (Detailed in Appendix). We further assess its capabilities in tasks requiring deeper language understanding and fine-grained localization. These include Referring Object Detection on RefCOCOg [47] and HumanRef [24], Visual Prompting on FSC147 [57] and other detection benchmarks, and Object Pointing across a consolidated set of datasets. We also evaluate performance on several specialized grounding tasks: GUI Grounding on ScreenSpot-V2 [75] and ScreenSpot-Pro [31], Layout Grounding on DocLayNet [53] and M6Doc [8] and OCR on a collection of three datasets.

Evaluation Metrics. We primarily use F1-score as the evaluation metric for detection tasks, reporting results at IoU thresholds of 0.5, 0.95, and the mean IoU (mIoU). Since many MLLMs lack reliable confidence scores, this provides a fairer comparison than Average Precision (AP). Specific metrics for other tasks are as follows: *1) Object Pointing*: F1@Point, a F1-like metirc where a true positive is defined as a predicted point falling within the SAM-generated segmentation mask of a ground-truth box. *2) GUI Grounding:* Accuracy, where a prediction is correct if the point falls within the target's ground-truth bounding box. *3) Visual Prompting:* In addition to F1-score, Mean Absolute Error (MAE) is used to evaluate object counting ability.

Type	Method	Zero-Shot	Score Thresh.	Common Object Detection				Long-tailed Object Detection		
				COCO				LVIS		
				mAP	F1@IoU 0.5	F1@IoU 0.95	F1@ mIoU	F1@IoU 0.5	F1@IoU 0.95	F1@ mIoU
Closed-set	Faster RCNN-R50 [60]	No	0.42	38.4	60.6	7.1	48.1	-	-	-
	DETR-R50 [4]	No	0.78	41.5	65.9	13.6	48.3	-	-	-
	DyHead-R50 [11]	No	0.24	45.9	66.1	15.0	51.3	-	-	-
	DAB-DETR-R50 [38]	No	0.31	44.4	67.1	13.4	50.2	-	-	-
	Deformable-DETR-R50 [86]	No	0.34	49.4	69.7	17.7	54.7	-	-	-
	DINO-R50 [81]	No	0.30	51.7	68.8	21.1	55.6	-	-	-
	DINO-Swin-L [81]	No	0.32	**59.6**	**75.6**	**25.4**	**62.1**	-	-	-
Open-set	Grounding DINO-Swin-T [39]	Yes	0.37 / 0.21	51.5	69.8	23.0	56.6	47.7	**22.7**	38.8
MLLM	DeepSeek-VL2-Tiny [74]	UNK	-	-	38.2	8.1	26.3	32.0	11.2	21.2
	OVIS2.5-9B [44]	UNK	-	-	51.6	8.2	34.6	53.1	14.4	35.8
	Mimo-VL-7B [67]	UNK	-	-	56.5	6.7	35.9	49.5	8.8	31.4
	OVIS2.5-2B [44]	UNK	-	-	56.2	10.3	38.7	54.4	15.8	37.4
	DeepSeek-VL2-Small [74]	UNK	-	-	60.9	14.9	45.9	56.2	21.0	41.8
	Qwen2.5-VL-7B [2]	UNK	-	-	64.0	12.6	45.7	57.7	17.6	40.2
	Qwen2.5-VL-3B [2]	UNK	-	-	64.7	15.0	47.6	55.8	19.3	40.3
	SEED1.5-VL [16]	Yes	-	-	71.3	14.3	51.4	**65.6**	19.5	46.7
	Rex-Omni-SFT	Yes	-	-	68.2	15.8	50.4	60.3	20.7	44.2
	Rex-Omni	Yes	-	-	**72.0**	**15.9**	**52.9**	64.3	20.7	**46.9**

Table 1. Evaluation results on the COCO and LVIS bencharmsk. "UNK" signifies that the information was not reported in the papers.

