



香港中文大學

The Chinese University of Hong Kong

# VisionReasoner: Unified Reasoning-Integrated Visual Perception via Reinforcement Learning

Yuqi Liu<sup>1</sup>, Tianyuan Qu<sup>1</sup>, Zhisheng Zhong<sup>1</sup>, Bohao Peng<sup>1</sup>

Shu Liu<sup>2</sup>, Bei Yu<sup>1</sup>, Jiaya Jia<sup>2,3</sup>

The Chinese University of Hong Kong<sup>1</sup>

SmartMore<sup>2</sup>

The Hong Kong University of Science and Technology<sup>3</sup>

March 31, 2026



- ① Introduction
- ② Method
- ③ Experiment
- ④ Conclusion

## Section 1

### Introduction



- Through an analysis of diverse visual perception tasks, we observe that many can be categorized into three fundamental types: detection (e.g., object detection<sup>1</sup>), segmentation (e.g., referring expression segmentation<sup>2</sup>, reasoning segmentation<sup>3</sup>), and counting (e.g., object counting<sup>4</sup>).
- Notably, our analysis reveals that these three task types share a common structure as multi-object cognition problems, suggesting that they can be addressed through a unified framework.

---

<sup>1</sup>Tsung-Yi Lin et al. (2014). “Microsoft coco: Common objects in context”. In: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, pp. 740–755.

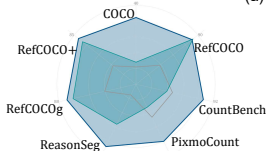
<sup>2</sup>Licheng Yu et al. (2016). “Modeling context in referring expressions”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, pp. 69–85.

<sup>3</sup>Xin Lai et al. (2024). “Lisa: Reasoning segmentation via large language model”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589.

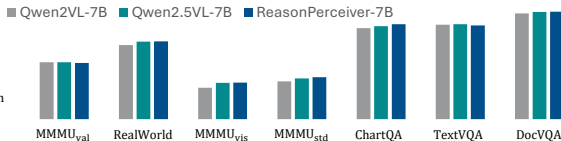
<sup>4</sup>Roni Paiss et al. (2023). “Teaching clip to count to ten”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180.

<p><b>Detection</b></p>		<p>&lt;think&gt; In the image, <i>the Statue of Liberty</i> is a prominent feature on the left side, and it is a <i>well-known symbol of freedom</i> and ,..., recognized as the Statue of Liberty. The Statue of Liberty is also a <i>common tourist attraction</i> and is often featured in photographs of New York City. &lt;/think&gt;</p>	
<p><b>Segmentation</b></p>		<p>&lt;think&gt;The task identifies objects that can provide shade ,..., typically <i>tents or canopies</i>. In the image, ,..., appear to be <i>tents</i>, as they have a <i>dome-like shape and are set up on the ground</i>. These structures are likely designed to provide shade for people who are gathered around them. &lt;/think&gt;</p>	
<p><b>Counting</b></p>		<p>&lt;think&gt; To determine how many objects can help with fishing,..., <i>boats are typically used for fishing</i>, and items like nets, rods, and other fishing gear are often <i>stored on the boats</i>. I'll <i>count the boats</i>.&lt;/think&gt; Total number is <i>8</i>.</p>	

(a) ReasonPerceiver addresses diverse visual tasks.



(b) Performance on visual tasks.



(c) Performance on VQA tasks.

Figure: (a) VisionReasoner addresses diverse tasks within a unified framework. It generates a reasoning process and outputs the expected result corresponding to each query. (b) VisionReasoner significantly outperforms Qwen2.5VL. (c) VisionReasoner retains strong VQA capabilities.

## Section 2

### Method

Although traditional vision models<sup>5,6</sup> achieve strong performance on standard visual perception benchmarks, they struggle with complex instructions.



Figure: VisionReasoner correctly localizes objects from a complex instruction, whereas both commercial DINO-X and open-source YOLO-World fail.

