VisRL: Intention-Driven Visual Perception via Reinforced Reasoning

Supplementary Material

In this supplementary material, we provide more technical details and experimental results, including 1) Additional results supplement Tab. 2 and Fig. 4 in the main text in Sec. 1; 2) Detailed descriptions of dataset used in Sec. 2 and Tab. 2; 3) Visual grounding ability tested on REC benchmarks in Sec. 3 and Tab. 3; 4) Our prompts designed for critics of data generation pipeline in Sec. 4; 5) More explanation of diversity controller and the cost in terms of data generation in Sec. 5; 6) More visualization of different datasets from Visual CoT benchmarks in Sec. 6; as well as 7) Limitation of our VisRL in Sec. 7.

1. Additional Results

Tab. 1 provides a detailed performance on the Visual CoT benchmark across various LMMs, serving as a supplementary analysis to Sec. 4.2 and Tab. 2 in the main text. In addition, we further provide two representative cases to illustrate and complement the quantitative analysis as shown in Fig. 1. VisCoT relies on SFT, so its capability is largely limited by the distribution of the training data. Our RL-based approach might be more robust and better generalizes to OOD cases, such as higher resolutions (1257*1553 in Fig. 1 (right)) not seen during training.

2. Dataset

2.1. VisCoT Dataset

We utilize the data from VisCoT [22] and follow its predefined training/testing split. Specifically, a subset of the training set is selected for training our VisRL model, as shown in Tab. 2. Besides, the test set remains consistent with VisCoT, as presented in Tab. 1, Tab. 2, Tab. 4 in the main text, etc..

Text/Doc: There are five text-related datasets—TextVQA [24], DocVQA [18], DUDE [26], TextCaps [23], and SROIE [6], covering text recognition and comprehension in various images and documents.

Fine-Grained Understanding: The Birds-200-2011 dataset (CUB) [27] is a widely used benchmark for fine-grained visual categorization. It includes rich visual data, detailed annotations of bird parts and attributes, and bounding boxes. To leverage this better for LMM, [22] design questions that challenge the model to identify specific bird characteristics, testing its ability to recognize fine-grained details.

General VQA: Flickr30k [21] and Visual7W [34] are used for general VQA tasks. Specifically, Flickr30k provides five captions per image and bounding boxes for most mentioned objects. [22] further use GPT-4 to generate questions focusing on small objects, while Visual7W has already included question-answer pairs with object-level grounding annotations.

Charts: InfographicsVQA [19] dataset features high-resolution infographics, to train LMMs in locating answers precisely.

Relation Reasoning: The Visual Spatial Reasoning (VSR) [12], GQA [7], and Open Images [9] datasets, which are rich in spatial relational information among image objects, are used for relation-reasoning tasks.

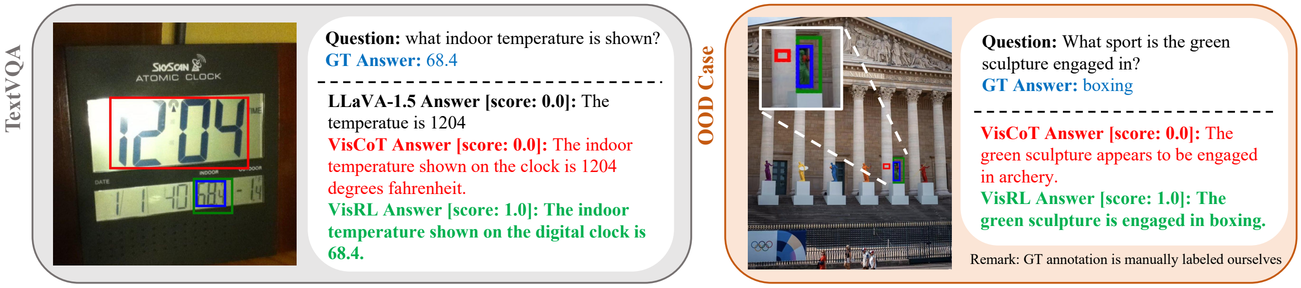
2.2. Comprehensive Benchmarks

We conducted evaluations on three further benchmarks as shown in Tab. 1 in the main text: MME [5], which comprehensively assesses perception and cognitive abilities across 14 sub-tasks; MMBench [16], a systematically designed objective benchmark for the robust and holistic evaluation, covering 20 capability dimensions; and POPE [10], which reframes hallucination evaluation as a series of binary questions requiring the model to determine the presence of objects in an image.

