VisRL: Intention-Driven Visual Perception via Reinforced Reasoning

Zhangquan Chen, Xufang Luo, Dongsheng Li

Tsinghua University

Microsoft

Code available

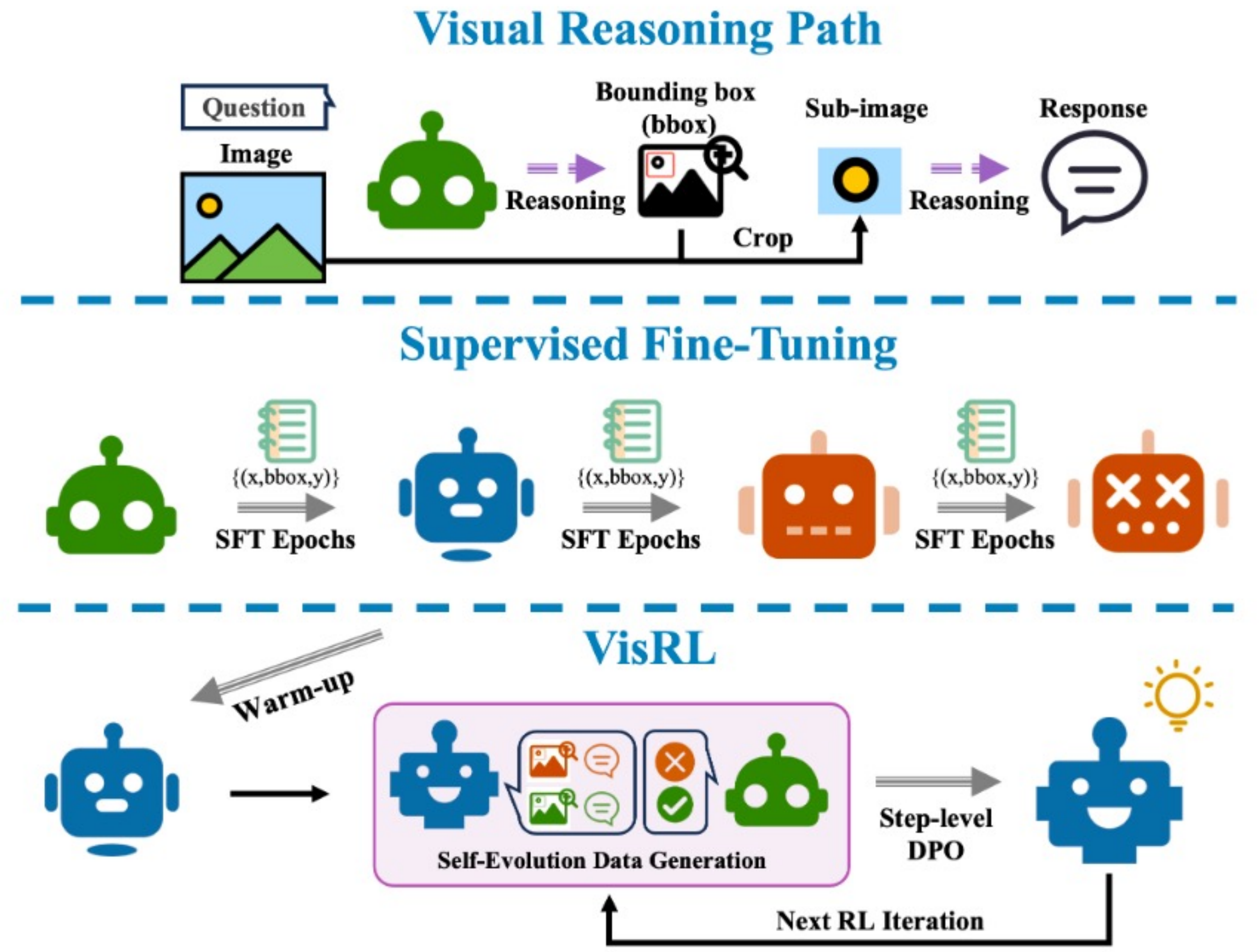
CVPR 2025 · CVPR Nashville JUNE 11-15, 2025