<sup>5</sup>Tianheng Cheng et al. (2024). “Yolo-world: Real-time open-vocabulary object detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911.

<sup>6</sup>Tianhe Ren et al. (2024). “Dino-x: A unified vision model for open-world object detection and understanding”. In: *arXiv preprint arXiv:2411.14347*.



Our analysis of vision perception tasks reveals that many of them can be categorized into three fundamental task types.

- **Detection.** Given an image  $\mathbf{I}$  and a text query  $\mathbf{T}$ , the detection task type aims to generate a set of bounding boxes  $\{\mathbf{B}_i\}_{i=1}^N$  that localize objects of interest.
- **Segmentation** Given an image  $\mathbf{I}$  and a text query  $\mathbf{T}$ , the segmentation task type aims to generate a set of binary segmentation masks  $\{\mathbf{M}_i\}_{i=1}^N$  that identify the regions of interest.
- **Counting** Given an image  $\mathbf{I}$  and a text query  $\mathbf{T}$ , the counting task type aims to estimate the number of target objects specified by the query.

Our VisionReasoner  $\mathcal{F}$  model incorporates a reasoning module, which processing image and locates targeted objects, and a segmentation module that produces segmentation masks if needed.

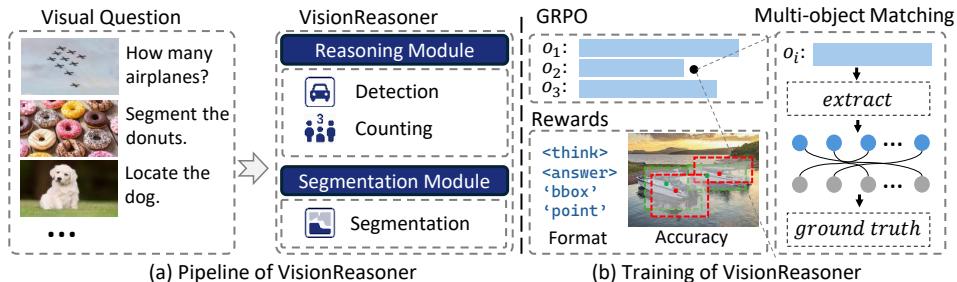


Figure: Illustration of VisionReasoner. (a) For a given image  $\mathbf{I}$  and text instruction  $\mathbf{T}$ , our model generates the expected output corresponding to the instruction. (b) For each observation  $o_i$ , we calculate the rewards and attain the optimal match of multi-objects.



Given an image  $\mathbf{I}$  and a text query  $\mathbf{T}$ , the VisionReasoner  $\mathcal{F}$  generates an interpretable reasoning process, and then produces the bounding boxes  $\{\mathbf{B}_i\}_{i=1}^N$  and binary masks  $\{\mathbf{M}_i\}_{i=1}^N$  if needed. The inference process can be formulated as:

$$(\{\mathbf{B}_i, \mathbf{M}_i\})_{i=1}^N = \mathcal{F}(\mathbf{I}, \mathbf{T}). \quad (1)$$

During inference, we define a specified task type

$\mathbf{C} \in \{\text{detection, segmentation, counting}\}$ . The system then produces the expected outputs as follows:

$$\text{Output} = \begin{cases} \{\mathbf{B}_i\}_{i=1}^N, & \text{if } \mathbf{C} \text{ is detection,} \\ \{\mathbf{M}_i\}_{i=1}^N, & \text{if } \mathbf{C} \text{ is segmentation,} \\ N, & \text{if } \mathbf{C} \text{ is counting.} \end{cases} \quad (2)$$



We design a unified reward mechanism for visual perception tasks, including format rewards and accuracy rewards.