Type	Method	Score Thresh.	Dense Object Detection						Referring Object Detection					
			Dense200			VisDrone			HumanRef			RefCOCOg test		
			F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU
Open-set	Grounding DINO-T	0.25	36.9	**19.7**	33.1	55.2	**3.9**	**38.5**	28.0	16.5	25.2	52.9	20.9	45.9
MLLM	DeepSeek-VL2-T	-	2.2	0.3	1.5	4.3	0.1	1.8	39.1	16.9	31.4	69.3	16.9	52.1
	OVIS2.5-9B	-	14.0	0.0	5.1	15.8	0.1	6.5	73.1	12.4	52.8	88.7	24.2	72.6
	OVIS2.5-2B	-	17.9	0.0	6.7	21.0	0.1	9.2	70.6	12.3	50.0	87.6	30.5	73.8
	MiMo-VL-7B	-	29.7	0.4	15.9	27.7	0.3	14.3	77.6	26.4	63.4	84.6	14.9	65.5
	Qwen2.5-VL-3B	-	0.8	0.1	0.5	31.5	1.9	20.4	66.7	46.8	60.5	83.8	31.8	70.1
	Qwen2.5-VL-7B	-	1.1	0.1	0.6	34.5	1.6	21.7	72.9	42.9	64.1	85.7	28.4	70.4
	DeepSeek-VL2-S	-	16.0	3.9	12.7	35.8	1.7	23.3	72.0	46.5	64.7	**91.8**	**47.0**	**81.6**
	SEED1.5-VL	-	76.9	5.3	53.2	55.9	0.6	27.4	**88.2**	60.0	**81.6**	85.2	32.1	73.2
	Rex-Omni-SFT	-	60.2	10.6	46.4	55.6	1.9	32.4	83.3	64.3	77.9	85.2	35.2	72.4
	Rex-Omni	-	**78.4**	10.3	**58.3**	**61.6**	1.5	35.8	85.4	**65.4**	79.9	86.8	36.6	74.3

Table 2. Evaluation results on dense object detection benchmarks and referring object detection benchmarks.

Evaluation Settings. We evaluate two variants: **Rex-Omni-SFT**, which undergoes only the first stage of supervised fine-tuning, and the full **Rex-Omni** model, which undergoes both SFT and the subsequent GRPO-based reinforcement post-training. We compare them against three baseline types: closed-set detector, open-set detector, and other MLLMs. For closed-set detectors, we input images and retain only the predicted bounding boxes whose categories match the ground-truth (GT) labels in each image. For open-set detectors, we provide all GT categories as text prompts and keep the corresponding results. For MLLMs, we adopt two prompting strategies: (1) querying one GT category at a time (e.g., "Detect dog in this image"), and (2) querying all GT categories simultaneously (e.g., "Detect dog, cat, person in this image"). Although the latter is more practical in real-world scenarios, most MLLMs exhibit a performance drop when handling multiple categories simultaneously. Therefore, except for SEED1.5-VL, Rex-Omni-SFT and Rex-Omni, we use the single-category strategy.

6.1. Evaluation Results

Common and Long-tailed Object Detection. The results are presented in Table 1. On COCO, Rex-Omni surpasses leading traditional open-set and closed-set detectors (e.g., Grounding DINO, DINO-R50) at an IoU of 0.5. Crucially, Rex-Omni achieves this in a zero-shot setting (without training on COCO data), thereby indicating that MLLM-based detection methods can indeed surpass traditional regression-based models in scenarios where precise bounding box localization is not critical. On the long-tailed LVIS benchmark, MLLMs generally exhibit an advantage over traditional open-set detectors, leveraging their stronger linguistic reasoning for better generalization to rare categories. Rex-Omni shows competitive performance, achieving SOTA mIoU performance, which reflects its superior bounding box precision. Across both benchmarks, the consistent improvement from Rex-Omni-SFT to the full Rex-Omni model validates the effectiveness of our GRPO-based post-training.

Dense and Tiny Object Detection. As shown in Table 2, most MLLMs, including our SFT-only variant, struggle with dense and tiny object detection. We identify two critical failure modes in all MLLMs: large-box prediction (a single box covering multiple objects) and structured duplicate predictions (repetitive coordinates). We attribute these issues to the teacher-forced nature of the SFT stage, which prevents the model from learning to autonomously regulate its output structure. While the full Rex-Omni model shows significant improvement, a detailed analysis of how our GRPO-based post-training mitigates these SFT-induced deficiencies is presented in Section 6.2.