1. Motivation

Goal: Learn **intelligent visual perceptron** from **task feedbacks**

Challenges:	Desiderata :
Lack of intermediate reasoning annotation	Human-like manner
Human trial-and-error learning	Robustness
Model adaptability variance	Free from data-driven supervision
Existence of hallucination	Effectiveness

Solution: To develop an intrinsic reinforced reasoning

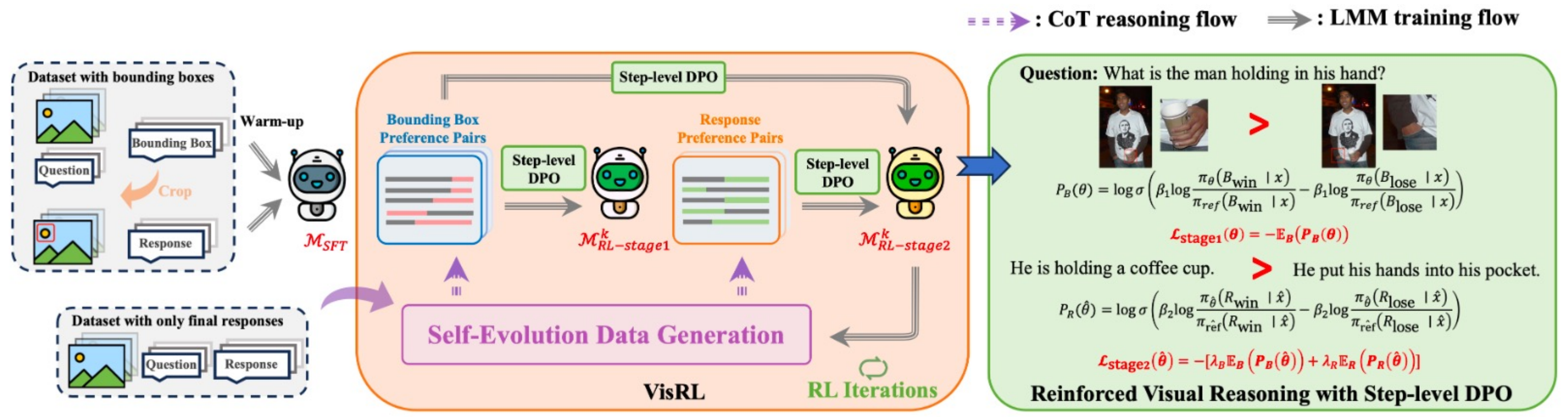


Contributions:

- ***VisRL***: the first RL-based framework for **intention-driven visual perception**, removing reliance on **dense annotations**.
- **Self-Evolution Pipeline:** a novel data generation pipeline, integrating a **diversity controller** and **step-level DPO optimization**.
- **Effectiveness: outperforms** strong baselines and generalizes well.