- **Thinking Format Reward** This reward is 1.0 if the model output a thinking process between `<think>` and `</think>` tags, and output the final answer between the `<answer>` and `</answer>` tags.
- **Answer Format Reward** We use bounding boxes  $\{\mathbf{B}_i\}_{i=1}^N$  and points  $\{\mathbf{P}_i\}_{i=1}^N$  as the answer as it has better training efficiency. So this reward restrict the model output answer in  $[\{\text{'bbox\_2d' : [x_1, y_1, x_2, y_2]}, \text{'point\_2d' : [x_1, y_1]}\}, \dots]$ .
- **Non Repeat Format Reward** We split the reasoning process into sentences to detect repeated pattern. A reward of 1.0 is assigned for those with unique or non-repetitive thinking processes.



We design a unified reward mechanism for visual perception tasks, including format rewards and accuracy rewards.

- **Bboxes IoU Reward.** Given a set of  $N$  ground-truth bounding boxes and  $K$  predicted bounding boxes, this reward computes their optimal one-to-one matched Intersection-over-Union (IoU) scores. For each IoU exceeding 0.5, we increment the reward by  $\frac{1}{\max\{N, K\}}$ .
- **Bboxes L1 Reward** Given a set of  $N$  ground-truth bounding boxes and  $K$  predicted bounding boxes, this reward computes their one-to-one matched L1 distance. For each L1 distance below the threshold of 10 pixel, we increment the reward by  $\frac{1}{\max\{N, K\}}$ .
- **Points L1 Reward** Given a set of  $N$  ground-truth points and  $K$  predicted points, this reward computes their one-to-one matched L1 distance. For each L1 distance below the threshold of 30 pixel, we increment the reward by  $\frac{1}{\max\{N, K\}}$ .



---

## Algorithm 1: Multi-object Matching

---

**Input:** pred bboxes  $\mathbf{b}_{\text{pred}} \in \mathbb{R}^{K \times 4}$ ; pred points  $\mathbf{p}_{\text{pred}} \in \mathbb{R}^{K \times 2}$ ; GT bboxes  $\mathbf{b}_{\text{gt}} \in \mathbb{R}^{N \times 4}$ ; GT points  $\mathbf{p}_{\text{gt}} \in \mathbb{R}^{N \times 2}$

**Function** *AccuracyReward*( $\mathbf{b}_{\text{pred}}, \mathbf{p}_{\text{pred}}, \mathbf{b}_{\text{gt}}, \mathbf{p}_{\text{gt}}$ ):

$r \leftarrow 0$ ;  $L_{\text{max}} \leftarrow \max(K, N)$ ;

$IoU \leftarrow \text{BatchIoU}(\mathbf{b}_{\text{pred}}, \mathbf{b}_{\text{gt}}) \in \mathbb{R}^{K \times N}$

$BL1 \leftarrow \text{BatchBoxL1Distance}(\mathbf{b}_{\text{pred}}, \mathbf{b}_{\text{gt}}) \in \mathbb{R}^{K \times N}$

$PL1 \leftarrow \text{BatchPointL1Distance}(\mathbf{p}_{\text{pred}}, \mathbf{p}_{\text{gt}}) \in \mathbb{R}^{K \times N}$

$R_{IoU} \leftarrow [IoU > IoU \text{ threshold}]$

$R_{BL1} \leftarrow [BL1 < \text{Box L1 threshold}]$

$R_{PL1} \leftarrow [PL1 < \text{Point L1 threshold}]$

$C \leftarrow (3 - (R_{IoU} + R_{BL1} + R_{PL1})) \in \mathbb{R}^{K \times N}$

$(\mathbf{r}, \mathbf{c}) \leftarrow \text{Hungarian}(C)$

$\text{total} \leftarrow 3|\mathbf{r}| - \sum_t C_{r_t, c_t}$

$r \leftarrow \text{total}/L_{\text{max}}$ ; **return**  $r$

**Output:** Accuracy reward  $r$

---

## Section 3

### Experiment



We use ten benchmarks to evaluate model performance across general vision perception tasks, including three fundamental task types: detection, segmentation and counting.

