Perception-R1: Pioneering Perception Policy with Reinforcement Learning

En Yu[1,¶], Kangheng Lin[2,¶], Liang Zhao[3,¶],
Jisheng Yin[3], Yana Wei[4], Yuang Peng[5], Haoran Wei[3], Jianjian Sun[3],
Chunrui Han[3], Zheng Ge[3], Xiangyu Zhang[3], Daxin Jiang[3], Jingyu Wang[2], Wenbing Tao[1†]
[1]Huazhong University of Science and Technology
[2]Beijing University of Posts and Telecommunications [3]StepFun
[4]Johns Hopkins University [5]Tingshua University
`{yuen, wenbingtao}@hust.edu.cn`

Abstract

Inspired by the success of DeepSeek-R1, we explore the potential of rule-based reinforcement learning (RL) in MLLM post-training for perception policy learning. While promising, our initial experiments reveal that incorporating a thinking process through RL does not consistently lead to performance gains across all visual perception tasks. This leads us to delve into the essential role of RL in the context of visual perception. In this work, we return to the fundamentals and explore the effects of RL on different perception tasks. We observe that the *perceptual perplexity* is a major factor in determining the effectiveness of RL. We also observe that reward design plays a crucial role in further approaching the upper limit of model perception. To leverage these findings, we propose ***Perception-R1***, a scalable RL framework using GRPO during MLLM post-training. With a standard Qwen2-VL-2B-Instruct, Perception-R1 achieves $+\mathbf{4.2}\%$ on RefCOCO+, $+\mathbf{17.9}\%$ on PixMo-Count, $+\mathbf{4.2}\%$ on PageOCR, and notably, $\mathbf{31.9}\%$ **AP** on COCO2017 `val` for the first time, establishing a strong baseline for perception policy learning. Project code is available at `https://github.com/linkangheng/PR1`.

1 Introduction

"We do not see the world as it is, but as we are — or as we are conditioned to see it."

Stephen R. Covey

The landscape of large language model (LLM) has undergone a paradigm shift from non-reasoning foundation model, *e.g.*, GPT-4/4o [44, 19], DeepSeek-V3 [33], to strongly reasoning model, *e.g.*, OpenAI o1/o3 [45], DeepSeek-R1 [12], and Kimi-1.5 [57]. DeepSeek-R1, in particular, introduced a simple yet effective rule-based reinforcement learning (RL) approach [55], enabling emergent reasoning patterns without relying on traditional scaffolding techniques such as Monte Carlo Tree Search (MCTS) [17, 67] or Process Reward Models (PRM) [31]. This has catalyzed a new revolution in LLM post-training techniques, prompting researchers to develop more powerful reasoning language models [42, 24].

Despite these advancements, current explorations predominantly focus on the purely linguistic domain, and the unimodal nature of these reasoning models limits their ability to engage with the world in a truly perceptive way. To bridge this gap, this work takes a pioneering step in exploring

†Corresponding author, ¶ Core contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

the potential of *perception policy learning* within multimodal LLMs [61, 3] from lens of RL. While transferring RL techniques with reasoning processes, *i.e.*, chain-of-thought [66], from the language domain shows promise on certain visual tasks, our empirical studies reveal that this approach is not universally effective. This inevitably prompts us to reexamine the *role that RL play in visual perception tasks, and how the utilization of RL can lead to better and scalable perception policy.*

Current understanding of reinforcement learning as a post-training technique is primarily rooted in linguistic tasks [24] and language-centric multimodal tasks [10]. This work, however, posits that *perception* is a critical prerequisite for visual reasoning. We argue that *only by fully unlocking the perceptual patterns of Multimodal LLMs (MLLMs) can these models achieve complex visual reasoning.* Visual perception tasks, fundamentally distinct from natural language tasks, necessitate a revised understanding of RL in this context due to two unique properties:

- *Visual perception is grounded in the objective physical world.* It possesses definite physical truth values (*e.g.*, points, lines, bounding boxes) but lacks the semantic depth of language.
- *Visual perception tasks, such as visual grounding and counting, are often "single-step" direct predictions.* This limits the structured reasoning search space typically explored by RL.

These characteristics indicate that applying RL to visual perception will yield different properties, motivating our exploration of a *perception-first* RL cognition. This work investigates the RL post-training of MLLMs in visual perception, complementing and extending current understanding. Through extensive experimental analysis, we have identified several bitter yet valuable lessons.

- *Explicit thinking process (CoT) during RL is not necessary for current perception policy.* (§ 5.2) We observe that the model without thinking process performs better than the one with thinking process.
- *Reward design plays a pivotal role in perception policy learning.* (§ 5.3) An appropriate reward function will lead to a healthier learning curve and explore stronger perceptual patterns of MLLM.
- *Perceptual perplexity determines RL superiority over SFT.* (§ 5.2) We observe that RL can bring more significant improvement compared to SFT on more complex visual tasks, *e.g.*, object detection.

Driven by these findings, we present a simple, effective, and scalable RL framework, *i.e.*, ***Perception-R1***, for efficient perception policy learning. Inspired by mainstream language reasoning models [12, 57], Perception-R1 applies rule-based RL algorithm GRPO [55] during MLLM post-training stage. With a vanilla Qwen2-VL-2B-Instruct [61], Perception-R1 achieves significant improvement on multiple visual perception benchmarks, *e.g.*, $+\mathbf{4.2}\%$ on RefCOCO+ [40], $+\mathbf{17.9}\%$ on PixMo-Count [13], and $+\mathbf{4.2}\%$ F1-score on PageOCR [34]. More importantly, Perception-R1 serves as the first time to enable a pure MLLM to reach $\mathbf{31.9}\%$ mAP on the object detection benchmark COCO2017 [32] `val`, showcasing the great potential of general foundation models to surpass expert models in mainstream visual tasks. We hope our method, results, and analysis will inspire future research on perception policy learning with RL.