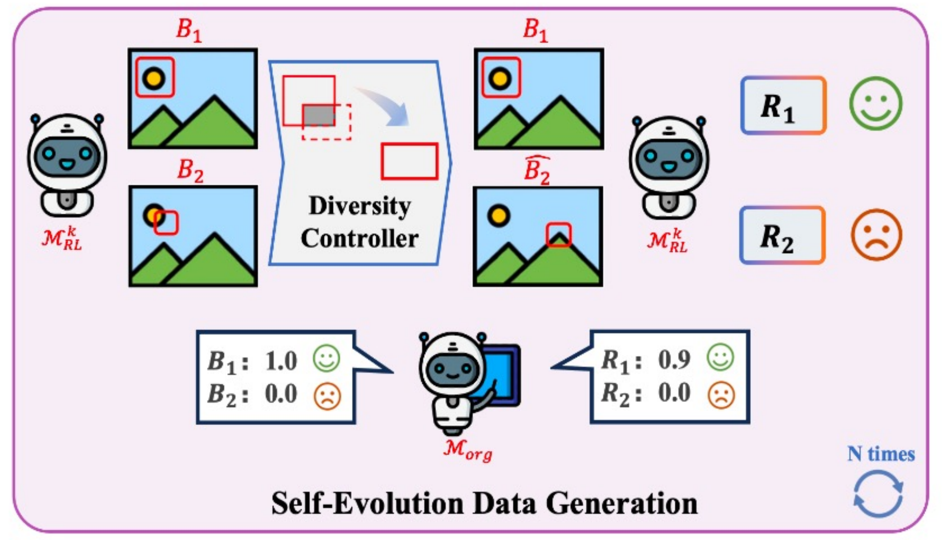
2. Method

Schematic illustration of *VisRL*



We first conduct a small-scale SFT warm-up, then perform RL training on large-scale data without bounding box annotations. The RL phase iterates between **self-evolution data generation** and **step-level DPO optimization**, ensuring reasoning improvement **without external models or annotations**.

Data Generation



VisRL self-evolves by sampling M_{SFT} for diverse CoT data and using M_{org} for self-criticism. This enables intrinsic learning, refining probability distributions without external dependencies.

$$P_{win} = \{p_i \mid s_i^b \geq \mathcal{T}_{\max}^b \text{ and } s_i^r \geq \mathcal{T}_{\max}^r\}$$

$$P_{lose} = \{p_i \mid s_i^b < \mathcal{T}_{\min}^b \text{ and } s_i^r < \mathcal{T}_{\min}^r\}$$

Step-level DPO

VisRL uses a step-level DPO method **in two stages**.

Stage 1: optimizes the bounding box

$$P_B(\theta) = \log\sigma\left(\beta_1 \log\frac{\pi_\theta(B_{win} \mid x)}{\pi_{ref}(B_{win} \mid x)} - \beta_1 \log\frac{\pi_\theta(B_{lose} \mid x)}{\pi_{ref}(B_{lose} \mid x)}\right)$$

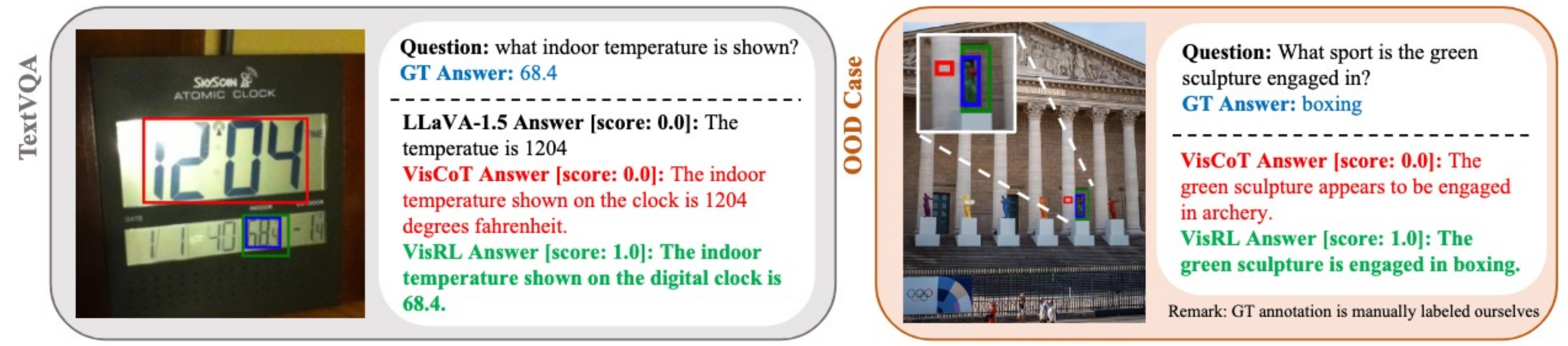
$$\mathcal{L}_{stage1}(\theta) = -\mathbb{E}_{(x, B_{win}, B_{lose}) \sim D_P}(P_B(\theta))$$