Type	Method	Score Thresh.	FSC147-test		COCO		LVIS	
			F1@IoU mIoU	MAE	F1@IoU mIoU	MAE	F1@IoU mIoU	MAE
Counting	BMNet+ [65]	-	-	14.6	-	-	-	-
	CountTR [37]	-	-	12.0	-	-	-	-
	DAVE [50]	-	-	8.7	-	-	-	-
Open-set	T-Rex2 [20]	0.3	**73.3**	10.9	57.8	**4.0**	**58.8**	**3.4**
MLLM	Rex-Omni-SFT	-	64.6	7.8	58.4	11.7	56.0	11.7
	Rex-Omni	-	62.8	**7.0**	**61.3**	4.5	57.0	5.2

Table 3. Evaluation results on the visual prompting task.

Method	COCO	LVIS	Dense200	VisDrone	HumanRef	RefCOCOg test
	F1@Point	F1@Point	F1@Point	F1@Point	F1@Point	F1@Point
OVIS2.5-2B	73.4	52.8	36.4	23.8	72.5	83.1
Qwen2.5-VL-3B	65.9	48.3	4.3	13.9	64.1	77.8
Qwen2.5-VL-7B	61.1	56.5	2.0	14.2	65.1	79.4
OVIS2.5-9B	72.6	61.7	35.0	18.8	62.3	84.5
Molmo-7B-D	77.3	40.3	33.1	29.2	70.0	83.6
SEED1.5-VL	78.2	70.7	72.1	56.7	83.1	84.2
Rex-Omni-SFT	76.0	66.7	72.9	49.5	82.1	83.9
Rex-Omni	**80.5**	**70.8**	**82.5**	**58.9**	**83.8**	**85.1**

Table 4. Evaluation results on the object pointing task.

Output Format	Method	HierText [43]		TotalText [10]		SROIE [18]	
		F1@IoU 0.5	F1@IoU mIoU	F1@IoU 0.5	F1@IoU mIoU	F1@IoU 0.5	F1@IoU mIoU
BBOX	PaddleOCRv5	45.2	**30.5**	40.2	25.7	**77.7**	**58.6**
	SEED1.5-VL	27.1	12.0	35.0	19.5	51.9	28.1
	Rex-Omni-SFT	23.5	13.7	38.1	25.0	46.5	28.6
	Rex-Omni	**45.9**	28.0	**56.6**	**40.6**	72.0	44.8
POLY	PaddleOCRv5	41.5	**26.3**	34.1	18.4	70.5	**50.2**
	Rex-Omni-SFT	**43.2**	26.2	50.3	**25.7**	**73.8**	39.7
	Rex-Omni	40.2	20.2	**52.8**	25.6	60.3	19.2

Table 5. Evaluation results on the OCR task.

Type	Method	Score Thresh.	DocLayNet		M6Doc	
			F1@IoU 0.5	F1@IoU mIoU	F1@IoU 0.5	F1@IoU mIoU
Closed-Set	DocLayout-YOLO [84]	0.3	**91.2**	**81.1**	-	-
MLLM	Qwen2.5-VL-3B	-	17.5	9.1	13.3	8.4
	Qwen2.5-VL-7B	-	25.6	13.4	24.0	15.0
	SEED1.5-VL	-	54.9	28.7	48.0	28.0
	Rex-Omni-SFT	-	85.9	70.7	74.5	54.2
	Rex-Omni		89.5	70.7	**76.3**	**55.6**

Table 6. Evaluation results on the layout grounding task.

Referring Object Detection. Results are shown in Table 2. Open-set detection model notably struggles with this task, as evidenced by Grounding DINO's consistent underperformance across benchmarks. In contrast, MLLMs, leveraging their inherent strong language understanding capabilities, consistently excel at this task. On the challenging HumanRef benchmark, Rex-Omni achieves competitive results, second only to commercial model SEED1.5-VL, demonstrating that its 3B parameters provide sufficient language understanding for effective REC in real-world scenarios.

