VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning

Yuqi Liu[1*] **Tianyuan Qu**[1*] **Zhisheng Zhong**[1] **Bohao Peng**[1] **Shu Liu**[2] **Bei Yu**[1] **Jiaya Jia**[2,3]
CUHK[1] SmartMore[2] HKUST[3]
https://github.com/dvlab-research/VisionReasoner

Abstract

Large vision-language models exhibit inherent capabilities to handle diverse visual perception tasks. In this paper, we introduce VisionReasoner, a unified framework capable of reasoning and solving multiple visual perception tasks within a shared model. Specifically, by designing novel multi-object cognitive learning strategies and systematic task reformulation, VisionReasoner enhances its reasoning capabilities to analyze visual inputs, and addresses diverse perception tasks in a unified framework. The model generates a structured reasoning process before delivering the desired outputs responding to user queries. To rigorously assess unified visual perception capabilities, we evaluate VisionReasoner on ten diverse tasks spanning three critical domains: detection, segmentation, and counting. Experimental results show that VisionReasoner achieves superior performance as a unified model, outperforming Qwen2.5VL by relative margins of 29.1% on COCO (detection), 22.1% on ReasonSeg (segmentation), and 15.3% on CountBench (counting).

1 Introduction

Recent advances in large vision-language models (LVLMs) [1, 44, 7, 31] have demonstrated remarkable capabilities in visual conversations. As the field progresses, researchers are increasingly applying LVLMs to a wider range of visual perception tasks, such as visual grounding [35] and reasoning segmentation [12, 24] , often incorporating task-specific modules or techniques.

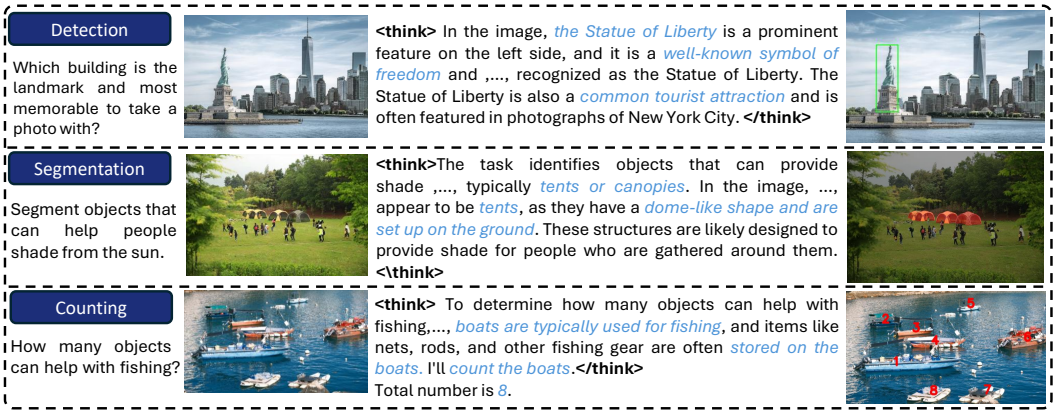
Inspired by the emergent test-time reasoning capabilities of large language models (LLMs) [8, 30], recent studies have explored the integration of reinforcement learning (RL) into LVLMs [43, 25, 24, 50]. Works such as VisualRFT [25] and Seg-Zero [24] demonstrate that RL can enhance reasoning in visual perception tasks. However, these approaches often employ RL in a task-specific manner, relying on distinct reward functions for different tasks, which limits their scalability and generalizability.

Through an analysis of diverse visual perception tasks, we observe that many can be categorized into three fundamental types: detection (e.g., object detection [17], visual grounding [48]), segmentation (e.g., referring expression segmentation [48], reasoning segmentation [12]), and counting (e.g., object counting [33]). 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.