Stage 2: optimizes both the bounding box and the final response

$$P_R(\hat{\theta}) = \log\sigma\left(\beta_2 \log\frac{\pi_{\hat{\theta}}(R_{win} \mid \hat{x})}{\pi_{r\hat{e}f}(R_{win} \mid \hat{x})} - \beta_2 \log\frac{\pi_{\hat{\theta}}(R_{lose} \mid \hat{x})}{\pi_{r\hat{e}f}(R_{lose} \mid \hat{x})}\right)$$

$$\mathcal{L}_{stage2}(\hat{\theta}) = -(\lambda_B \mathcal{L}_B(\hat{\theta}) + \lambda_R \mathcal{L}_R(\hat{\theta}))$$

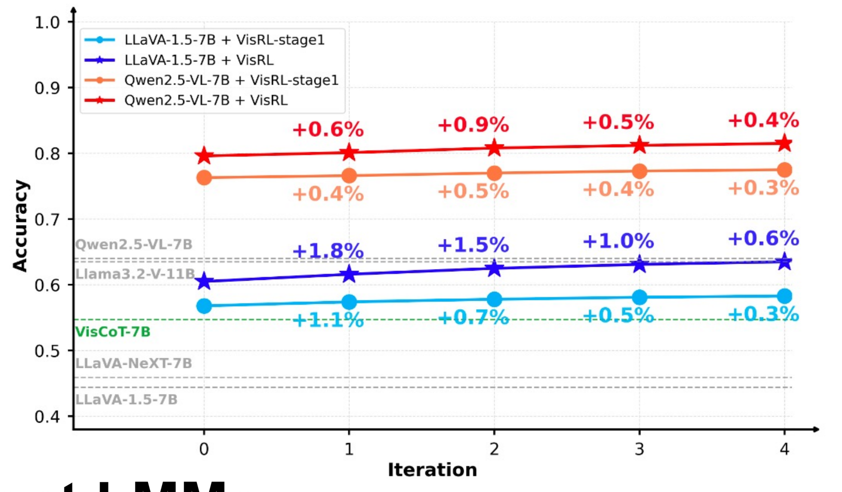
3. Results



Comparison with different baselines

Method	LLM	Vision Encoder	MME	MMBench	POPE	Dataset Num.
LLaVA [B] [34]	Vicuna-7B [13]	CLIP-ViT-L-14-224 [44]	1051.2	34.4	76.5	558K
SEAL [D] [61]	Vicuna-7B	CLIP-ViT-L-14-224	1128.9	33.1	**82.4**	558K + 387K [D]
LLaVA + P2G [T] [8]	Vicuna-7B	CLIP-ViT-L-14-224	1223.0	—	—	558K + 427K [D]
LLaVA + VisRL	Vicuna-7B	CLIP-ViT-L-14-224	1183.8	37.5	78.2	558K + 30K [D]+180K
LLaVA + VisRL– Iter1	Vicuna-7B	CLIP-ViT-L-14-224	**1238.3**	**38.6**	80.4	180K
LLaVA-1.5 [B] [33]	Vicuna-7B	CLIP-ViT-L-14-336	1510.7	64.3	85.8	558K
VisCoT [D] [48]	Vicuna-7B	CLIP-ViT-L-14-336	1453.6	67.9	86.0	558K + 376K [D]
LLaVA-1.5 + VisRL	Vicuna-7B	CLIP-ViT-L-14-336	1526.3	70.1	87.5	558K + 30K [D]+180K
LLaVA-1.5 + VisRL– Iter1	Vicuna-7B	CLIP-ViT-L-14-336	**1560.0**	**71.7**	**88.8**	180K
LLaVA-NeXT [B] [35]	Vicuna-7B-1.5 [72]	CLIP-ViT-L-14-336	1611.1	72.3	—	558K
VisionLLM v2 [D] [60]	Vicuna-7B-1.5	CLIP-ViT-L-14-336	1512.5	77.1	87.5	892K
Insight-V-LLaVA [T] [15]	Vicuna-7B-1.5	CLIP-ViT-L-14-336	1583.9	**81.7**	—	558K + 215K [D]
LLaVA-NeXT + VisRL	Vicuna-7B-1.5	CLIP-ViT-L-14-336	1619.2	78.8	88.4	558K + 30K [D]+180K
LLaVA-NeXT + VisRL– Iter1	Vicuna-7B-1.5	CLIP-ViT-L-14-336	**1637.0**	80.0	**89.3**	180K

