Dr. Seg: Revisiting GRPO Training for Visual Large Language Models through Perception-Oriented Design

Supplementary Material

We provide supplementary material related to the main paper, arranged as follows:

1. More Experiment Details and Ablations (Section A)
2. Qualitative Results (Section B)
3. Additional Results on REC (Section C)
4. Dataset Details (Section D)

A. More Experiment Details and Ablations

Details of Dr. Seg. Our project is built on SAM [7], VERL [16], EasyR1 [21], and VisionReasoner [14]. The global batch size is 16, with the micro batch size per device for updates and the micro batch size per device for experience collection both set to 2. The rollout number per sample is 8, and the KL-loss coefficient is 1×10^{-2}. We use the AdamW optimizer with a learning rate of 1×10^{-6} and a weight decay of 1×10^{-2}.

Details of VisionReasoner with Raw Reward. In Sec. 3.3, we replace the binary rewards in VisionReasoner with unnormalized continuous rewards. Concretely, for the box IoU reward, we directly use the continuous IoU value without applying any threshold. For the box L1 reward and point L1 reward, we adopt the same linear scaling function in Eq.15. All other settings are kept the same as in the original VisionReasoner configuration.

COCO Precision/Recall Metrics. Following VisionReasoner [14], we compute mAP by assigning each predicted box a proxy score defined as the ratio of the predicted box area to the image area. We also evaluate Precision/Recall/F1 to avoid such approximation.

Method	AP@0.5	P@0.5	R@0.5	F1
Qwen2-VL	43.5	72.1	41.3	52.5
Qwen2.5-VL	47.6	72.4	46.6	56.7
VisionReasoner	57.3	62.6	63.9	63.3
Dr. Seg	59.0	69.6	64.3	66.8

Table 1. More detection metrics on COCO.

Ablation on Response Length. To isolate whether the gains come from simply generating longer responses, we compare several test-time prompting variants that increase the average response length while keeping the underlying model unchanged. As shown in Tab. 2, merely elongating the output does not improve performance: both a generic Chain-of-Thought prompt and a Look-to-Confirm style prompt substantially increase the response length but degrade gIoU (e.g., 65.5 $\rightarrow$ 63.1 and 59.6). In contrast, Dr. Seg achieves the best gIoU (67.8) with a similar length range, indicating that the improvement cannot be attributed to longer responses. Length denotes the word count, measured on the ReasonSeg val set.

Method	Length	gIoU
Baseline	77.8	65.5
+ Chain-of-Thought prompt	117.0	63.1
+ Look-to-Confirm prompt	137.3	59.6
Dr. Seg	148.3	67.8

Table 2. Ablation on response length using test-time prompts.

Ablation on Format Reward. To further verify whether the gain could come from a particular output structure (e.g., the $\langle$look$\rangle$ tag) or learned semantic guidance, We conduct a controlled ablation that matches structure by retraining with the same format reward R_{format}, while changing the keyword inside the tag (e.g., $\langle$look$\rangle$, $\langle$wait$\rangle$) or using unrelated keywords. As shown in Tab. 3, adding R_{format} with random/wait keywords yields only marginal changes (64.9–65.6 gIoU), suggesting that format constraints alone are insufficient. Moreover, removing R_{format} (Look*) fails to stably generate the tag and degenerates toward the baseline, indicating R_{format} is necessary for stable structured behavior.

Keyword	R_{format}	Length	gIoU
Random	✓	89.6	64.9
Wait	✓	95.3	65.6
Look*		73.6	65.8
Look	✓	148.3	67.8

Table 3. Ablation on response structure by retraining with/without the format reward R_{format} and varying the tag keyword. Look*: without R_{format}, the model fails to stably generate $\langle$look$\rangle$ and degrades toward the baseline.

Ablation on Accuracy Reward Components. To clarify the contribution of each accuracy component, we ablate the reward composition used in R_{accuracy} by progressively adding x_1 (IoU), x_2 (count), and x_3 (point) in Tab. 4.