Visual Prompting and Object Pointing. For the visual prompting task (Table 3), despite simplifying the task into a text-based coordinate prediction problem, Rex-Omni achieves highly competitive performance, comparable to the specialist T-Rex2. In the object pointing task (Table 4), Rex-Omni consistently outperforms other MLLMs across a range of scenarios, including dense object scenes and referring expressions. Together, these strong results across two distinct, fine-grained localization tasks underscore Rex-Omni's strong task generalization capabilities.

OCR, Layout, and GUI Grounding. Rex-Omni also demonstrates strong performance in structured scene understanding. For OCR (Table 5), it achieves competitive results for both bounding box (BBOX) and polygonal (POLY) outputs, significantly outperforming SEED1.5-VL and rivaling the specialized expert model PaddleOCRv5 in several key aspects. For Layout Grounding (Table 6), Rex-Omni substantially outperforms other MLLMs. While a performance gap remains compared to specialized closed-set models, Rex-Omni's inherent open-set capability allows it to generalize to unseen domains and novel layout structures, establishing it as a more versatile solution. This strong performance also extends to GUI Grounding (Table 7), where Rex-Omni achieves SOTA comparable accuracy among models of a similar size.

Method	ScreenSpot-v2	ScreenSpot-Pro
	Avg	Avg
Qwen2.5-VL-3B [2]	80.9	25.9
UI-R1-3B [45]	85.4	17.8
InfiGUI-R1-3B [41]	-	35.7
JEDI-3B [76]	**88.6**	36.1
Rex-Omni-SFT	86.4	32.6
Rex-Omin-GRPO	88.4	**36.8**

Table 7. Evaluation results for GUI Grounding task.

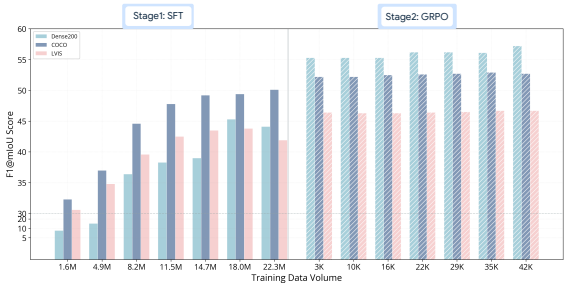


Figure 4. Model performance across different stages.

6.2. In-depth Analysis of Rex-Omni

In this section, we conduct an in-depth analysis to investigate why GRPO-based post-training is crucial for enhancing Rex-Omni's performance. We first show its performance trajectory during the SFT and GRPO stages (Figure 4). During SFT, performance improves steadily and eventually plateauing. In contrast, the GRPO stage triggers a rapid performance jump with only a small number of training data. This suggests that GRPO's primary role is not simply continued learning, but rather unlocking the strong latent capabilities already present in the SFT-trained model. GRPO achieves this by reshaping model behavior through sequence-level, reward-guided feedback. In the following sections, we delve deeper into the specific mechanisms behind this improvement.

6.2.1. Behavioral Correction via GRPO

A key benefit of GRPO is its ability to correct behavioral deficiencies learned during SFT. We analyze two major error patterns:

Duplicate Predictions. SFT's teacher-forced training prevents the model from learning to avoid repetitive outputs. In contrast, GRPO requires autonomous generation and penalizes repeated coordinates with low rewards. To verify this, we analyzed and removed duplicate predictions from both SFT and GRPO model outputs. As shown in Table 8, the SFT-only model's performance improved significantly after duplicate removal (e.g., +15.3% on VisDrone), whereas the GRPO model showed minimal gains. This indicates that GRPO effectively suppresses duplicate predictions, a phenomenon visualized in Figure 5a.

Remove Duplicate	COCO				VisDrone			
	SFT		GRPO		SFT		GRPO	
	F1@0.5	Removv.	F1@0.5	Removv.	F1@0.5	Removv.	F1@0.5	Removv.
No	68.2	-	72.0	-	55.6	-	61.6	-
Yes	**70.1**	1.23%	**72.6**	0.08%	**62.3**	15.3%	**62.1**	0.1%

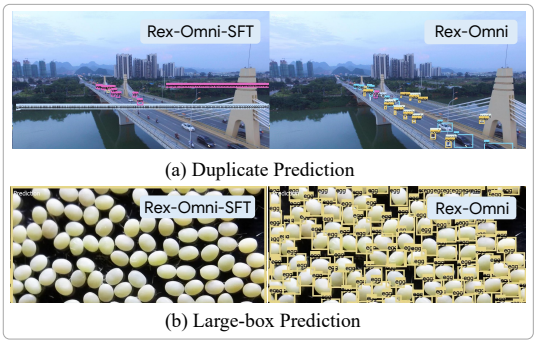
Table 8. Performance comparison of SFT and GRPO models before and after removing duplicate predictions. This highlights GRPO's effectiveness in mitigating repetitive outputs.