2 Related Works

Multimodal Foundation and Reasoning Models. Recently, vision-language models [37, 3, 73, 70] have demonstrated remarkable capabilities in visual comprehension [64, 68] and generation [14, 48] through large-scale pretraining [2, 61] and visual instruction tuning [37, 35]. These models integrate visual modalities into a unified semantic space via visual encoders [49] and adapters [11, 37], while leveraging auto-regressive large language models [59, 1] as decoders for output generation. Despite the advancements in multimodal foundation models, their visual reasoning capabilities remain in an early developmental stage. Recent approaches [8, 39, 41] have explored reinforcement learning (RL) post-training to enhance visual reasoning. However, they primarily focus on language-centric tasks such as ambiguous reference resolution [39] and geometric problem-solving [41], while overlooking critical aspects of perception-driven reasoning. In this work, we take a pioneering step in utilizing RL for perception policy learning, aiming to bridge this gap and advance multimodal reasoning.

Visual Perception in Multimodal Models. Visual perception, a core concept in computer vision [21, 52, 20, 69, 29], involves interpreting and understanding sensory (visual) information from the real world. In the context of MLLMs, visual perception is crucial for integrating, comprehending, and reasoning about visual data from images or videos. Current MLLMs typically bolster their visual perception by employing advanced visual architectures [63, 64], optimized visual-language modeling strategies [70, 68], and sophisticated post-training techniques [74]. This work explores the potential of reinforcement learning (RL) to further enhance these visual perception capabilities.

RL-based Post-training in LLMs and MLLMs. Reinforcement learning (RL) has emerged as a pivotal paradigm for refining LLMs through alignment with human preferences and task-specific objectives. Prominent approaches like Reinforcement Learning from Human Feedback (RLHF) [46] and Direct Preference Optimization (DPO) [50] have demonstrated remarkable success in enhancing safety, coherence, and instruction-following capabilities of LLMs [43, 47, 44] and MLLMs [74, 60]. Recently, rule-based RL techniques, represented by GRPO [55], have demonstrated the potential for large-scale RL applications. LLMs have officially entered the era of strongly reasoning models. Subsequently, MLLMs [8, 39, 41] have also quickly followed this technology. However, so far, there has been no exciting, true "Aha Moment" in the multimodal domain. This study aims to investigate the potential contributions of RL to multimodal models, focusing on visual perception.

3 Preliminaries

Perception Policy Definition. The goal of perception policy in visual-language context is enabling the model to first (***i***) extract and understand visual information from the environment [37, 68], then (***ii***) perform logical reasoning based on this understanding [73, 70] to (***iii***) accomplish specific tasks and further interact with the environment [5, 22]. In this work, we aim to empower the model to deal with a series of pure visual, *e.g.*, *counting, detection*, and visual-language, *e.g.*, *grounding, optical character recognition (OCR)*, tasks through perception policy learning.

Group Relative Policy Optimization (GRPO [55]) is a rule-based reinforcement learning algorithm tailored for post-training LLMs. Its core idea is to use group relative rewards to optimize the policy, eliminating the need for a separate critic model [54]. Specifically, GRPO samples multiple outputs ($\mathbf{o_1} \sim \mathbf{o_g}$ in Figure 1) from the old policy for the same input, calculates the average reward of these outputs as the baseline, and uses the relative rewards to guide policy updates. The optimization objective of GRPO can be formulated as following:

$$
\begin{aligned}
\mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)]} \\
&\quad \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta^{i,t}}{\pi_{\theta_{\text{old}}}^{i,t}}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta^{i,t}}{\pi_{\theta_{\text{old}}}^{i,t}}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right\},
\end{aligned}
$$

$$
\mathbb{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] = \frac{\pi_{\text{ref}}\left(o_{i,t}|q, o_{i,<t}\right)}{\pi_\theta\left(o_{i,t}|q, o_{i,<t}\right)} - \log\frac{\pi_{\text{ref}}\left(o_{i,t}|q, o_{i,<t}\right)}{\pi_\theta\left(o_{i,t}|q, o_{i,<t}\right)} - 1, \tag{1}
$$

where ϵ and β are hyper-parameters, and $\hat{A}_{i,t}$ is the advantage, computed using a group of rewards $\{r_1, r_2, \cdots, r_G\}$ corresponding to the outputs within each group. Refer to [12, 55] for more details.

4 Perception-R1

In a nutshell, our Perception-R1 applies the rule-based RL algorithm GRPO [55] to the post-training stage of MLLM and optimizes the reward modeling to support perception policy learning. Figure 1 illustrates the idea, more approach and implementation details introduced next.

4.1 Rule-based Reward Modeling

The reward function serves as the principal training signal in reinforcement learning (RL), directing the optimization process. Existing LLM methods [12, 57, 24] basically apply a highly resilient, rule-based reward system consisting of only two reward types: Format Reward and Answer Reward.

Format Reward. In existing LLM and MLLM, the output format is comprised of two essential components: the final output format and the intermediate reasoning process format. The reward for the final output is defined in accordance with specific task requirements and is typically encapsulated within `<answer></answer>` tags, whereas the reward for the intermediate reasoning process generally mandates that the reasoning steps be enclosed within `<think></think>` tags. Formally,

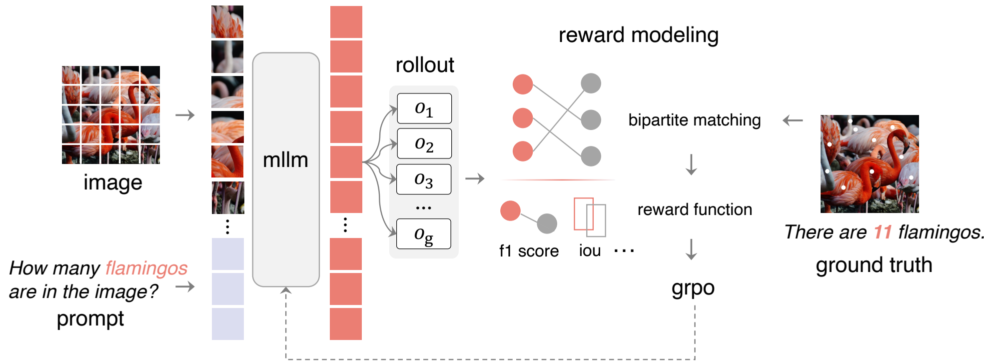


Figure 1: **Illustration of Perception-R1 framework.** Following DeepSeek-R1 [12], we prompt MLLM model to generate several rollout responses, conduct reward modeling, and then apply GRPO [55] during post-training stage.

$$S_{\text{format}} = \begin{cases} 1, & \text{if format is correct} \\ -1, & \text{if format is incorrect} \end{cases} \tag{2}$$

In Perception-R1, we follow this setting. A subtle difference emerges that visual perception task frequently require the output of object coordinates, *e.g.*, bounding box, lines, or points. Consequently, the output format must be strictly constrained to the `[x1,y1,x2,y2]` structure.