B. Qualitative Results

We provide a qualitative example in Fig. 1 and more visualizations in Fig. 5. As illustrated in Fig. 1, prior methods [14] struggle in complex multi-object segmentation scenarios, often miscounting target instances due to over- or under-segmentation. For example, in the piano half-step case, the

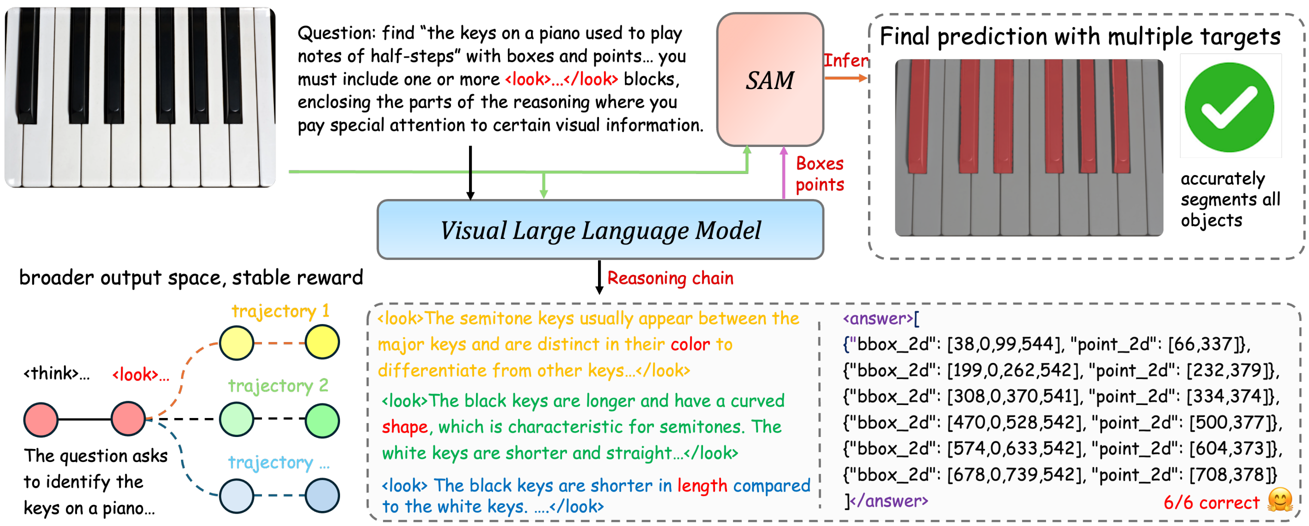

Figure 1. Supplementary illustration of the Dr. Seg architecture. The model first performs a Look-to-Confirm exploration, generating multiple reasoning trajectories that describe diverse visual cues. During inference, these trajectories are used to prompt SAM to produce mask proposals. The example shown here demonstrates successful segmentation of all six semitone keys (6/6 correct).

Setting	RefCOCO testA	ReasonSeg val
w	35.6	24.1
$w\, x_1$	72.9	57.8
$w\, x_1, x_2$	79.8	58.3
$w\, x_1, x_2, x_3$	80.2	67.8

Table 4. Ablation on R_{accuracy}.

baseline model correctly identifies only 4 out of 6 candidate keys. We argue that these errors stem from two coupled factors: (1) binary rewards fail to reflect fine-grained performance differences, providing only coarse supervision; and (2) the restricted output space limits the model's ability to explore diverse reasoning trajectories, causing incorrect object enumeration. Our method alleviates these issues: the proposed Look-to-Confirm mechanism expands the model's output space by encouraging exploration of alternative visual cues, while the Distribution-Ranked Reward delivers stable fine-grained feedback independent of metric scales. Together, these components yield substantially more accurate object counts. We provide more visualizations in Fig. 5, highlighting the effectiveness of our method in challenging multi-object settings.

C. Additional Results on REC

We also evaluate our models' performance on the referring expression comprehension (REC) task, with results reported in Tab. 6. Our model also achieves superior performance compared with previous methods [14] and base models [1, 17]. Specifically, it outperforms the baseline VisionReasoner on all three REC datasets, with an average improvement of 0.7 absolute points.

D. Dataset Details

Training Dataset Statistics. We use the *VisionReasoner_multi_object_7k_840* dataset [14], which was constructed by the authors of VisionReasoner from four sources: LVIS [6], RefCOCOg [19], gRefCOCO [10], and LISA++ [18]. It is formed by sampling roughly 1,800 examples from the training split of each dataset, without any special data partitioning criteria, resulting in a balanced mixture across the four sources.