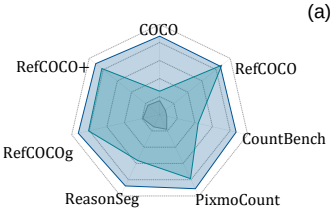
Building on this insight, we propose VisionReasoner, a unified framework that addresses diverse visual perception tasks through a shared architecture. The framework's core capabilities, which include advanced reasoning and multi-object cognition, are enabled through a carefully designed reward mechanism. Format rewards, including thinking rewards that promote structured reasoning and non-repeat rewards that prevent redundant reasoning patterns. Accuracy rewards, comprising multi-object IoU rewards and L1 rewards for precise localization, strengthen multi-object cognition. Unlike previous approaches like Kosmos [35] that use cross-entropy loss, our RL framework requires

[*]Equal contribution.

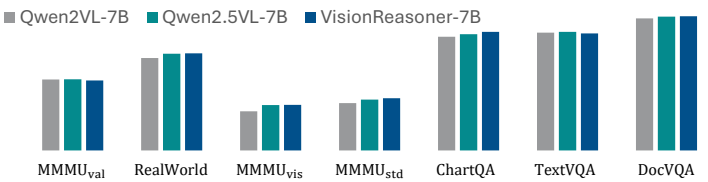
Preprint.



(a) VisionReasoner addresses diverse visual tasks.



(b) Performance on visual tasks.



(c) Performance on VQA tasks.

Figure 1: (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.

optimal prediction-to-ground-truth matching. We address this challenge by implementing an efficient matching pipeline combining the batch computing and the Hungarian algorithm, significantly improving computational efficiency while maintaining matching accuracy.

To comprehensively evaluate model performance, we conduct extensive experiments with VisionReasoner across 10 diverse tasks spanning three fundamental types: detection, segmentation, and counting. Remarkably, our VisionReasoner-7B model achieves strong performance despite being trained on only 7,000 samples, demonstrating both robust test-time reasoning capabilities and effective multi-task generalization, as shown in Figure 1 (a)-(b). Experimental results show significant improvements over baseline models, with relative gains of 29.1% on COCO-val (detection), 22.1% on ReasonSeg-test (segmentation), and 15.3% on CountBench-test (counting), validating the effectiveness of our unified approach. Additionally, VisionReasoner exhibits visual question answering capabilities comparable to state-of-the-art models, as shown in Figure 1 (c).

Our contributions are summarized as follows:

- We propose VisionReasoner, a unified framework for visual perception tasks. Through carefully crafted rewards and training strategy, VisionReasoner has strong multi-task capability, addressing diverse visual perception tasks within a shared model.
- Experimental results show that VisionReasoner achieves superior performance across ten diverse visual perception tasks within a single unified framework, outperforming baseline models by a significant margin.
- Through extensive ablation studies, we validate the effectiveness of our design and offer critical insights into the application of RL in LVLMs.

2 Related Works

2.1 Large Vision-language Models

Following LLaVA's [20] pioneering work on visual instruction tuning for large vision-language models, subsequent studies [44, 28, 31, 1, 16, 51] have adopted this paradigm for vision-language

conversation. Beyond visual conversation tasks, LVLMs have been extended to diverse vision applications, including visual grounding [35] and reasoning segmentation [12]. Notely, the recent GPT-4.1 [31] demonstrates state-of-the-art performance in multi-modal information processing and visual reasoning. Although these models are evaluated on specific tasks, their performance has not been systematically evaluated under a unified visual perception framework.

2.2 Reinforcement Learning in Large Models

In the field of large language model (LLMs), various reinforcement learning (RL) algorithms are used to enhance model performance, such as reinforcment learning from human feedback (RLHF) [32], direct preference optimization (DPO) [36] and proximal policy optimization (PPO) [40]. The recent DeepSeek R1 [8], trained using GRPO [41], demonstrates remarkable test-time scaling capabilities, significantly improving reasoning ability and overall performance. Building on these advances, researchers try to apply these RL techniques to LVLMs. Notable efforts include Visual-RFT [25], EasyR1 [50] and Seg-Zero[24], all of which exhibit strong reasoning capabilities and achieve impressive performance.