Type	Data	# of samples
Det	COCO	36,781
	RefCOCO	5,786
	RefCOCO+	5,060
	RefCOCOg	7,596
Seg	RefCOCO	1,975
	RefCOCO+	1,975
	RefCOCOg	5,023
	ReasonSeg	979
Count	Pixmo-Count	1,064
	CountBench	504
<b>SUM</b>		<b>66,023</b>

**Table:** Statistics of evaluation benchmarks. We report the number of instances for detection and segmentation tasks. The reported numbers combine validation and test splits where applicable.



Method	Detection								Avg.
	COCO		RefCOCO		RefCOCO+		RefCOCog		
	val	testA	val	testA	val	testA	val	test	
<i>Task-specific Models</i>									
VGTR	-	79.0	82.3	63.9	70.1	65.7	67.2	-	
TransVG	-	81.0	82.7	64.8	70.7	68.7	67.7	-	
RefTR	-	85.7	88.7	77.6	82.3	79.3	80.0	-	
MDETR	-	86.8	89.6	79.5	84.1	81.6	80.9	-	
GLIP-T	46.6	50.4	54.3	49.5	52.8	66.1	66.9	55.2	
G-DINO-T	48.4	74.0	74.9	66.8	69.9	71.1	72.1	68.2	
DQ-DETR	50.2	88.6	91.0	81.7	86.2	82.8	83.4	<b>80.6</b>	
<i>Large Vision-language Models</i>									
Shikra-7B	-	87.0	90.6	81.6	87.4	82.3	82.2	-	
InternVL2-8B	-	87.1	91.1	79.8	87.9	82.7	82.7	-	
Qwen2-VL-7B	28.3	80.8	83.9	72.5	76.5	77.3	78.2	71.1	
Qwen2.5-VL-7B	29.2	88.8	91.7	82.3	88.2	84.7	85.7	78.6	
VisionReasoner-7B	37.7	88.6	90.6	83.6	87.9	86.1	87.5	<b>80.3</b>	

Table: Performance comparison on detection tasks.



Method	Segmentation					Avg.
	ReasonSeg		RCO	RCO+	RCOg	
	val	test	testA	testA	test	
<i>Task-specific Models</i>						
LAVT	-	-	75.8	68.4	62.1	-
ReLA	22.4	21.3	76.5	71.0	66.0	51.4
<i>Large Vision-language Models</i>						
LISA-7B	44.4	36.8	76.5	67.4	68.5	58.7
GLaMM-7B	-	-	58.1	47.1	55.6	-
PixelLM-7B	-	-	76.5	71.7	70.5	-
Seg-Zero-7B	62.6	57.5	80.3	76.2	72.6	69.8
Qwen2-VL-7B	44.5	38.7	68.7	65.7	63.5	56.2
Qwen2.5-VL-7B	56.9	52.1	79.9	76.8	72.8	67.7
VisionReasoner-7B	66.3	63.6	78.9	74.9	71.3	<b>71.0</b>

**Table:** Performance comparison on segmentation tasks. We use SAM2 for vision-language models if necessary.



Method	Counting			Avg.
	Pixmo		Count	
	val	test	test	
<i>Large Vision-language Models</i>				
LLaVA-OV-7B	55.8	53.7	78.8	62.8
GLaMM-7B	-	-	-	-
PixelLM-7B	-	-	-	-
Seg-Zero-7B	-	-	-	-
Qwen2-VL-7B	61.6	56.3	80.4	66.1
Qwen2.5-VL-7B	58.1	53.1	78.8	63.6
VisionReasoner-7B	70.1	70.7	89.2	<b>76.7</b>

Table: Performance comparison on counting tasks.



We compare the results with LVLMs and task-specific models on each of the three fundamental task types. It is worthy note that our VisionReasoner is capable of handling different tasks within the same model.