Evaluation Dataset Statistics. Table 5 summarizes the statistics of all evaluation datasets used in this study. For a fair comparison, we follow prior work [13, 14] and adopt the same data splits as in their experiments; we report the number of evaluation samples. The benchmark suite covers (i) referring expression segmentation (RefCOCO, RefCOCO+, RefCOCOg [19]); (ii) reasoning-oriented segmentation (ReasonSeg [8]); (iii) our self-constructed multi-object split (COCONut; details in the next paragraph); (iv) object detection (MS COCO [9]); (v) object counting (Pixmo-Count [4], CountBench [15]) and (vi) referring expression comprehension (RefCOCO, RefCOCO+, RefCOCOg).

Evaluation Metrics. For object detection on COCO, we adopt the standard AP metric computed using the COCO API. For referring expression comprehension on RefCOCO series, we use bbox AP, which measures detection accuracy at an IoU threshold of 0.5. For object segmentation on RefCOCO series and ReasonSeg, we use gIoU, computed as the mean IoU across all segmentation masks. For counting tasks, we use count accuracy as the evaluation metric.

Table 5. Statistics of evaluation benchmarks. Numbers combine validation and test splits where applicable. DET, REC, SEG, and CNT denote Detection, Referring Expression Comprehension, Referring Expression Segmentation, and Counting, respectively.

Dataset	Split	# of samples
DET	COCO	36,781
REC	RefCOCO	5,786
	RefCOCO+	5,060
	RefCOCOg	7,596
SEG	RefCOCO	1,975
	RefCOCO+	1,975
	RefCOCOg	5,023
	ReasonSeg	979
	COCONut	665
CNT	Pixmo-Count	1,064
	CountBench	504
SUM		**67,408**

Details of COCONut Benchmark. Since publicly available datasets for complex, multi-object segmentation remain scarce, we additionally construct a benchmark from COCONut val split [5], a dataset that couples panoptic segmentation with grounded captions.

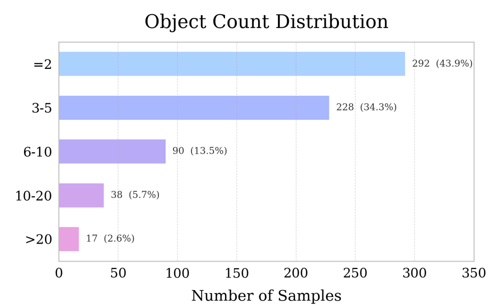

Figure 2. Object count distribution of the COCONut multi-object split.

Figure 3. Categories included in our COCONut dataset, all of which are subclasses of the original COCO label set.

Sampling Protocol. We sample 1,000 images from the official validation split and, after filtering, retain 665 images. The procedure is as follows:

1. **Candidate filtering**: We discard images without annotations. For each remaining image, we count per-class instance frequencies and keep only those images in which at least one class appears at least twice. For each retained image, we collect the set of classes that satisfy this criterion and randomly choose one target class from this set.
2. **Union-mask construction**: For the chosen class, we build a binary mask for every instance of that class and take the pixel-wise OR over all such masks to obtain the class-level union mask across the full image.
3. **Text prompt**: For each sample, we form the instruction `All <category>`, indicating "segment all instances of the specified class."
4. **Quality Control**: Finally, we perform a light manual pass to remove low-quality cases (e.g., extremely small targets or ambiguous class references). The remainder constitutes our dataset.

Goal and Characteristics. This benchmark evaluates class-level multi-instance union reasoning: given a natural image and a concise class prompt, the model must output the combined region covering all instances of that class. Unlike typical single-instance prompts, this setting stresses coverage completeness over multiple same-class objects, which better reflects crowded or overlapping real-world scenes. We also provide summary statistics and qualitative examples in Fig. 2 and Fig. 4.

Table 6. Performance on the referring expression comprehension task.

Method	Referring Expression Comprehension						
	RefCOCO		RefCOCO+		RefCOCOg		avg
	val	*testA*	*val*	*testA*	*val*	*test*	
Task-specific Models							
DQ-DETR [11]	88.6	91.0	81.7	86.2	82.8	83.4	85.6
Grounding-DINO-T(ft) [12]	89.2	91.9	81.1	87.4	84.2	84.9	86.5
Grounding-DINO-L(ft) [12]	**90.6**	**93.2**	82.8	**89.0**	86.1	87.0	**88.1**
7B Visual Large Language Models							
Qwen2-VL [17]	80.8	83.9	72.5	76.5	77.3	78.2	78.2
Qwen2.5-VL [1]	88.8	91.7	82.3	88.2	84.7	85.7	86.9
Shikra [2]	87.0	90.6	81.6	87.4	82.3	82.2	85.2
Ferret-v2 [20]	87.5	91.3	80.8	87.4	83.9	84.8	86.0
InternVL2-8B [3]	87.1	91.1	79.8	87.9	82.7	82.7	85.2
VisionReasoner [14]	88.6	90.6	**83.6**	87.9	86.1	87.5	87.4
Dr. Seg (Ours)	89.4	91.5	**83.6**	88.0	**88.1**	**88.0**	**88.1**