3 Method

To develop a unified visual perception model capable of solving diverse vision tasks, we first identify and analyze the representative visual perception tasks, then reformulate their inputs and outputs into a set of three fundamental task categories (Section 3.1). Next, we detail the architecture of our VisionReasoner model (Section 3.2). Additionally, we present the reward functions employed for training our model (Section 3.3). Finally, we elaborate on our novel training strategy of multi-object cognition (Section 3.4).

3.1 Task Reformulation and Categorization

Our analysis of vision tasks listed in *Papers With Code* [34] reveals that approximately 50 task types (constituting 10% of the roughly 500 cataloged vision task types) can be categorized into three fundamental task types. This suggests that a single model capable of addressing these fundamental task types could potentially solve 10% of existing vision tasks. Further details are provided in the *supplementary materials*.

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. This type requires multi-object cognition ability. This category includes tasks such as Visual Grounding [48, 11] and Object Detection [17].

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. We address this type by detect-then-segment paradigm. This category includes tasks such as Referring Expression Segmentation [11, 48] and Reasoning Segmentation [12, 46].

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. We address this type by detect-then-count paradigm. This category includes tasks such as Object Counting [5, 33].

3.2 VisionReasoner Model

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. The whole architecture is shown in Figure 2 (a). We implement our model based on Seg-Zero [24], due to its superior performance in single-object segmentation tasks. To broaden its applicability, we augment it with multi-object cognition capabilities. This critical enhancement enables VisionReasoner to address three fundamental task types: detection, segmentation, counting. Specifically, given an image $\mathbf{I}$, a text query $\mathbf{T}$, the VisionReasoner $\mathcal{F}$ generates an interpretable reasoning process, and then produces the expected output corresponding to $\mathbf{T}$. The model output is represented in a structured format $(\{\mathbf{B}_i, \mathbf{M}_i\})_{i=1}^N$, where $\mathbf{B}_i$ denotes the bounding box (bbox) and $\mathbf{M}_i$ represents the binary mask of the targeted object. Note that $\mathbf{M}_i$ is produced by the segmentation module using $\mathbf{B}_i$ and central

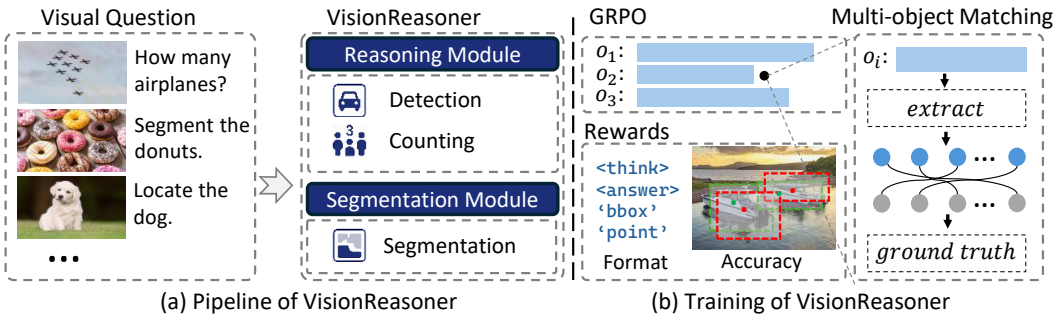


Figure 2: Illustration of VisionReasoner. (a) For a given image $\mathbf{I}$ and text instruction $\mathbf{T}$, our model generates the expected output corresponding to the instruction. (b) For each observation o_i, we calculate the rewards (Section 3.3) and attain the optimal match of multi-objects (Section 3.4)

points $\mathbf{P}_i$ as prompts. This process can be formulated as:

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

During inference, the user provide the input image $\mathbf{I}$ and text prompt $\mathbf{T}$, and define a specified task type $\mathbf{C} \in \{\text{detection, segmentation, counting}\}$. The system then produces the expected outputs as follows:

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

In this way, our VisionReasoner can process diverse perception tasks in a unified manner within a shared framework.

3.3 Reward Functions

Our reward functions consist of two types: format rewards and accuracy rewards. Following Seg-Zero [24], we use target object bboxes and center points to calculate the rewards rather than binary masks. These rewards jointly guide the optimization process by reinforcing both structural correctness and multi-object recognition performance.