3. Visual Grounding

Furthermore, we conducted additional evaluations of our VisRL on REC benchmarks. Specifically, we tested different methods on RefCOCO [8] and RefCOCO+ [17], both of which were collected in an interactive gaming interface and follow the validation/test-A/test-B split. In these two datasets, test-A always consists of images containing multiple people, whereas test-B includes all other objects. Additionally, compared to RefCOCO, queries in RefCOCO+ do not contain absolute spatial terms, such as references to an object's location within the image (e.g., "on the right side"). RefCOCOg [17] was another dataset collected in a non-interactive setting, and its queries are generally longer than those in RefCOCO and RefCOCO+.

As shown in Tab. 3, VisRL surpasses all previous generalist models, even outperforming models with significantly larger parameters. Moreover, in most of the cases, our method exceeds the performance of previous state-of-the-art specialist models, e.g. G-DINO-L [15] and UNINEXT [30]. This demonstrates the exceptional capability of our approach in accurately predicting bounding boxes. Notably, our model achieves improvements of **1%** to **5%** to VisCoT. "Top-1 Accuracy@0.5," refers to the accuracy of a model in correctly predicting the bounding box as the top-ranked output when the IoU between the predicted and GT bounding boxes is at least 50%.



Question: what indoor temperature is shown?
GT Answer: 68.4

LLaVA-1.5 Answer [score: 0.0]: The temperatue is 1204
VisCoT Answer [score: 0.0]: The indoor temperature shown on the clock is 1204 degrees fahrenheit.
VisRL Answer [score: 1.0]: The indoor temperature shown on the digital clock is 68.4.

Question: What sport is the green sculpture engaged in?
GT Answer: boxing

VisCoT Answer [score: 0.0]: The green sculpture appears to be engaged in archery.
VisRL Answer [score: 1.0]: The green sculpture is engaged in boxing.

Remark: GT annotation is manually labeled ourselves

Figure 1. Visualization of TextVQA (left) and OOD case (right).

Table 1. Performance on the different benchmarks. The amount of dense-labeled CoT data with bounding box annotations used is indicated in []. The **best** results from different LMMs are highlighted.

		Doc/Text					Chart	General VQA	Relation Reasoning			Fine-grained	
LMM	**Training Phase**	DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfogVQA	Flickr30k	GQA	Open images	VSR	CUB	**Avg**
LLaVA-1.5-7B [13]	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] [22]	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 [14]	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 [20]	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 [31]	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 [25]	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 [33]	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 [2]	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**

Table 2. We detail the number of samples used on each dataset during the SFT and RL training stages in terms of Qwen2.5-VL-7B. Specifically, SFT is trained on data with bounding box labels, while RL utilizes only the image-question-answer pairs without any additional annotations. After our preference dataset construction, the RL data is distilled from 180k to 30k samples. Moreover, the datasets used for SFT and RL are independent (no overlap), while RL1 and RL2 share the same training dataset.

Dataset	**Category**	**SFT (w. bounding box CoT)**	**RL (w/o bounding box label)**	**RL (after data generation)**
Flickr30k [21]	General VQA	4000	32500	4626
GQA [7]	Relation Reasoning	4000	30000	5173
InfographicsVQA [19]	Charts	4055	11000	2358
Open Images [9]	Relation Reasoning	3053	40000	5019
TextCaps [23]	Text/Doc	4152	28000	4869
TextVQA [24]	Text/Doc	3524	15000	2458
VSR [12]	Relation Reasoning	1876	1500	998
CUB [27]	Fine-Grained Understanding	1987	2000	1278
Visual7W [34]	General VQA	4000	20000	3267
Total		30647	180000	30046

Table 3. Performance (Top-1 Accuracy@0.5) on Referring Expression Comprehension (REC) tasks. [S] refers to specialist models, while [G] refers to generalist models. The best is **highlighted**, while the second-best is underlined.