Answer Reward. The Answer Reward pertains to the correctness of model-generated responses, serving as a central consideration in reward design. Typically, outputs from language models are abstract and semantically rich, requiring validation through external mechanisms such as code-based ADE [12] or mathematical answer verification [55]. In contrast, visual perception tasks benefit from clearly defined physical ground truths, which simplify the development of a robust reward function.

Perception-R1 diverges from LLM approaches by anchoring the reward mechanism in visual discrimination. This departure is pivotal, as it replaces the often implicit and subjective feedback mechanisms typical of language models with an explicit, quantifiable metric. Formally, discriminative reward r_i can be represented as:

$$r_i = \Phi(o_i, z), \tag{3}$$

where $\Phi(\cdot)$ indicates the discriminative function, for example, IoU for bounding box and euclidean distance for point. By leveraging visual discrimination, we provide the model with a clear and objective feedback signal, ensuring the model's policy update with precise measured margin.

4.2 Multi-Subject Reward Matching

In natural environments, physical objects rarely appear in isolation and instead frequently co-occur in groups. This inherent complexity gives rise to a challenge we define as *reward matching*, which entails aligning the model's output with the corresponding ground truth before reward computation. Specifically, when prompting the model to predict the attributes of multiple subjects within an image, *e.g.*, points and bounding box, it is necessary to determine the appropriate ground truth reference for each subject to ensure accurate reward assignment.

Formally, let $y = \{y_i\}_{i=1}^{N}$denote the set of predicted attributes for N subjects, and let $z = \{z_j\}_{j=1}^{M}$ represent the corresponding ground truth attributes.We model the reward matching problem as a bipartite graph matching task, where one set of nodes corresponds to predictions and the other to ground truths. The edge weight between a prediction y_i and a ground truth t_j is determined by the reward function $\Phi(y_i, z_j)$ defined in Eq. 3, which measures their similarity or compatibility. The objective is to find the optimal assignment that maximizes the total reward:

$$\hat{\sigma} = \underset{\sigma \in \Omega_N}{\arg\max} \sum_{i=1}^{N} \Phi\left(y_i, z_{\sigma(i)}\right), \tag{4}$$

where Ω_N is the set of all valid assignments between predictions and ground truths. To solve this optimization problem efficiently, we employ the Hungarian algorithm [27], a well-established method for bipartite graph matching that guarantees the optimal pairing by maximizing the overall reward (or equivalently, minimizing the cost). This ensures that each predicted attribute is accurately matched with its corresponding ground truth, thereby optimizing the reward computation process.

After the optimal reward assignment is determined, we calculate the answer reward by aggregating the individual rewards for each subject. Mathematically, the overall reward score is defined as:

$$\begin{aligned} S_{\text{answer}} &= \frac{1}{N} \sum_{i=1}^{N} \Phi\left(y_i, z_{\hat{\sigma}(i)}\right), \\ S_{\text{total}} &= S_{\text{format}} + S_{\text{answer}} \end{aligned} \tag{5}$$

where $\hat{\sigma}$ is the optimal assignment obtained via the Hungarian algorithm.In Perception-R1, we primarily use reward matching for visual counting and object detection tasks, as these involve multiple objects.

4.3 Perception-R1 Configuration

Model Setting. Our model implementation follows Qwen2-VL [61]. We mainly use the Qwen2-VL-Instruct-2B as the baseline model.We also utilize Qwen2.5-VL-3B-Instruct [3] for training object detection tasks, due to its specialized optimization for localizing bounding boxes. The input image resolution for Qwen2-VL is dynamic cooperated with 2D-RoPE [56].

Task and Data Setting. Given that Perception-R1 is primarily oriented towards pure visual and visual-language tasks, we select several mainstream and representative downstream tasks for perception policy learning, specifically including *visual grounding*, *e.g.*, refCOCO [71] / + [71] / g [40], *OCR*, *i.e.*, PageOCR [34], *visual counting*, *i.e.*, Pixmo-Count [13], and *object detection*, *i.e.*, COCO2017 [32]. For each task, a subset ($5k \sim 10k$) of samples are respectively extracted as base data for individual post-training. More details are in the appendix A.1.

Training Setting. We focus on the RL-based post-training stage of MLLM. All the selected base models have already undergone pre-training and SFT stage. During RL stage, the initial learning rate is set as $1e-6$ with 8 rollouts by default and a batch size of 1. The following are some important hyper-parameters during post-training. Prompts detailed settings are in the appendix A.1.

Gradient Accmulation	Rollout G	KL Coefficient	Max Response Len	Temperature
2	8	0.04	2048	1.0

Reward Setting. We tailor distinct discriminative rewards for various visual perception tasks. For the grounding task, the reward is based on the Intersection over Union (IoU) between the predicted output and the ground truth. In the counting task, we adopt a paradigm similar to Qwen2.5-VL, which first detects points and then counts them. Here, the reward is derived from the Euclidean distance computed during reward matching. For OCR, the edit distance serves as the primary reward metric. Lastly, in object detection, we combine multiple rewards: an object number reward based on the F1 score, a location reward using IoU, and a binary classification reward with a missing penalty.

Sampling Setting. Following Kimi-1.5 [57], we adopt a curriculum sampling strategy that begins with easier data and gradually transitions to more challenging examples. Specifically, for the object detection task, we first conduct offline training on the COCO dataset to compute reward values. Based on the selected rewards, *i.e.*, number reward, we partition the dataset accordingly. As training advances, we progressively replace the data with more difficult samples (*i.e.*, those associated with lower rewards) while concurrently increasing the rollout to broaden the model's exploration space.

Table 1: **Visual grounding benchmark evaluation.** To comprehensively assess the model's grounding capability, we select referring experssion comprehension (REC) benchmark, *i.e.*, RefCOCO [71], RefCOCO+[71], and RefCOCOg[40] for evaluation. The expert model is denoted in gray.