Thinking Format Reward. This reward constraint 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 $[\{'\text{bbox\_2d}' : [x_1, y_1, x_2, y_2], '\text{ point\_2d}' : [x_1, y_1]\}, ...]$.

Non Repeat Reward. To avoid repeated patterns, we split the reasoning process into individual sentences and prioritize those with unique or non-repetitive thinking processes.

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 $\frac{1}{\max\{N, K\}}$.

3.4 Multi-object Cognition

Multi-object Data Preparation. We derive bboxes and points directly from the original mask annotations in existing segmentation datasets (e.g., RefCOCOg [48], LISA++[46]). Specifically, for a given binary mask of an object, we determine its bounding box by extracting the leftmost, topmost, rightmost, and bottommost pixel coordinates. Additionally, we compute the center point coordinates of the mask. Different from Seg-Zero's [24] single-object formulation, we process multiple objects per image by: (i) using one central point (ii) joining all textual descriptions with the conjunction 'and', and (iii) concatenating all associated bounding boxes and center points into list per image.

Multi-object Matching. A key challenge in training VisionReasoner is its multi-object matching mechanism. Our framework addresses this through batch computation and the Hungarian algorithm, which optimally solves the many-to-many matching problem for bounding boxes IoU rewards, bounding boxes L1 rewards, and points L1 rewards. As shown in Figure 2 (b), for each observation o_j, which contains a list of bboxes $(\mathbf{B}_{\mathbf{pred}_i})_{i=1}^{K}$ and points $(\mathbf{P}_{\mathbf{pred}_i})_{i=1}^{K}$, we calculate its reward scores with the ground-truth bboxes $(\mathbf{B}_{\mathbf{GT}_i})_{i=1}^{N}$ and points $(\mathbf{P}_{\mathbf{GT}_i})_{i=1}^{N}$ by implementing batch computation. We then calculate the optimal one-to-one matching with using Hungarian algorithm. These design guarantees optimal assignment between predictions and ground truth annotations while achieving high computational efficiency.

4 Experiments

4.1 Evaluation Benchmarks

We use ten benchmarks to evaluate model performance across general vision perception tasks. Our evaluation includes three fundamental task types: detection, segmentation and counting. Specifically, we employ COCO [17] and RefCOCO(+/g) [48] for detection evaluation; RefCOCO(+/g) and ReasonSeg [12] for segmentation evaluation; PixMo-Count [5] and CountBench [33] for counting evaluation.

Table 1: 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.

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
SUM		66,023

Annotation Preparation. To ensure consistency across all evaluation tasks, we standardize the evaluation data by converting all samples into a unified multi-modal conversation format and removing potential information leakage. This preprocessing involves: converting numeric class labels to textual descriptions in COCO [17]; removing explicit numerical references from text descriptions in CountBench [33]; applying consistent formatting across all datasets to maintain evaluation fairness.

Evaluation Metrics. For object detection on COCO, we adopt the standard AP metric computed using the COCO API [42]. For referring object grounding on RefCOCO(+/g), we use bbox AP, which measures detection accuracy at an IoU threshold of 0.5. For object segmentation on RefCOCO(+/g) and ReasonSeg, we use gIoU, computed as the mean IoU across all segmentation masks. For counting tasks, we use count accuracy as evaluation metric.

Statistics and Visualization. We show the statistic data in Table 1. For detection and segmentation tasks, we report the number of valid instances. For counting tasks, we provide the total number of test samples. Our evaluation comprises a total of 66,023 test samples, covering three fundamental visual perception task types and 10 specific tasks. We visualize some examples in Figure 3.