Method	Res.	RefCOCO [8]			RefCOCO+ [17]			RefCOCOg [17]	
		val	test-A	test-B	val	test-A	test-B	val-u	test-u
UNINEXT [S] [30]	640^2	92.64	94.33	**91.46**	85.24	89.63	79.79	88.73	**89.37**
G-DINO-L [S] [15]	384^2	90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02
OFA-L [G] [28]	480^2	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58
Shikra 7B [G] [4]	224^2	87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19
MiniGPT-v2-7B [G] [3]	448^2	88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66
Qwen-VL-7B [G] [1]	448^2	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
Ferret-7B [G] [32]	336^2	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
u-LLaVA-7B [G] [29]	224^2	80.41	82.73	77.82	72.21	76.61	66.79	74.77	75.63
SPHINX-13B [G] [11]	224^2	89.15	91.37	85.13	82.77	87.29	76.85	84.87	83.65
VisCoT-7B [22]	336^2	91.77	94.25	87.46	87.46	92.05	81.18	88.38	88.34
LLaVA-1.5-7B [13] + VisRL	336^2	**92.72**	**96.18**	90.21	**90.23**	**94.10**	**85.77**	**91.17**	89.28

4. Instruction for Critics

4.1. Evaluation of Generated Bounding Box

Fig. 2 illustrates how we design the instruction to evaluate the bounding boxes generated by $\mathcal{M}_{SFT}$ based on $\mathcal{M}_{org}$. Specifically, given the VQA data (Q, I, R_{GT}), $\mathcal{M}_{SFT}$ first outputs the bounding box based on Q and I, which is then used to crop the sub-images I^s. Subsequently, we assess the correlation between the generated bounding box and the GT response by prompting (Q, R_{GT}, I^s) to $\mathcal{M}_{org}$ (shown in Fig. 2). Thus, we achieve the evaluated score of bounding box solely based on the GT response, without the need for extra bounding box annotations.

You are responsible for verifying the relevance of the image based on the provided question and standard answer, you need to assess whether the image aligns with the standard answer.
The full score is 1 point and the minimum score is 0 points. Please directly provide the score in JSON format, for example, {{"score": 0.8}}, without showing the intermediate process.
The evaluation criteria is that, the higher score will be if the image effectively encompasses the information provided in the standard answer based on question.

Question: {Question}
Standard answer: {GT_Response}

Figure 2. Prompt for the bounding box critics.

4.2. Evaluation of Generated Response

Fig. 3 presents the evaluation of responses along the sampled paths. Specifically, we prompt the model $\mathcal{M}_{org}$ to assess the generated response with GT response based on the given question/image, and assign the score accordingly.

You are responsible for proofreading the answers, you need to give the score to the model's answer by referring to the standard answer, based on the given question and image.
The full score is 1 point and the minimum score is 0 points. Please directly provide the score in JSON format, for example, {{"score": 0.8}}, without showing the intermediate process.
The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score.

Question: {Question}
Standard answer: {GT_Response}
Model's answer: {Response}

Figure 3. Prompt for the response critics.

5. Data Generation

5.1. Diversity Controller

For the bounding box (b-box) diversity, our controller (Eqn. (3) in the main text) replaces B_2 with a new random box outside the B_1's region when their IoU exceeds the threshold $\mathcal{T}$. For overall diversity, we sample N candidate pairs with two paths in each pair are different, and select the pairs with the lowest and highest scores as the final preference pair. According to results in Tab. 4, we empirically set $N = 5$ for optimal computational efficiency and accuracy. We simply chose a reasonable value for the hyperparameters in data generation without further tuning.

Table 4. Effect of varying N on generated bounding box.

N	3	4	5	6	7
Accuracy	63.01%	64.08%	64.64%	64.68%	64.69%

5.2. Cost

Our data generation pipeline further alleviates the cost of bounding box annotation, both in terms of manual effort and overall annotation expense. 1) Manual annotation can be infeasible due to unclear standards and language ambiguity, while VisRL generates and refines b-boxes and answers autonomously, enabling self-evolution. 2) With $N = 5$ and vLLM acceleration, one A100-80GB GPU generates about 100 preference pairs per minute (1.6 pairs/s), with no extra costs.