		RefCOCO											
method	size	val$_{@50}$	testA$_{@50}$	testB$_{@50}$	val$_{@75}$	testA$_{@75}$	testB$_{@75}$	val$_{@95}$	testA$_{@95}$	testB$_{@95}$	val$_{Avg}$	testA$_{Avg}$	testB$_{Avg}$
MDETR [25]	-	87.5	90.4	82.6	-	-	-	-	-	-	-	-	-
OFA [62]	-	88.4	90.6	83.3	-	-	-	-	-	-	-	-	-
LLaVA-1.5 [35]	7B	49.1	54.9	43.3	10.7	13.6	6.9	0.4	0.3	0.3	20.1	22.9	16.8
LLaVA-NeXT [36]	7B	82.5	88.4	74.0	45.7	54.8	35.6	1.9	2.6	0.7	43.4	48.6	36.8
LLaVA-OV [28]	7B	73.0	82.3	63.5	24.2	29.6	15.9	0.5	0.5	0.5	32.6	37.5	26.6
Qwen2-VL [61]	2B	86.8	89.6	82.0	77.2	80.6	70.1	33.0	35.7	26.9	65.7	68.6	59.7
Perception-R1	2B	**89.1**	**91.4**	**84.5**	**79.5**	**83.6**	**72.4**	**35.0**	**38.5**	**28.8**	**67.9**	**71.2**	**61.9**

		RefCOCO+											
method	size	val$_{@50}$	testA$_{@50}$	testB$_{@50}$	val$_{@75}$	testA$_{@75}$	testB$_{@75}$	val$_{@95}$	testA$_{@95}$	testB$_{@95}$	val$_{Avg}$	testA$_{Avg}$	testB$_{Avg}$
MDETR [25]	-	81.1	85.5	72.9	-	-	-	-	-	-	-	-	-
OFA [62]	-	81.3	87.1	74.2	-	-	-	-	-	-	-	-	-
LLaVA-1.5 [35]	7B	42.4	49.7	36.4	9.8	12.4	6.4	0.5	0.5	0.2	17.6	20.8	14.3
LLaVA-NeXT [36]	7B	74.5	84.0	64.7	41.5	51.8	30.0	1.9	2.7	1.0	39.3	46.2	31.9
LLaVA-OV [28]	7B	65.8	79.0	57.2	23.6	28.8	15.3	0.6	0.6	0.4	30.0	36.1	24.3
Qwen2-VL [61]	2B	77.1	82.5	70.1	68.7	73.8	60.0	29.4	32.3	23.0	58.4	62.9	51.0
Perception-R1	2B	**81.7**	**86.8**	**74.3**	**73.6**	**79.3**	**64.2**	**32.6**	**36.9**	**26.7**	**62.6**	**67.7**	**55.1**

		RefCOCOg							
method	size	val$_{@50}$	test$_{@50}$	val$_{@75}$	test$_{@75}$	val$_{@95}$	test$_{@95}$	val$_{Avg}$	test$_{Avg}$
MDETR [25]	-	83.3	83.3	-	-	-	-	-	-
OFA [62]	-	82.2	82.3	-	-	-	-	-	-
LLaVA-1.5 [35]	7B	43.2	45.1	8.5	9.3	0.3	0.3	17.3	18.2
LLaVA-NeXT [36]	7B	77.5	77.1	40.7	39.9	1.8	1.7	40.0	39.6
LLaVA-OV [28]	7B	70.8	70.8	23.3	23.6	0.6	0.7	31.6	31.7
Qwen2-VL [61]	2B	83.3	83.1	72.7	73.0	28.9	27.9	61.6	61.3
Perception-R1	2B	**85.7**	**85.4**	**75.7**	**76.0**	**32.1**	**33.1**	**64.5**	**64.8**

Table 2: **PageOCR evaluation,** compared with various strong expert and general models. "en" means English and "zh" means Chinese.

		Edit Distance ↓		F1-score ↑		Precision ↑		Recall ↑		BLEU ↑		METEOR ↑	
	size	en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
Nougat [4]	250M	25.5	-	74.5	-	72.0	-	80.9	-	66.5	-	76.1	-
DocOwl1.5 [23]	7B	25.8	-	86.2	-	83.5	-	96.2	-	78.8	-	85.8	-
GOT [65]	580M	3.5	3.8	97.2	98.0	97.1	98.2	97.3	97.8	94.7	87.8	95.8	93.9
Qwen2-VL [61]	2B	8.0	10.0	94.4	93.0	96.9	96.1	93.0	90.5	90.9	78.0	94.1	87.2
LLaVA-NeXT [36]	7B	43.0	-	64.7	-	57.3	-	88.1	-	47.8	-	58.2	-
Perception-R1	2B	**2.8**	**8.4**	**98.2**	**96.9**	**98.6**	**97.2**	**97.8**	**96.7**	**96.6**	74.7	**98.1**	**93.8**

5 Experiments

The experimental section evaluates Perception-R1's performance on visual perception tasks (§ 5.1), followed by analytical experiments exploring reinforcement learning (RL)'s role in perception policy learning (§ 5.2). Finally, it discusses the interplay between visual perception and RL, along with key insights for perception policy learning (§ 5.3).

5.1 Performance Landscape in Perception Tasks

We evaluate Perception-R1 on mainstream perception tasks: visual grounding, counting, OCR, and object detection. Experiments use the datasets described in § 4.3 and benchmarks for image understanding. Results are in Tables 1–4.

Visual Grounding is a task that involves localizing visual objects based on linguistic descriptions. Specifically, given a language prompt, the model is required to output the spatial coordinates of the subject (typically a single entity) described in the prompt. As shown in Table 1, we evaluate

Table 3: **Mainstream visual tasks evaluation** including (a) visual object counting and (b) challenging general object detection. Notably, the results of expert model in (b) are copied from MMDetection [7]. † means Perception-R1 for object detection is build based on Qwen2.5-VL-3B-Instruct [3].

		Visual Counting	
method	size	$\text{Pixmo}_{\texttt{val}}$	$\text{Pixmo}_{\texttt{test}}$
LLaVA-1.5 [35]	7B	33.3	31.0
LLaVA-1.6 [58]	7B	32.7	31.9
LLaVA-OV [28]	7B	55.8	53.7
Qwen2-VL [61]	2B	60.2	50.5
Perception-R1	2B	**78.1**	**75.6**

(a) **Visual counting evaluation** on Pixmo-Count [13] `val` set and `test` set.