4.2 Experimental Settings

Training Data. The training data is sourced from four datasets: LVIS [9], RefCOCOg [48], gRefCOCO [18], and LISA++ [46], following the strategy outlined in Section 3.4. These datasets provide diverse text annotations: LVIS employs simple class names as text, RefCOCOg contains referring

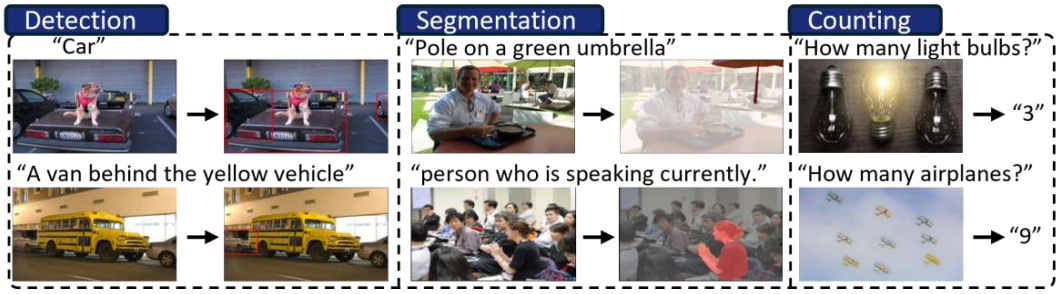


Figure 3: Examples from evaluation benchmarks. Zoom in for better viewing.

Table 2: Performance comparison on detection tasks.

Method	Detection							Avg.
	COCO	RefCOCO		RefCOCO+		RefCOCOg		
	val	val	testA	val	testA	val	test	
Task-specific Models								
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	-
OWL-ViT	30.9	-	-	-	-	-	-	-
YOLO-World-S	37.6	-	-	-	-	-	-	-
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	**80.6**
Large Vision-language Models								
Shikra-7B	-	87.0	90.6	81.6	87.4	82.3	82.2	-
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	**80.3**

expressions where each text corresponds to a single object, gRefCOCO includes expressions that may refer to multiple objects, and LISA++ features texts requiring reasoning. Together, these datasets comprise a diverse range of text types, with approximately 7,000 training samples in total.

Reinforcement Learning. We train VisionReasoner using the GRPO algorithm [41]. During training, the policy model produces several response samples for each input. These samples are assessed by reward functions, and the policy model optimizes its parameters to maximize reward while maintaining proximity to the reference model via KL-divergence regularization.

Implementation Details. We initialize VisionReasoner with the similar settings in Seg-Zero [24]. We employ a batch size of 16 and a learning rate of 1e-6. The entire training process takes 6 hours.

4.3 Main Results

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 and is evaluated in a zero-shot manner.

Detection. We compare VisionReasoner with several state-of-the-art LVLMs, including Shikra [2], Qwen2-VL-7B [44] and Qwen2.5VL-7B [1]. For task-specific models, we evaluate against VGTR [4], TransVG [6], RefTR [15], MDETR [10], OWL-ViT [29], YOLO-World [3], GroundingDINO [22], DQ-DETR [21], GLIP [14]. Since LVLMs do not output confidence score, we approximate it using the ratio of the bounding box area to the total image area (bbox_area / image_area) to enable compatibility with COCOAPI [42]. However, this coarse approximation leads to underestimated AP scores. As shown in Table 2, VisionReasoner achieves superior performance among LVLMs. While

Table 3: Performance comparison on segmentation tasks and counting tasks. We use SAM2 for vision-language models if necessary in segmentation tasks.

Method	Segmentation					Avg.	Counting			Avg.
	ReasonSeg		RCO	RCO+	RCOg		Pixmo		Count	
	val	test	testA	testA	test		val	test	test	
Task-specific Models										
LAVT	-	-	75.8	68.4	62.1	-	-	-	-	-
ReLA	22.4	21.3	76.5	71.0	66.0	51.4	-	-	-	-
Large Vision-language Models										
LISA-7B	44.4	36.8	76.5	67.4	68.5	58.7	-	-	-	-
LLaVA-OV-7B	-	-	-	-	-	-	55.8	53.7	78.8	62.8
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	29.9	48.0	76.5	51.5
Qwen2.5-VL-7B	56.9	52.1	79.9	76.8	72.8	67.7	63.3	67.9	76.0	69.1
VisionReasoner-7B	66.3	63.6	78.9	74.9	71.3	**71.0**	70.1	69.5	87.6	**75.7**

Table 4: Comparison on the reasoning length. Complex instructs require longer reasoning.