6. More visualization

In Fig. 4, 5, 6 and 7, we provide more visualization results of our VisRL compared with VisCoT, while using the same base model – LLaVA-1.5-7B.

7. Limitation

VisRL is not yet optimized for multi-object or ultra-high-resolution scenarios, which may fail in such settings. Additionally, the two-stage reasoning framework also leaves room for improved inference efficiency.

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. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. 2

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478, 2023. 3

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023. 3

[5] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 1

[6] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1516–1520. IEEE, 2019. 1

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019. 1, 2

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787–798, 2014. 1, 3

[9] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International journal of computer vision, 128(7):1956–1981, 2020. 1, 2

[10] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023. 1

[11] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023. 3

[12] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. Transactions of the Association for Computational Linguistics, 11:635–651, 2023. 1, 2

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3

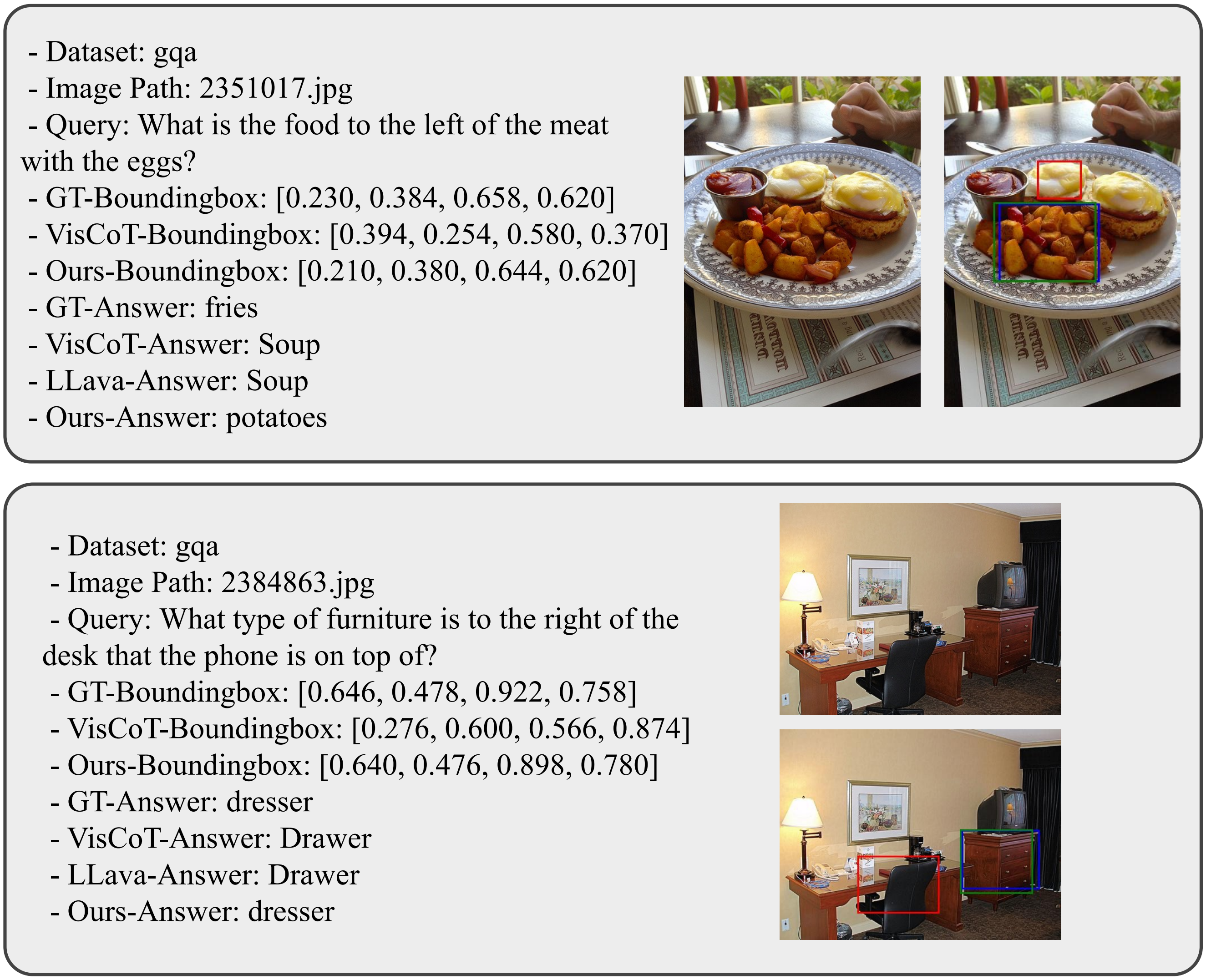
[14] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pages 38–55. Springer, 2024. 1, 3