			Object Detection		
method	size	epoch	AP	AP_{50}	AP_{75}
YOLOv3 [51]	-	273	27.9	49.2	28.3
Faster-RCNN [52]	-	12	35.6	55.7	37.9
DETR [6]	41M	500	42.0	62.4	44.2
Qwen2.5-VL [3]	3B	1	16.1	23.7	16.7
Perception-R1†	3B	1	**31.9**	**46.7**	**33.4**

(b) **Object detection evaluation** on COCO2017 [32] `validation` set.

Table 4: **General image understanding and reasoning evaluation,** compared with various baselines. We select 8 mainstream multimodal benchamrks, *i.e.*, MMBench [38], MMVet [72], MMStar [9], ScienceQA [53], SeedBench [18], MME [16], LLaVA-Bench [37], and ai2D [26] for the comprehensive understanding. We use the model after RL training in the counting tasks for the eval.

		MMBench	MMVet	MMStar	ScienceQA	SeedBench	MME		LLaVA-Bench	AI2D
	llm	Avg	Avg	Avg	Avg	Avg	Cognition	Perception	Avg	Avg
LLaVA1.5 [35]	Vicuna1.5-7B	62.8	32.8	32.6	65.4	60.1	302.1	1338.3	52.6	51.9
LLaVA-NeXT [36]	Vicuna1.5-7B	66.0	37.9	37.7	68.2	69.1	195. 7	1419.5	52.7	67.4
Qwen2-VL [61]	Qwen2-2B	**71.9**	45.6	**46.3**	**74.0**	72.7	418.5	1471.1	46.5	71.6
Perception-R1	Qwen2-2B	71.8	**48.9**	45.7	73.4	**73.0**	**430.0**	**1473.9**	**58.2**	**71.8**

Perception-R1 on three mainstream benchmarks, refCOCO / + / g, and report Acc@0.5, Acc@0.75, and Acc@0.95 to comprehensively assess its visual grounding capability. We surprisingly find that several SoTA MLLMs exhibit poor performance on the more challenging Acc@0.95 metric, with scores even below 1%. In contrast, Perception-R1 achieves a stable performance of over 30% on this metric. This observation suggests that the community should prioritize reporting more discriminative results in future evaluations. The experimental results demonstrate that Perception-R1 exhibits strong competitiveness compared to both specialized and general-purpose models.

Optical Character Recognition (OCR) represents a critical task in visual perception due to its substantial practical value. Current methodologies predominantly adopt either expert models or fine-tuned generalist models for OCR. Perception-R1 pioneers the utilization of RL to further unlock the OCR capabilities of MLLM. As shown in Table 2, our proposed Perception-R1 achieves SoTA performance on the highly challenging OCR benchmark, *i.e.*, PageOCR [34], demonstrating significant superiority over existing expert models, *e.g.*, GOT ($\mathbf{98.2}\%$ vs. $\mathbf{97.2}\%$ F1-score) and robust generalist models, *e.g.*, LLaVA-NeXT ($\mathbf{98.2}\%$ vs. $\mathbf{64.7}\%$ F1-score). Notably, Perception-R1 does not use the Chinese OCR data for training so it is a zero-shot performance for Chinese metric. This breakthrough substantiates the formidable potential of RL applications in OCR tasks, establishing new frontiers for enhancing textual understanding and recognition in complex visual environments.

Visual Counting, as a fundamental vision task, necessitates models to accurately quantify category-specific instances within images, requiring robust *visual logic* to identify and enumerate targets through structured recognition patterns. In Perception-R1, we adopt a detect-then-count paradigm that reformulates the counting problem into a point detection process. As shown in Table 3a, Perception-R1 achieves remarkable counting performance, surpassing the current strong baselines by a substantial margin ($\mathbf{17.9}\%$ improvement compared to Qwen2-VL in Pixmo `val` set). This advancement substantiates that RL effectively stimulates models to explore intrinsic *visual logic* mechanisms (Although counting yields deterministic results, the sequence of counting can exhibit distinct patterns), thereby enhancing their capacity to resolve complex vision tasks.

General Object Detection, widely regarded as the crown jewel of computer vision tasks, has long been considered one of the most challenging problems in visual perception. As a pioneering endeavor to integrate RL into object detection, Perception-R1 achieves a groundbreaking milestone, **serving as the first pure MLLM to surpass the 30+ AP threshold, *i.e.*, $\mathbf{31.9}$ AP in Table 3b, on the COCO 2017 `val` set**, matching or even exceeding the performance of specialized expert models. This achievement underscores rule-based RL's immense potential in addressing complex vision tasks requiring sophisticated visual-logic integration.

Table 5: **Ablation Study of Perception-R1.** We perform ablation studies to investigate key properties of Perception-R1 across a range of visual perception tasks. Specifically, we report the Acc@0.5 for RefCOCO / + / g `val` set, the F1-score for PageOCR, the average scores for Pixmo-Count, and the AP metric for COCO2017 `val` set. **w/o** means without. Notably, there is no reward matching applied to visual grounding and OCR tasks, as these tasks do not involve the multi-subject reward.

	Visual Grounding			OCR	Visual Counting		Detection
case	RefCOCO	RefCOCO+	RefCOCOg	PageOCR	$\text{Pixmo}_{\texttt{val}}$	$\text{Pixmo}_{\texttt{test}}$	COCO2017
Perception-R1	89.1	81.7	85.7	98.2	78.1	75.6	31.9
w/o reward matching	-	-	-	-	77.1	75.4	23.5
w/o RL	86.8	77.1	83.3	98.2	60.2	50.5	16.1
w thinking	75.1	67.9	71.3	93.8	74.9	72.8	25.7
w/o thinking	**89.1**	**81.7**	**85.7**	98.2	**78.1**	**75.6**	28.1
RL only	**89.1**	**81.7**	**85.7**	98.2	**78.1**	**75.6**	**31.9**
SFT only	88.2	80.7	84.6	97.2	58.0	59.9	25.9
SFT+RL	88.4	80.7	85.1	**98.3**	77.1	75.4	30.8

Table 6: **Reward design analysis of Perception-R1.** cls reward indicates binary classification reward and missing reward is a penalty to penalize missed detections. To facilitate rapid experimentation, we randomly sampled 10k data from COCO2017 `train` set for this experiment.