Data	Avg. Len (# words)
COCO	62
RefCOCOg	65
ReasonSeg	71

Table 5: Comparison of multi-object matching. Our code achieves a 6×10^{35} times speedup.

Hungarian	Batch Comp	Time (s)
		$> 3 \times 10^{32}$
✓		2×10^{-3}
✓	✓	5×10^{-4}

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. We evaluate VisionReasoner against state-of-the-art LVLMs, including LISA [12], GLaMM [37], PixelLM [39], Seg-Zero [24], Qwen2-VL [44] and Qwen2.5VL [1]. For these LVLMs, we first extract bounding box predictions and subsequently send them into SAM2[38] to generate segmentation masks. We also compare task-specific models, including LAVT [47] and ReLA [19]. For models that do not report gIoU, we report their cIoU as an alternative. As shown in Table 3, VisionReasoner achieves state-of-the-art performance, outperforming both general-purpose LVLMs and task-specific approaches.

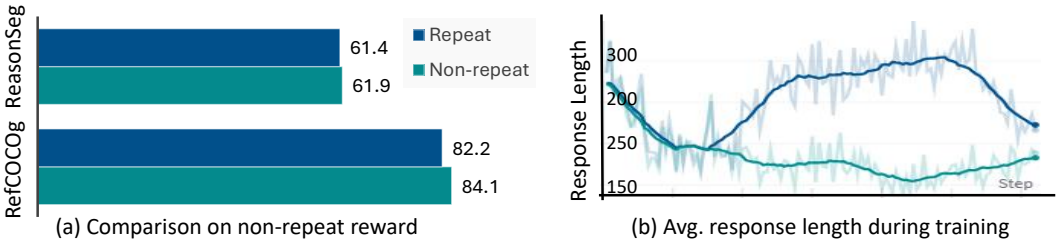
Counting. We evaluate VisionReasoner against state-of-the-art LVLMs, including LLaVA-OneVision [13], Qwen2-VL-7B [44] and Qwen2.5VL-7B [1]. We evaluate these LVLMs in a first-detect-then-count manner. As shown in Table 3, VisionReasoner achieves state-of-the-art performance.

4.4 Ablation Study

We perform ablation studies to assess the effectiveness of our design and validate the optimal hyper-parameter selection and training recipe design for VisionReasoner. We also evaluate VisionReasoner on VQA tasks.

Reasoning Length. As shown in Table 4, our analysis reveals that the model's reasoning length adapts dynamically to text query complexity. Specifically, for simple class names in COCO and short phrases in RefCOCOg, the reasoning process is relatively concise. In contrast, complex reasoning-intensive queries in ReasonSeg require longer reasoning processes.

Multi-object Matching. We quantitatively assess the efficiency of our two key design choices for multi-object matching: the Hungarian algorithm and batch computation. In a scenario with 30 objects, Table 5 demonstrates that random permutation and simple matching require over 3×10^{32} seconds (i.e., 8×10^{24} years) to complete, while our optimized approach achieves matching in just 5×10^{-4} seconds - a 6×10^{35} times speedup.



(a) Comparison on non-repeat reward (b) Avg. response length during training

Figure 4: Ablation on non-repeat reward. (a) Consistent performance gain across different datasets using non-repeated reward. (b) Non-repeat rewards lead to shorter response lengths.

Table 6: Performance comparison on different training data.

Method	Training Data				Det	Seg	Avg.
	RefCOCOg	gRefCOCO	LVIS	LISA++	RefCOCOg-val	ReasonSeg-val	
VisionReasoner-7B	✓				84.1	61.9	73.0
	✓	✓			85.8	63.8	74.8
	✓	✓	✓		85.5	64.2	74.9
	✓	✓	✓	✓	**86.1**	**66.3**	**76.2**

Table 7: Performance comparison on VQA tasks.