VisRL over multiple iterations



Performance on the VisCoT dataset across different LMMs

LMM	Training Phase	Doc/Text DocVQA	Doc/Text TextCaps	Doc/Text TextVQA	Doc/Text DUDE	Doc/Text SROIE	Chart InfogVQA	General VQA Flickr30k	Relation Reasoning GQA	Relation Reasoning Open images	Relation Reasoning VSR	Fine-grained CUB	Avg
LLaVA-1.5-7B [33]	Base (w/o CoT)	0.244	0.597	0.588	0.290	0.136	0.400	0.581	0.534	0.412	0.572	0.530	0.444
	VisCoT [438k] [48]	0.355	0.610	0.719	0.279	0.341	0.356	0.671	0.616	0.833	0.682	0.556	0.547
	SFT [30k]	0.336	0.597	0.715	0.270	0.308	0.336	0.671	0.617	0.833	0.676	0.559	0.538
	SFT+RL1	0.382	0.612	0.724	0.300	0.378	0.406	0.674	0.639	0.838	0.715	0.579	0.568
	SFT+RL1+RL2	**0.419**	**0.641**	**0.759**	**0.394**	**0.411**	**0.497**	**0.675**	**0.666**	**0.848**	**0.748**	**0.598**	**0.605**
LLaVA-NeXT-7B [35]	Base (w/o CoT)	0.431	0.586	0.570	0.332	0.114	0.361	0.525	0.559	0.462	0.594	0.520	0.459
	SFT [30k]	0.423	0.580	0.722	0.330	0.293	0.356	0.589	0.684	0.821	0.767	0.551	0.556
	SFT+RL1	0.474	0.611	0.728	0.373	0.350	0.447	**0.592**	0.707	0.826	0.837	0.573	0.593
	SFT+RL1+RL2	**0.508**	**0.655**	**0.743**	**0.474**	**0.379**	**0.525**	**0.592**	**0.738**	**0.837**	**0.871**	**0.587**	**0.628**
Llama-3.2-V-11B [39]	Base (w/o CoT)	0.797	0.771	0.879	0.588	0.629	0.637	0.601	0.484	0.335	0.589	0.674	0.635
	SFT [30k]	0.776	0.762	0.880	0.584	0.634	0.633	0.712	0.683	0.728	0.720	0.855	0.724
	SFT+RL1	0.811	0.791	0.890	0.599	0.698	0.688	0.724	0.707	0.731	0.738	0.864	0.749
	SFT+RL1+RL2	**0.844**	**0.835**	**0.897**	**0.638**	**0.733**	**0.714**	**0.731**	**0.757**	**0.794**	**0.822**	**0.884**	**0.786**
MiniCPM-o-2.6-8B [66]	Base (w/o CoT)	0.528	0.504	0.548	0.125	0.114	0.220	0.534	0.561	0.462	0.585	0.529	0.428
	SFT [30k]	0.518	0.498	0.551	0.134	0.133	0.239	0.615	0.727	0.789	0.787	0.715	0.519
	SFT+RL1	0.551	0.533	0.561	0.150	0.182	0.286	0.630	0.737	0.799	0.824	0.734	0.544
	SFT+RL1+RL2	**0.596**	**0.600**	**0.565**	**0.209**	**0.251**	**0.353**	**0.639**	**0.793**	**0.870**	**0.864**	**0.756**	**0.591**
PaliGemma2-10B [50]	Base (w/o CoT)	0.017	0.498	0.536	0.129	0.114	0.197	0.529	0.558	0.486	0.543	0.541	0.377
	SFT [30k]	0.110	0.498	0.544	0.134	0.133	0.225	0.611	0.718	0.800	0.770	0.724	0.479
	SFT+RL1	0.169	0.527	0.549	0.163	0.179	0.272	0.621	0.731	0.811	0.822	0.736	0.507
	SFT+RL1+RL2	**0.303**	**0.585**	**0.560**	**0.229**	**0.248**	**0.336**	**0.639**	**0.789**	**0.884**	**0.847**	**0.764**	**0.562**
Yi-VL-6B [67]	Base (w/o CoT)	0.115	0.522	0.551	0.130	0.122	0.205	0.522	0.561	0.468	0.587	0.497	0.389
	SFT [30k]	0.168	0.521	0.598	0.139	0.152	0.247	0.606	0.721	0.772	0.792	0.695	0.492
	SFT+RL1	0.208	0.564	0.610	0.174	0.182	0.294	0.613	0.747	0.799	0.844	0.713	0.523
	SFT+RL1+RL2	**0.318**	**0.611**	**0.627**	**0.234**	**0.280**	**0.358**	**0.620**	**0.804**	**0.853**	**0.871**	**0.726**	**0.573**
Qwen2.5-VL-7B [3]	Base (w/o CoT)	0.836	0.760	0.847	0.606	0.789	0.685	0.601	0.467	0.289	0.581	0.583	0.640
	SFT [30k]	0.807	0.720	0.886	0.580	0.719	0.635	0.630	0.626	0.764	0.782	0.876	0.730
	SFT+RL1	0.842	0.768	0.895	0.600	0.784	0.692	0.642	0.669	0.788	0.822	0.888	0.763
	SFT+RL1+RL2	**0.874**	**0.819**	**0.897**	**0.640**	**0.829**	**0.753**	**0.675**	**0.700**	**0.814**	**0.864**	**0.892**	**0.796**