[16] 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? In European conference on computer vision, pages 216–233. Springer, 2024. 1

[17] 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. 1, 3

[18] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021. 1



- Dataset: gqa
- Image Path: 2351017.jpg
- Query: What is the food to the left of the meat with the eggs?
- GT-Boundingbox: [0.230, 0.384, 0.658, 0.620]
- VisCoT-Boundingbox: [0.394, 0.254, 0.580, 0.370]
- Ours-Boundingbox: [0.210, 0.380, 0.644, 0.620]
- GT-Answer: fries
- VisCoT-Answer: Soup
- LLava-Answer: Soup
- Ours-Answer: potatoes

- Dataset: gqa
- Image Path: 2384863.jpg
- Query: What type of furniture is to the right of the desk that the phone is on top of?
- GT-Boundingbox: [0.646, 0.478, 0.922, 0.758]
- VisCoT-Boundingbox: [0.276, 0.600, 0.566, 0.874]
- Ours-Boundingbox: [0.640, 0.476, 0.898, 0.780]
- GT-Answer: dresser
- VisCoT-Answer: Drawer
- LLava-Answer: Drawer
- Ours-Answer: dresser

Figure 4. More visualization results of LLaVa-1.5 vs. VisCoT vs. VisRL (based on LLaVa-1.5). Ground truth (GT) bounding boxes are shown in blue, VisCoT-generated bounding boxes are shown in red, while Ours-generated bounding boxes are in green.