- **Detection.** VisionReasoner achieves superior performance among LVLMs. While our model shows a performance gap compared to some task-specific baselines on COCO datasets, it maintains competitive advantages due to its superior generalization capability.
- **Segmentation.** VisionReasoner achieves state-of-the-art performance, outperforming both general-purpose LVLMs and task-specific approaches.
- **Counting.** VisionReasoner achieves state-of-the-art performance on counting tasks.



VisionReasoner achieves a slight performance gain even though we do not train on VQA data.

Method	OCRBench	RealworldQA	MMMUPro <sub>vision</sub>	MMMUPro <sub>std</sub>
Qwen2VL-7B	809	66.1	28.0	33.8
Qwen2.5VL-7B	822	69.2	32.4	36.4
VisionReasoner-7B	<b>825</b>	<b>69.5</b>	<b>32.6</b>	<b>37.4</b>

Table: Performance comparison on VQA tasks.

<p><b>COCO</b></p> <p>Oven</p>		<p><b>&lt;think&gt;</b> An oven is typically a built-in appliance used for baking and cooking food. In the image, there are <i>two appliances that appear to be ovens, one on the left and one on the right</i>. The left one is a standalone oven with a visible control panel and oven door, while the right one is integrated into the cabinetry and has a similar design. <b>&lt;/think&gt;</b></p>	
<p><b>RefCOCO</b></p> <p>Man on yellow bike.</p>		<p><b>&lt;think&gt;</b> The yellow motorcycle is clearly visible <i>in the center of the image</i>, with a rider <i>wearing a black and orange suit</i>. ... The rider on the yellow motorcycle is wearing a helmet and gloves, and the <i>motorcycle has a distinct yellow color</i> with some green and black accents. <b>&lt;/think&gt;</b></p>	
<p><b>RefCOCO+</b></p> <p>Who is holding the blue mug?</p>		<p><b>&lt;think&gt;</b> The task involves identifying the person holding the blue mug. I need to <i>look for the blue mug and then determine who is holding it</i>. The blue mug is clearly visible on the table, and I can see a hand holding it. The <i>hand is wearing a beige sweater</i>, which helps in identifying the person. <b>&lt;/think&gt;</b></p>	
<p><b>RefCOCOg</b></p> <p>A green truck is on the grass.</p>		<p><b>&lt;think&gt;</b> The image contains two trucks, one green and one blue, both parked on grass. The <i>green truck is on the left side of the image</i>, and the blue truck is on the right. The <i>green truck is clearly visible and distinct in color</i>, making it the object to focus on for the query. <b>&lt;/think&gt;</b></p>	
<p><b>ReasonSeg</b></p> <p>Where the garbage should be put?</p>		<p><b>&lt;think&gt;</b> The task involves identifying where garbage should be put, which typically <i>refers to a designated trash bin or receptacle</i>. In the image, there is a <i>gray trash bin on the right side of the garage door</i>. This is the most likely place for garbage to be put, as it is a common location for such items. <b>&lt;/think&gt;</b></p>	
<p><b>Pixmo-Count</b></p> <p>People</p>		<p><b>&lt;think&gt;</b> The task involves identifying and bounding the people in the image, ..., closely matched ones. The people in the image are clearly visible and <i>can be identified by their human-like features and clothing</i>. ... The people are the main subjects of the image. <b>&lt;/think&gt;</b> Total number is 4.</p>	
<p><b>CountBench</b></p> <p>How many nail polishes are there in the image?</p>		<p><b>&lt;think&gt;</b> The task involves identifying the number of nail polishes in the image and comparing the objects to find the most closely matched ones...., I will <i>count the nail polishes</i>. <b>&lt;/think&gt;</b> Total number is 5.</p>	

Figure: Qualitative results on different tasks.

Section 4

Conclusion



- We present VisionReasoner, a unified vision-language framework for reasoning visual perception tasks.
- VisionReasoner demonstrates strong capabilities in analyzing visual inputs, generating structured reasoning processes and delivering task-specific outputs.
- Experiments across ten diverse tasks, spanning detection, segmentation and counting, validates the robustness and versatility of our approach.

**THANK YOU!**