	COCO2017		
reward function	AP	AP_{50}	AP_{75}
format reward	-	-	-
format reward + location reward (IoU)	18.8	25.3	20.1
format reward + location reward (IoU) + cls reward	20.2	27.3	21.4
format reward + location reward (IoU) + cls reward + recall reward (F1)	27.6	42.0	28.7
format reward + location reward (IoU) + cls reward + recall reward (F1) + missing reward	**28.1**	**42.0**	**29.6**

General Visual Comprehension extends beyond pure perceptual tasks, and we evaluate Perception-R1 on multiple multimodal benchmarks. As shown in Table 4, we observe an intriguing phenomenon that models trained with RL for vision-specific tasks, *e.g.*, counting task, exhibit concurrent performance gains in generic comprehension benchmarks. We attribute this cross-task enhancement to the perception policy learning, which drives the model to discover superior image interpretation patterns.

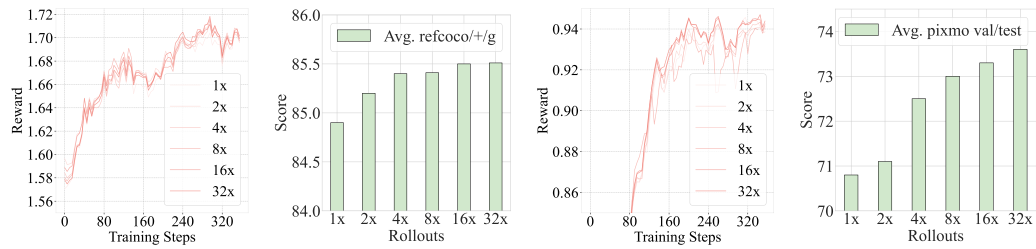
5.2 Ablation Study of Perception-R1

In this section, we aim to conduct a comprehensive ablation study to systematically investigate the contributions of critical components within Perception-R1. Experimental results are shown in Table 5. From the experimental results, we can derive three principal empirical findings:

Reward matching enhances the explorability of multi-subject visual perception. As evidenced by the comparative results between row 1 and 2 in Table 5, replacing the bipartitle matching with sequential matching leads to substantial performance degradation in both visual counting and object detection task. This suggests that sequential matching constrains the RL exploration space. On the contrast, the bipartite matching mechanism provides more possibility in reward assignment, enabling the model to explore optimal visual perception patterns.

Explicit thinking processes prove non-essential for contemporary visual perception. Comparative analysis of row 3 and 4 reveals consistent performance degradation across all four evaluated perception tasks when incorporating an explicit thinking process during both training and inference phases. Similar phenomenon also emerges in image classification tasks [30]. We posit that this phenomenon arises because *current visual perception tasks are more oriented toward visual logic rather than semantic logic.* This shift implies that explicit language-centric reasoning processes are unnecessary, as models tend to focus more on learning ***implicit*** visual patterns.

Perceptual perplexity dictates RL's advantage over SFT. A comparison of post-training methods (SFT, RL, and SFT+RL) across four perception tasks (Table 5, rows 6-8) reveals that RL offers superior performance enhancement in tasks with high perceptual perplexity, such as counting and multi-object/category detection. Conversely, for low-perplexity tasks like grounding and OCR, RL performs comparably to or even worse than SFT. This suggests that high perceptual perplexity is a significant factor influencing RL's effectiveness, indicating that RL techniques are most beneficial for tasks with greater perceptual complexity and a larger exploration space for the perception policy. Further analysis of perceptual perplexity is provided in the appendix A.2.



(a) Grounding reward (b) Grounding performance (c) Counting reward (d) Counting performance

Figure 2: **Scalability analysis of Perception-R1.** We select two primary tasks: grounding and counting. We visualize the training reward curves under varying numbers of rollouts and evaluate the final performance of each task. All experiments are conducted with $5k$ sampled data. And the default rollout number setting ($1\times$) is 8.

5.3 More In-depth Analysis

In this section, we explore several key properties of Perception-R1 to further enhance our understanding of Perception Policy Learning with RL.

Analysis of reward design for perception policy learning. We introduced the details of reward function of Perception-R1 in § 4.3. In this part, we examine the influence of these reward functions on perception policy learning. Specifically, using object detection as a case study, we incrementally integrate the designed answer reward into the format reward, as illustrated in Table 6. The results indicate that the progressive introduction of refined reward functions leads to consistent improvements in detection performance, ultimately exceeding the performance of expert models. This underscores the critical role of reward design in perception policy learning. Furthermore, it identifies a promising avenue for future research: *the development of more refined and task-specific reward functions to enhance perception policy learning.*

Analysis of scaling up rollout for perception policy learning. The scalability of RL is a key concern of existing LLM post-training. In this part, we analyze the scalability of Perception-R1, focusing specifically on scaling up the number of rollouts. As shown in Figure 2, we conduct rollout-scaling experiments in two tasks: visual grounding and visual counting. The results indicate that increasing rollout count enhances reward optimization and final performance. This demonstrates Perception-R1's strong scaling properties and underscores the critical role of rollout quantity in scaling perception policies. By generating sufficient rollouts, the model broadens its exploration space, increasing the diversity of candidate solutions for reward evaluation. This expansion accelerates convergence to optimal visual perception patterns.

6 Limitation and Conclusion

"What can RL bring to MLLM?" is a public question since the propose of DeepSeek-R1. Several latest works attempt to apply RL from the perspective of language-centric visual reasoning [39, 15, 41]. However, in this paper, we take a different pathway and argue that perception is a crucial prerequisite for visual reasoning. *Only by fully unlocking the perception patterns of MLLMs can the models possess the ability to reason about complex visual tasks.* Nevertheless, we regrettably find that many current perception tasks are overly simplistic, which limits the exploration space for RL. This, in turn, restricts the possibility of MLLMs achieving a perceptual "Aha moment" through thinking process. Finding more appropriate perception tasks, *aka., meta task*, may be the key to addressing this issue.