[19] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022. 1, 2

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

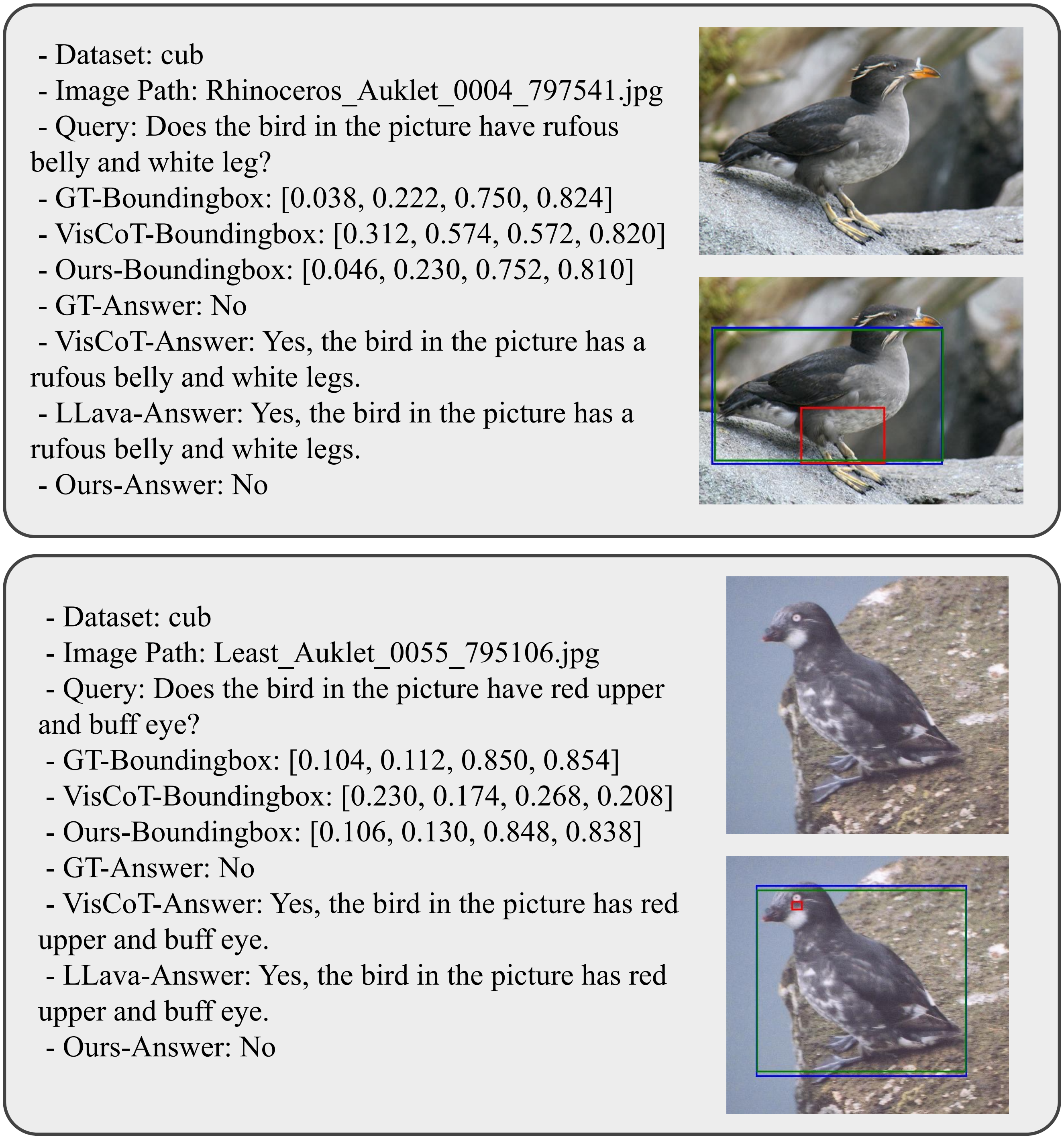
[21] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015. 1, 2

[22] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models, 2024. 1, 2, 3

[23] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 742–758. Springer, 2020. 1, 2

[24] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326, 2019. 1, 2

[25] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang



- Dataset: cub
- Image Path: Rhinoceros_Auklet_0004_797541.jpg
- Query: Does the bird in the picture have rufous belly and white leg?
- GT-Boundingbox: [0.038, 0.222, 0.750, 0.824]
- VisCoT-Boundingbox: [0.312, 0.574, 0.572, 0.820]
- Ours-Boundingbox: [0.046, 0.230, 0.752, 0.810]
- GT-Answer: No
- VisCoT-Answer: Yes, the bird in the picture has a rufous belly and white legs.
- LLava-Answer: Yes, the bird in the picture has a rufous belly and white legs.
- Ours-Answer: No