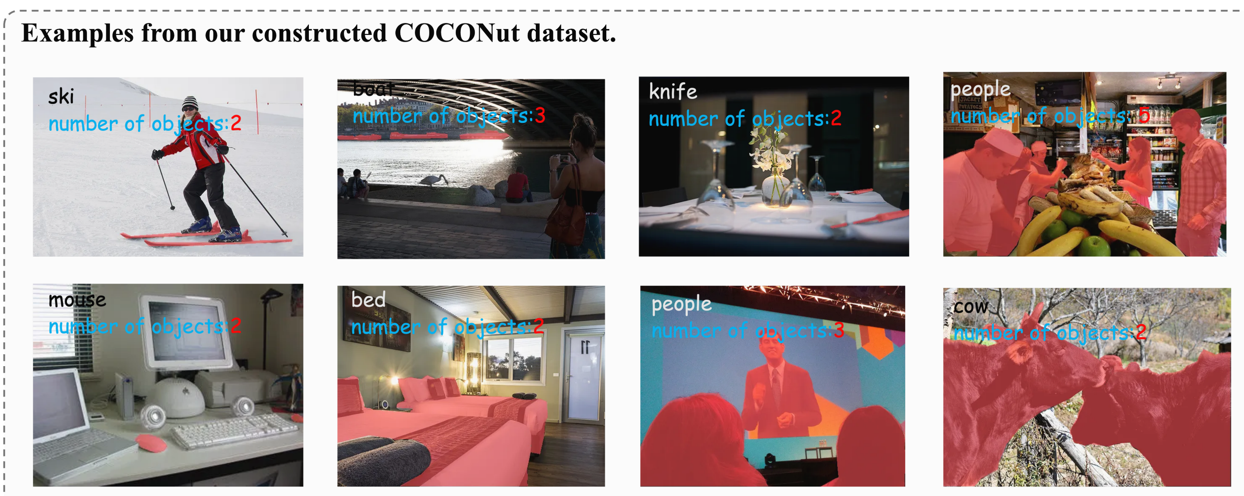

Figure 4. Visualization of our COCONut dataset, illustrating diverse object categories and a wide range of object counts per image.

Examples of our method compared with previous.

Origin image | Visionreasoner | Dr. Seg (ours)

Question: "someone with bare arms"

VisionReasoner: <think> The person in the image is the one with bare arms, which is the biker riding the bike in the foreground. The other biker riding the purple bike is wearing a t-shirt and jeans, which might not necessarily be bare arms, but still, they also could be a close match. The two people visible are the two bikers.</think><answer>[{"bbox_2d": [262,83,385,300], "point_2d": [283,179]}, {"bbox_2d": [363,135,553,528], "point_2d": [417,289]}]</answer>

Dr. Seg (Ours): <think> [The person with bare arms is the one riding the motorcycle in the foreground. They are wearing a sleeveless shirt, which suggests they have bare arms. The other person is wearing a helmet and appears to be moving away from the camera, wearing a full jacket, which might cover their arms partially.] <look> [The person in the foreground is clearly not wearing a shirt, showing their bare arms, which aligns with the description of having bare arms. The other person is distanced and dressed differently, with a helmet, which makes it less likely for them to have bare arms.] </look> There might be another potential match for someone with bare arms in the background, but observing their attire and distance from the camera, they do not seem to fit the description as well.] </think><answer>[{"bbox_2d": [263,84,385,301], "point_2d": [289,183]}]</answer>

Figure 5. A representative failure case of prior methods. The model initially attends to the correct person but drifts away as the chain-of-thought unfolds, leading to over or under counting the target instances. We attribute this behavior to excessive noise in the reward design: collapsing raw rewards into binary signals discards information and introduces noise. This noise is then further amplified when heterogeneous metric scales are summed together.

Examples of our method compared with previous.