Remove Large box	Dense200							
	SFT				GRPO			
	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU	Removv.	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU	Removv.
No	59.1	8.8	44.9	-	78.4	10.3	58.3	-
Yes	74.6	11.2	56.7	20.5%	81.8	9.4	60.0	3.5%

Table 9. Impact of large box prediction removal on F1-score for SFT and GRPO models on the Dense200 dataset.

Stage	COCO		LVIS		HumanRef	
	F1@0.5	F1@mIoU	F1@0.5	F1@mIoU	F1@0.5	RF1@mIoU
SFT	80.5	63.0	74.2	56.6	85.2	60.0
GRPO	81.1	63.5	75.0	56.9	86.4	61.2

Table 10. Impact of GRPO on coordinate precision. The table reports F1-scores for instances where both models achieve consistent ground-truth matching.



(a) Duplicate Prediction

(b) Large-box Prediction

Figure 5. Visualization of the two failure modes.

Large-box Predictions. In dense scenes, SFT-trained models often predict a single large box encompassing multiple objects. We conducted an experiment on the Dense200 dataset, defining a "large box" as a single prediction covering more than 95% area of the image. As shown in Table 9, 20.5% of the SFT model's predictions were large boxes, and removing them significantly boosted its performance. In contrast, only 3.5% of the GRPO model's predictions were large boxes. This confirms that GRPO's behavior-aware optimization discourages such overly large predictions, a failure mode visualized in Figure 5b.

6.2.2. Improvement in Coordinate Precision

We then validate whether GRPO's geometry-aware rewards refine coordinate precision over SFT's cross-entropy loss. To do this, we isolate the analysis to focus purely on coordinate precision. We filter for only "perfectly matched" predictions, which are instances where both SFT and GRPO models predicted the same number of boxes, each with a high IoU against its ground-truth. As shown in Table 10, GRPO yields only modest gains over SFT in this controlled setting (e.g., F1@mIoU increases by just +0.5 on COCO and +0.3 on LVIS). This suggests that SFT alone is largely sufficient for learning accurate coordinates. Therefore, we conclude that GRPO's primary advantage lies not in significantly boosting raw coordinate precision, but in correcting the behavioral deficiencies discussed previously.

Method	COCO		
	F1@IoU 0.5	F1@IoU 0.95	F1@IoU mIoU
SFT	68.2	15.8	50.4
GRPO	72.0	15.9	52.9
SFT-Sampling	72.6	16.8	54.0

Table 11. Impact of GRPO on the likelihood of sampling correct predictions. This table compares the F1-scores (IoU=0.5, IoU=0.95, mIoU) of SFT, GRPO, SFT-Sampling on COCO, dataset. It illustrates GRPO's role in enhancing the probability and inherent quality of correct outputs, and explores SFT's potential under various sampling strategies.

6.2.3. Elevating the Likelihood of Correct Predictions

Finally, we analyze if GRPO's benefit comes from elevating the sampling probability of correct predictions. We investigate the idea that while SFT models can generate accurate coordinates, the likelihood of sampling these optimal outputs is low, and GRPO's reward-guided exploration aims to increase this probability. To validate this on COCO, we conducted high-temperature sampling (temperature 1.2, top-k 50, top-p 0.99) on the SFT model, generating 8 candidate predictions for each test instance. From these, we derived SFT-Sampling: for each test sample, we select the best prediction (highest F1-score) from its 8 candidates, then aggregate these sample-wise best predictions for the overall score. This estimates SFT's maximal performance if optimal predictions could be reliably selected at the sample level. As shown in Table 11, the SFT-Sampling score (72.6 F1@0.5) notably surpasses both the base SFT (68.2) and even the GRPO model (72.0). This result compellingly demonstrates that the SFT model indeed possesses the latent capability to generate high-quality predictions. It further suggests that GRPO's primary contribution is to significantly increase the likelihood of sampling these correct outputs through its refined policy, effectively making the model's strong potential more consistently accessible.