Method	OCRBench	RealworldQA	MMMU$_{vision}$	MMMU$_{std}$	ChartQA	DocVQA
Qwen2VL-7B	809	66.1	28.0	33.8	81.4	94.5
Qwen2.5VL-7B	822	69.2	32.4	36.4	83.1	95.7
VisionReasoner-7B	**825**	**69.5**	**32.6**	**37.4**	**84.9**	**96.0**

Non Repeat Reward. Figure 4 (a) presents the performance comparison with and without the non-repeat reward. Models are trained only on 2,000 samples from RefCOCOg. The model achieves better results when trained using the non-repeat reward. Additionally, model without non-repeat reward tends to generate longer reasoning processes, as shown in Figure 4 (b), and we observe repetitive reasoning patterns during inference.

Different Training Data. We conduct an ablation study on different training datasets, with results presented in Table 6. The four datasets provide diverse text annotations: LVIS uses simple class names, RefCOCOg contains single-object referring expressions, gRefCOCO includes expressions that may refer to multiple objects, and LISA++ features texts requiring reasoning. Our experiments demonstrate that these datasets consistently improve model performance.

Visual QA Ability. We also compare VisionReasoner's VQA [26, 27, 23, 49, 45] ability with Qwen2VL [44] and the baseline model Qwen2.5VL[1]. As shown in Table 7, VisionReasoner achieves a slight performance gain even thought we do not train on VQA data.

Sampling Number. Figure 5 presents performance comparison with different sampling number. Models are trained using all 7,000 training samples. We observe an initial performance gain followed by a notable decline with larger sampling number, suggesting that excessive sampling may induce overfitting to the training distribution and consequently degrade generalization capability.

Reasoning. Figure 6 compares the performance of models with and without reasoning capabilities. We implement no reasoning by removing thinking reward. Models are trained using all 7,000 training samples. Our results show that both approaches outperform the baseline. And the reasoning-enhanced model demonstrates significant gain on intricate reasoning segmentation data.

4.5 Qualitative Results

We visualize some results on Figure 7. Notably, VisionReasoner addresses several visual perception tasks within a shared model. Our model generates comprehensive reasoning processes for all

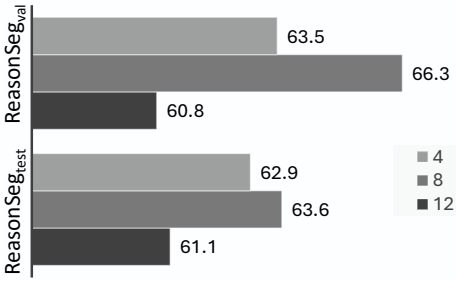


Figure 5: Different sampling number.