Origin image	Visionreasoner	Dr. Seg (ours)

Question: " the person with the highest status"

<think> [The person with the highest status in a Tibetan Buddhist context, typically a monk, is often dressed in yellow robes, which can be seen in the image. However, it's important to consider the context and the surrounding statues, as the individual seated prominently might also have a significant position within the community. The statue behind the group, though, appears to be a Buddha statue, which could imply that the individual is not a monk but a religious leader or a respected figure in the Buddhist community.] <look> [The yellow-robed figure is positioned behind the group, suggesting a higher status within the context of Tibetan Buddhism, which may not match the individual identified as having the highest status within the image.] </look> Therefore, the person wearing the yellow robe might be the one with the highest status.] </think><answer>[{"bbox_2d": [318,159,481,472], "point_2d": [410,304]}]</answer>

Question: " What object in the picture can be used to display the names and prices?"

<think> [The image shows a bunch of fish with yellow labels on them, which contain prices. These labels are the most relevant objects for displaying names and prices of different fish. They are typically used in fish markets to provide information to potential buyers. The rest of the fish and the table they are placed on are not directly used to display the information.] <look> [The labels are rectangular yellow tags with black text that contain numerical values presumably representing prices, and they are directly visible on the fish, which are the primary elements being sold. The labels are designed to quickly convey important information to the audience in the context of a fish market.] </look> The labels can be used to display the names and prices of different fish in a fish market setting.] </think><answer>[{"bbox_2d": [2,139,107,265], "point_2d": [49,207]}, {"bbox_2d": [101,190,222,363], "point_2d": [167,288]}, {"bbox_2d": [210,147,327,258], "point_2d": [270,209]}, {"bbox_2d": [335,172,474,290], "point_2d": [404,235]}, {"bbox_2d": [499,142,647,289], "point_2d": [582,234]}, {"bbox_2d": [662,165,787,233], "point_2d": [730,200]}, {"bbox_2d": [339,49,448,137], "point_2d": [399,95]}, {"bbox_2d": [309,367,465,590], "point_2d": [392,485]}, {"bbox_2d": [629,375,740,518], "point_2d": [698,450]}, {"bbox_2d": [180,456,315,688], "point_2d": [248,569]}]</answer>

Figure 6. Visualization of the proposed Dr. Seg compared with VisionReasoner.

Examples of our method compared with previous.

Origin image | Visionreasoner | Dr. Seg (ours)

Question: " Some people enjoy having a view of the outside world while still being comfortable inside their homes. What part of the house in the picture can provide such an experience?"

<think> [The part of the house with a bay window allows people to enjoy a view of the outside world while being comfortable inside. The bay window design in the image suggests that it is a larger window that protrudes from the wall, providing ample natural light and an expansive view of the surroundings. The presence of ivy and the curved design of the window emphasize the aesthetic appeal of this feature, making it a pleasant and functional space for homeowners.] <look> [The window and the view outside would be the prominent objects for someone to enjoy the outside while remaining inside. The bay window, with its curved elements, can provide a unique and interesting view for the inhabitants, while the greenery outside suggests a pleasant environment to view.] </look> The most closely matched object in the image is the window.</think><answer>[{"bbox_2d": [362,164,505,569], "point_2d": [434,357]}, {"bbox_2d": [528,180,649,576], "point_2d": [596,396]}, {"bbox_2d": [209,171,334,571], "point_2d": [275,340]}]</answer>

Question: " someone who is most likely to be the player"

<think> [The person wearing a black shirt, white pants, and a helmet is likely the player because they are dressed appropriately for a baseball player and are positioned on the field where players typically are. The other individuals, while also present in the scene, do not have the typical attire or position of a baseball player, indicating that the one with the helmet is the most likely player.] <look> [The visual focus is on the person with the helmet, who is wearing a baseball outfit and is positioned as a potential player, which suggests they are the one most likely to be the player in the game. The other individuals either do not have the necessary attire or are not in the correct position on the field to be considered players at this moment.] </look> The person with the helmet and baseball outfit is most likely the player.] </think><answer>[{"bbox_2d": [422,42,631,763], "point_2d": [573,275]}]</answer>

Figure 7. More visualization of the proposed Dr. Seg compared with VisionReasoner.

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, 4

[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. 4

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 4

[4] 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 vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 2

[5] Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*, 2025. 3

[6] 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. 2

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[8] 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. 2

[9] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[10] 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. 2

[11] 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. 4

[12] 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. 4

[13] 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. 2

[14] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 1, 2, 4

[15] 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. 2

[16] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 1

[17] 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. 2, 4

[18] 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. 2

[19] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer, 2016. 2

[20] 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. 4

[21] 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. 1