7. Conclusion

In this work, we introduced Rex-Omni, a 3B MLLM that bridges the gap between precise localization and deep language understanding. This is achieved through three key designs: efficient task formulation, large-scale data generation, and a SFT with GRPO two-stage training pipeline. Extensive experiments show Rex-Omni achieves state-of-the-art or competitive zero-shot performance across a wide array of perception tasks. Crucially, our analysis confirms that GRPO post-training is essential for correcting SFT-induced behavioral deficiencies, a key contribution towards robust MLLM detectors. In conclusion, Rex-Omni demonstrates that the behavioral and geometric limitations of MLLMs can be systematically overcome, paving the way for the next generation of versatile, language-aware perception systems.

Acknowledgment

This work was partly supported by the National Natural Science Foundation of China under Grant 62403012.

References

[1] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 3, 4, 6, 7

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 4

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3, 6

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 3

[6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1, 3

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2022. 1

[8] Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15138–15147, 2023. 5

[9] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 1, 3, 4

[10] Chee Kheng Chng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition, 2017. 7

[11] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, pages 7373–7382, 2021. 6

[12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 4

[13] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 5

[14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 3

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[16] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1, 2, 3, 6

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 5

[18] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 7

[19] Qing Jiang, Feng Li, Tianhe Ren, Shilong Liu, Zhaoyang Zeng, Kent Yu, and Lei Zhang. T-rex: Counting by visual prompting. *arXiv preprint arXiv:2311.13596*, 2023. 3

[20] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2024. 1, 3, 4, 7

[21] Qing Jiang, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, Lei Zhang, et al. Chatrex: Taming multimodal llm for joint perception and understanding. *arXiv preprint arXiv:2411.18363*, 2024. 1, 3

[22] Qing Jiang, Xingyu Chen, Zhaoyang Zeng, Junzhi Yu, and Lei Zhang. Rex-thinker: Grounded object referring via chain-of-thought reasoning. *arXiv preprint arXiv:2506.04034*, 2025. 3

[23] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2025. 3

[24] Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Qin Liu, and Lei Zhang. Referring to any person, 2025. 1, 3, 4, 5

[25] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1760–1770, 2021. 3

[26] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 1

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *arXiv: 2304.02643*, 2023. 3, 4, 5

[28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *arXiv: 1811.00982*, 2018. 4

[29] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3

[30] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 1

[31] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025. 5

[32] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 1, 3

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 5

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[37] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting, 2023. 7

[38] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 6

[39] 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. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2, 3, 6

[40] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1

[41] Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners, 2025. 7

[42] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 3

[43] 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*, pages 1049–1059, 2022. 7

[44] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 1, 6

[45] Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanjing Xiong, and Hongsheng Li. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning, 2025. 7

[46] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024. 1, 3

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 5

[48] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023. 1

[49] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1

[50] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. Dave – a detect-and-verify paradigm for low-shot counting, 2024. 7

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[52] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 4

[53] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3743–3751, 2022. 5

[54] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1

[57] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 5

[58] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[59] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 3

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1, 3, 6

[61] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 3, 4

[62] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 4

[63] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 4

[64] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. 2, 5

[65] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting, 2022. 7

[66] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 1

[67] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. 6

[68] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv: 2312.11805*, 2023. 1

[69] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 3

[70] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[71] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 3

[72] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[73] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024. 3

[74] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 6

[75] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024. 5

[76] Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, Yiheng Xu, Junli Wang, Doyen Sahoo, Tao Yu, and Caiming Xiong. Scaling computer-use grounding via user interface decomposition and synthesis, 2025. 7

[77] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 1, 3

[78] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024.

[79] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*, pages 405–422. Springer, 2025. 1, 3

[80] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation, 2023. 3

[81] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1, 2, 3, 6

[82] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 1, 3

[83] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16965–16974, 2024. 1

[84] Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024. 7

[85] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1

[86] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 6