In a summary, this work takes a pioneering step in exploring the potential of rule-based RL in MLLM post-training for perception policy learning. Through extensive experimental analysis, we establish several valuable cognition about perception policy learning with RL. Driven by these findings, we build ***Perception-R1***, a simple, effective, and scalable RL framework for efficient perception policy learning. Perception-R1 sets new SoTAs across multiple visual perception tasks, particularly in object detection tasks. By introducing a novel paradigm, it achieves and even surpasses the performance of expert models, showing the great potential of perception policy learning with RL.

Acknowledgements

This research is supported in part by the National Key Research and Development Program of China (Grant No. 2023ZD0121300), and National Natural Science Foundation of China (Grant No. 62495092).

This research is also supported by the National Natural Science Foundation of China under Grant 62576144 and Grant 62176096.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[3] 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.

[4] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[6] 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.

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[8] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V`, 2025.

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli

Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[13] 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.

[14] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[17] Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. Interpretable contrastive monte carlo tree search reasoning, 2024.

[18] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[19] GPT-4o. Hello gpt-4o, 2024.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024.

[23] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[24] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. `https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero`, 2025.

[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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

[26] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[27] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

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

[29] Jinyang Li, En Yu, Sijia Chen, and Wenbing Tao. Ovtr: End-to-end open-vocabulary multiple object tracking with transformer. *arXiv preprint arXiv:2503.10616*, 2025.

[30] Ming Li, Shitian Zhao, Jike Zhong, Yuxiang Lai, and Kaipeng Zhang. Cls-rl: Image classification with rule-based reinforcement learning. *arXiv preprint arXiv:2503.16188*, 2025.

[31] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[32] 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.

[33] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[34] Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[38] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[39] 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.

[40] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[41] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[42] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[43] OpenAI. Chatgpt. `https://openai.com/blog/chatgpt`, 2022.

[44] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[45] OpenAI. Learning to reason with llms, September 2024.

[46] 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.

[47] 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.

[48] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*, 2024.

[49] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[50] 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, 2024.

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

[52] 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.

[53] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

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

[55] 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.

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[57] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[58] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[60] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*, 2024.

[61] 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.

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022.

[63] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2024.

[64] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Small language model meets with reinforced vision vocabulary. *arXiv preprint arXiv:2401.12503*, 2024.

[65] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[67] Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search, 2024.

[68] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.

[69] En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023.

[70] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. *arXiv preprint arXiv:2312.00589*, 2023.

[71] 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.

[72] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[73] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023.

[74] Zining Zhu, Liang Zhao, Kangheng Lin, Jinze Yang, En Yu, Chenglong Liu, Haoran Wei, Jianjian Sun, Zheng Ge, and Xiangyu Zhang. Perpo: Perceptual preference optimization via discriminative rewarding. *arXiv preprint arXiv:2502.04371*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope. We assert that this paper pioneers the exploration of rule-based Reinforcement Learning (RL) in Multimodal Large Language Model (MLLM) post-training for perception policy learning. Furthermore, we establish several key insights regarding perception policy learning with RL. Driven by these findings, we developed Perception-R1, which achieves new state-of-the-art (SoTA) results across multiple visual perception tasks.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion of limitation is presented in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the full set of assumptions and a complete (and correct) proof in Section 3 and Section 4 including preliminaries and equations and detailed explanation.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All reproduce information including source code are provided in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the source code and detailed readme in supplemental material and we also provide detailed data source in the appendices.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We present the experimental setting in section 4.3.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars in this work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report it on Section 4.3 and all experiments are conducted on NVIDIA A100 Tensor Core GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work aligns with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work does not have potential negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no high risk for misuse of our models and datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All code and data are following licenses properly. And all related works are cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce new code and models in this paper. Both the code and mode cards are well-documented, with detailed information regarding usage, training procedures, licenses, and limitations provided alongside the assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This work is build based on LLMs and we provide a detailed introduction about them.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Technical Appendices and Supplementary Material

In this appendix, we provide additional details about ***Perception-R1***, which are omitted due to the 9-page limit of the main paper. Specifically, Section A.1 elaborates on the detailed dataset and training settings. Section A.2 presents additional explanation about perceptual perplexity. Section A.3 presents more experimental results.

A.1 Additional Details about Experimental Setting

More detailed dataset information of Perception-R1. In Section 4.3, we introduced what data was used for RL post-training of Perception-R1 on which tasks. In this part, we will provide more detailed information about the datasets, as shown in Table 7.

Table 7: **Training dataset statistics.** Notably, we do not mix the data from different perception tasks for joint training because the rewards for different tasks vary.

tasks	datasets	Original	Used	Ratio
visual grounding	RefCOCO / RefCOCO+ / RefCOCOg	320k	5k	1.56%
OCR	PageOCR	50k	5k	10%
visual counting	PixMo-Count	1.9M	10k	0.5%
object detection	COCO2017	110k	110k	100%
overall	-	2.38M	130k	-

More detailed training setting information of Perception-R1. Section 4.3 elaborates on several key parameters of Perception-R1. In this part, we further demonstrate the diverse prompts employed for distinct perception tasks, as shown in Table 8.

Table 8: **Prompts of Perception-R1.** The system prompt of Perception-R1 follows Qwen2-VL [61] and Qwen2.5-VL [3].

tasks	system prompt	user prompt
visual grounding	Qwen2-VL	Output the bounding box of the {question} in the image.
OCR	Qwen2-VL	OCR this image.
visual counting	Qwen2-VL	Output all the bounding boxes of the {label}
object detection	Qwen2.5-VL	Please output bbox coordinates and names of {90 categories of COCO}.

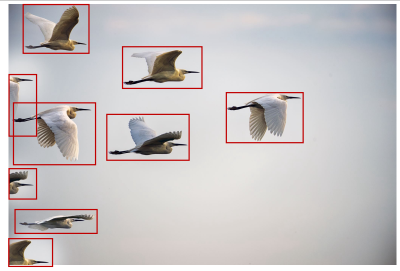
A.2 Additional Explanation about Perceptual Perplexity

In this work, we point out that *perceptual perplexity* is a major factor in determining the effectiveness of RL. And RL techniques are most beneficial for tasks with greater perceptual complexity and a larger exploration space for the perception policy. In this part, we aim to demonstrate the impact of perceptual perplexity across different tasks on the performance of RL through quantitative analysis. Specifically, we utilize the theoretical possibility of matching outcomes in reward matching to quantify the perceptual perplexity of the task, that is, how many theoretical matching results exist between the model's predictions and the ground truth. We also count the probabilities of actual matches across different datasets, that is, the average number of permutations of the ground-truth output per image.