Referring Expression Comprehension (REC) tasks

Method	Res.	RefCOCO [31] val	RefCOCO [31] test-A	RefCOCO [31] test-B	RefCOCO+ [49] val	RefCOCO+ [49] test-A	RefCOCO+ [49] test-B	RefCOCOg [49] val-u	RefCOCOg [49] test-u
UNINEXT [S] [84]	640^2	92.64	94.33	**91.46**	85.24	89.63	79.79	88.73	**89.37**
G-DINO-L [S] [45]	384^2	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02
OFA-L [G] [74]	480^2	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58
Shikra 7B [G] [10]	224^2	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
MiniGPT-v2-7B [G] [8]	448^2	88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66
Qwen-VL-7B [G] [2]	448^2	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
Ferret-7B [G] [88]	336^2	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
u-LLaVA-7B [G] [83]	224^2	80.41	82.73	77.82	72.21	76.61	66.79	74.77	75.63
SPHINX-13B [G] [40]	224^2	89.15	91.37	85.13	82.77	87.29	76.85	84.87	83.65
VisCoT-7B [62]	336^2	91.77	94.25	87.46	87.46	92.05	81.18	88.38	88.34
LLaVA-1.5-7B [42] + VisRL	336^2	**92.72**	**96.18**	90.21	**90.23**	**94.10**	**85.77**	**91.17**	89.28

Ablation on data generation

	WP-LP	WP-LN	WN-LP	WN-LN	Data Num.
w GPT-4o-2024-11-20	0.00%	65.31%	1.32%	33.37%	**47k**
w SFTed Model	0.00%	54.68%	0.00%	45.32%	3k
w/o Bounding Box Critics	5.42%	31.02%	10.04%	53.51%	86k
w/o Diversity Controller	4.53%	52.02%	4.68%	38.77%	19k
VisRL-Full	0.43%	64.64%	1.64%	33.29%	30k
VisRL-Full-Iter1	0.45%	67.82%	0.82%	30.91%	33k
VisRL-Full-Iter2	**0.47%**	**70.12%**	**0.00%**	**29.41%**	35k