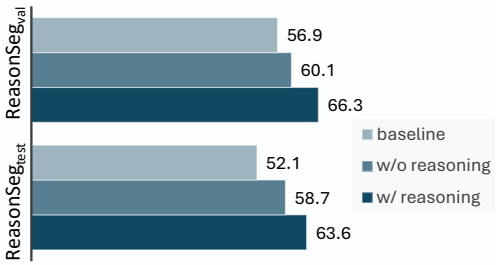


Figure 6: Reasoning vs. no reasoning

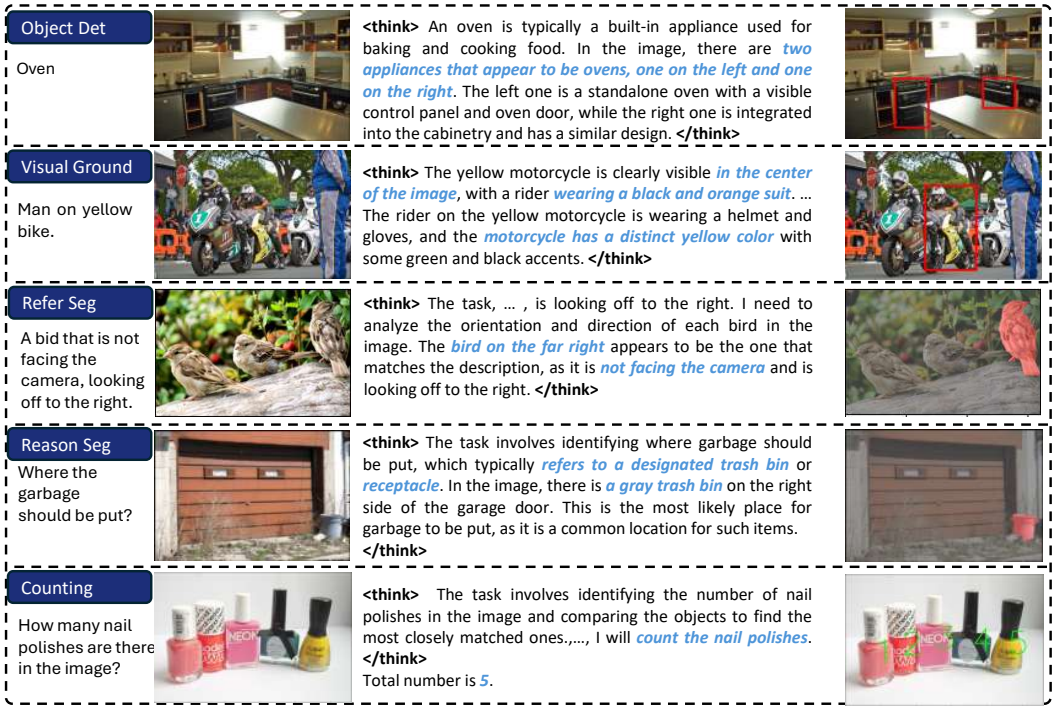


Figure 7: Qualitative results on different datasets. Zoom in for better visualization.

tasks while producing expected outputs. We find that VisionReasoner can effectively distinguish between similar objects, as shown in the visual grounding and referring expression segmentation. VisionReasoner also accurately localize multiple targets, as shown in object detection and counting. We also observe that the length of the reasoning process adapts dynamically: more intricate image-query pairs elicit detailed rationales, while simpler inputs result in concise explanations.

5 Conclusion

In this paper, we present VisionReasoner, a unified vision-language framework. With systematic task reformulation, VisionReasoner effectively addresses diverse visual perception tasks within a shared model. By introducing novel multi-object cognitive learning strategies and curated reward functions, 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. Notably, VisionReasoner achieves significant improvements over existing models such as Qwen2.5VL, with relative performance gains of 29.1% on COCO (detection), 22.1% on ReasonSeg (segmentation), and 15.3% on CountBench (counting). These results highlight the potential of unified frameworks in advancing multiple visual perception tasks.

References

[1] 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.

[2] 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.

[3] 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.

[4] Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19462–19472, 2023.

[5] 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.

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021.

[7] Google. Gemini pro 2.5. `https://deepmind.google/technologies/gemini/`, 2025.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[9] 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.

[10] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[12] 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.

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[14] 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.

[15] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

[16] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[17] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[18] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[19] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[21] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1728–1736, 2023.

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

[23] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[24] 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.

[25] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

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

[27] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[28] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024.

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

[30] OpenAI. OpenAI o1. `https://openai.com/o1/`, 2024.

[31] OpenAI. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`, 2025.

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[33] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3170–3180, 2023.

[34] Papers with Code. Datasets - papers with code. `https://paperswithcode.com/datasets?mod=images&page=1`, 2021.

[35] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[39] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[41] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[42] COCO Team. Microsoft COCO: Common objects in context. `https://github.com/cocodataset/cocoapi`, 2014.

[43] R1-V Team. R1-V. `https://github.com/Deep-Agent/R1-V?tab=readme-ov-file`, 2025.

[44] 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.

[45] xAI. Grok-1.5 vision. `https://x.ai/news/grok-1.5v`, 2024.

[46] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

[47] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022.

[48] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[49] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[50] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

[51] Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, et al. Lyra: An efficient and speech-centric framework for omni-cognition. *arXiv preprint arXiv:2412.09501*, 2024.