Table 9: **Perceptual perplexity analysis** of different perception tasks. n is the number of ground truth. The statistical value of perplexity is calculated based on sampled data.

tasks	dataset	theoretical	statistical
visual grounding	refCOCO	1	1
OCR	PageOCR	1	1
visual counting	Pixmo-Count	A_n^n	229,202
object detection	COCO2017	A_n^n	3.47e+83

As shown in Table 9, in visual grounding and OCR tasks, there is only one fixed target for the ground truth, leading to just one possible matching outcome. However, for visual counting and

Input:  **Output**: 

Question:

<image>Output all the bounding boxes of the bird.

Perception-R1:

(259,427),(466,591)(298,172),(497,315)(558,344),(759,524)(0,273),(61,466)(30,790),(222,875)(27,9),(208,207)(0,902),(95,998)(0,637),(51,714)(15,389),(215,599) Num boxes = 9

Figure 3: **Demo case of Percpetion-R1** on visual counting task.

Input:  **Output**: 

Question:

<image>Output the bounding box of the woman holding a child in the image.

Perception-R1:

The woman holding a child(157,459),(208,622)

Figure 4: **Demo case of Percpetion-R1** on visual grounding task.

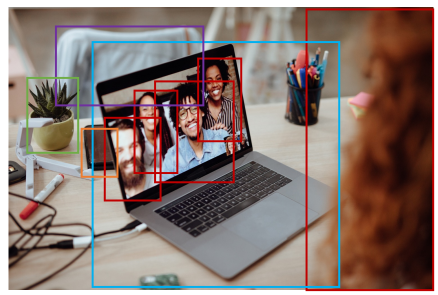
object detection tasks, multiple target objects exist in an image. The various permutations of these targets create numerous matching possibilities, offering ample exploration space for RL to optimize perception capabilities.

A.3 Additional Experimental Results

More qualitative results. In this section, we provide more qualitative analysis of Perception-R1 on multiple visual perception tasks. The selected cases are shown in Figure 3 - 5 including visual counting, visual grounding, object detection, and OCR task, respectively.

Input:

Output:



Question:

<image>Please output bbox coordinates and names of {90 categories of COCO}.

Perception-R1:

```
[
{"bbox_2d": [435, 1, 616, 415], "label": "person"},
{"bbox_2d": [218, 112, 323, 253], "label": "person"},
{"bbox_2d": [147, 165, 225, 280], "label": "person"},
{"bbox_2d": [183, 126, 239, 242], "label": "person"},
{"bbox_2d": [276, 77, 338, 192], "label": "person"},
{"bbox_2d": [127, 54, 471, 406], "label": "laptop"},
{"bbox_2d": [108, 174, 156, 243], "label": "cell phone"},
{"bbox_2d": [26, 107, 103, 214], "label": "potted plant"},
{"bbox_2d": [64, 29, 280, 137], "label": "chair"}
]
```

Figure 5: **Demo case of Percpetion-R1** on general object detection task. The color of bounding boxes correspond to the category.

Input:


G E A R L A B

Riding Dirty

A muddy mix of road- and mountain-bike racing, cyclocross doubles the thrill of both sports. Here's the gear to get you started.

Question:

<image>OCR this image.

Perception-R1:

G E A R L A B\n\nRiding Dirty\n\nA muddy mix of road- and mountain-bike racing, cyclocross doubles the thrill of both sports. Here's the gear to get you started.\n\nby Berne Broudy\n\nCyclocross is as cool as it looks: Racers pedal drop-handlebar bikes with trail-ready tires through a course that often includes pavement, dirt, mud, and grass, as well as obstacles that force you to dismount and sling your bike over your shoulder. "All you really need is to be tough and have a good attitude and a desire to suffer a little bit," says Stu Thorne, founder and director of the Cannondale professional cyclocross team. The right gear helps, too.\n\nBest for Entry Level\n\nCannondale CAADX 105\n\nWhat you sacrifice with extra weight you make up for with a highly capable aluminum frame that'll keep you going between races. A carbon-blade fork helps eat up some of the pounding the courses can dish out. The dependable 2x11 speed Shimano 105 drivetrain powers through any grade, and TRP cable disc brakes perform well even in mud. Meanwhile, the 35cc tires and stable geometry make the CAADX a superfun commuter if you're looking for a bike that can multitask.\n\ncannondale.com $1,970\n\nBest for Experienced Cyclists\n\nSpecialized CruX Elite X1\n\nIf you already log long miles on a premium road bike, this should be your new cross rig. The carbon frame is as light as it is.\n\nA low bottom bracket and relatively slack head angle, which make it stable, helped us sail through rocks and roots and corner quickly. The tires can be run tubeless to better resist flats, and extra clearance means they spin freely when caked with mud. The CruX Elite is playful and fast — and something you won't outgrow as you collect medals.\n\nspecialized.com $3,000\n\nACCESSORIES\n\nCraft Shield Glove\n\nThe cross season typically runs from September through February, so you'll need hearty gloves like these, with a fleece lining and a waterproof base, for warmth on wet race days. craftsports.us $78\n\nDarn Tough Micro Crew\n\nUnlike other bike races, cyclocross requires you to be on foot at times. So light, strong socks are key. These aren't likely to wear out, but Darn Tough will replace them if they do. darntough.com $18\n\nPark Tool Brush Set\n\nThe mud, dirt, and grime that builds up during off-road rides can damage key components. This kit does more than just keep your bike looking fresh; it keeps it healthy, too. parktool.com $80\n\nRapha Arm and Leg Warmer\n\nThese merino layers, which have a bit of Lycra for stretch, peel off easily when the weather warms up. And they dry quickly, whether you sweat profusely or get caught in a sudden squall. rapha.cc From $70\n\nTopeak SmartGauge D2\n\nFor peak performance, adjust your tire pressure to suit the terrain. (On soft trails, lower pressure makes it grip better.) The SmartGauge makes it a snap with a fast, readable result. topeak.com $40

Figure 6: **Demo case of Percpetion-R1** on OCR task.