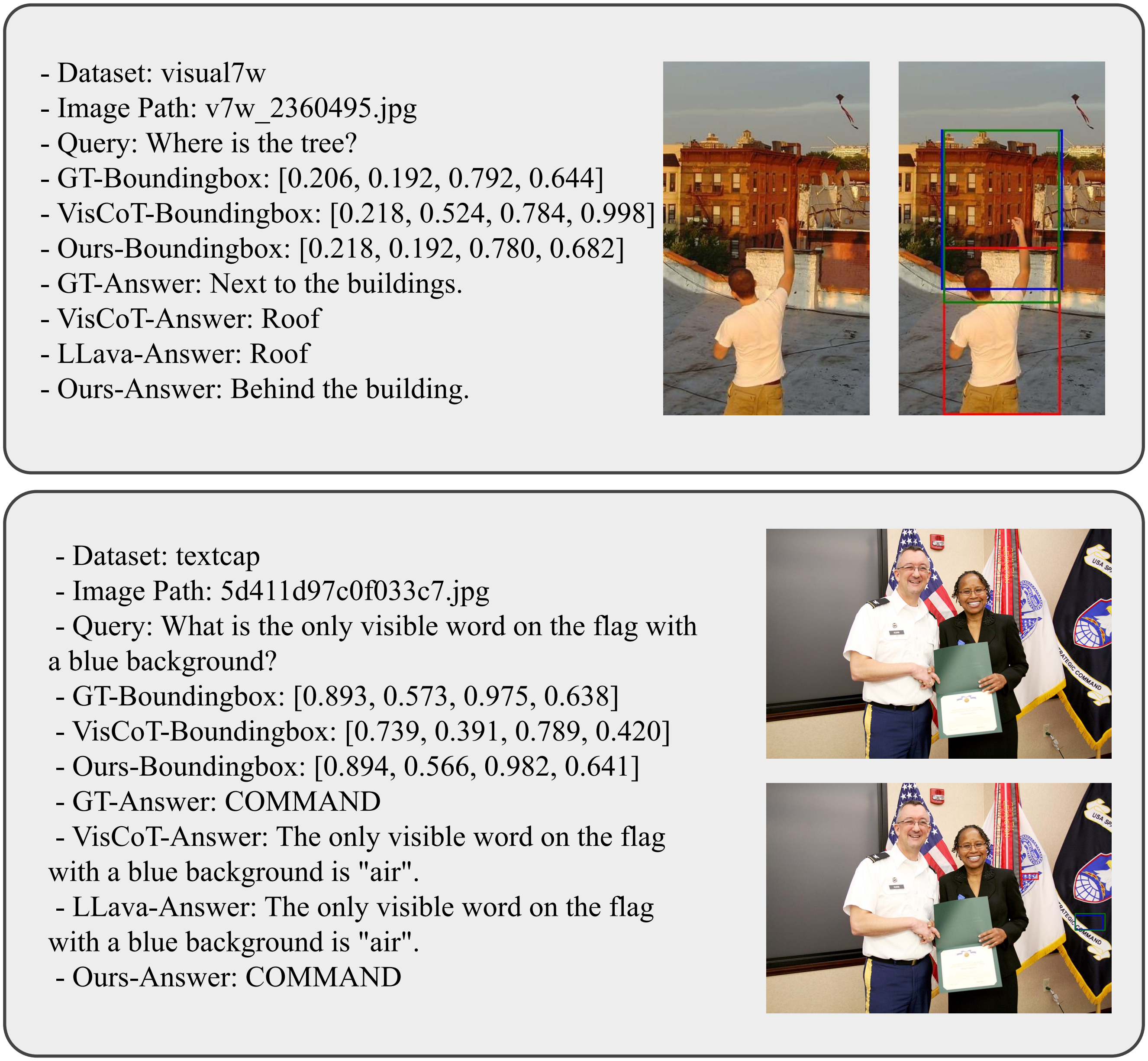
- Dataset: cub
- Image Path: Least_Auklet_0055_795106.jpg
- Query: Does the bird in the picture have red upper and buff eye?
- GT-Boundingbox: [0.104, 0.112, 0.850, 0.854]
- VisCoT-Boundingbox: [0.230, 0.174, 0.268, 0.208]
- Ours-Boundingbox: [0.106, 0.130, 0.848, 0.838]
- GT-Answer: No
- VisCoT-Answer: Yes, the bird in the picture has red upper and buff eye.
- LLava-Answer: Yes, the bird in the picture has red upper and buff eye.
- Ours-Answer: No

Figure 5. More visualization results of LLaVa-1.5 vs. VisCoT vs. VisRL (based on LLaVa-1.5). Ground truth (GT) bounding boxes are shown in blue, VisCoT-generated bounding boxes are shown in red, while Ours-generated bounding boxes are in green.

Long, et al. Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555, 2024. 2

[26] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19528–19540, 2023. 1

[27] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona,



- Dataset: visual7w
- Image Path: v7w_2360495.jpg
- Query: Where is the tree?
- GT-Boundingbox: [0.206, 0.192, 0.792, 0.644]
- VisCoT-Boundingbox: [0.218, 0.524, 0.784, 0.998]
- Ours-Boundingbox: [0.218, 0.192, 0.780, 0.682]
- GT-Answer: Next to the buildings.
- VisCoT-Answer: Roof
- LLava-Answer: Roof
- Ours-Answer: Behind the building.

- Dataset: textcap
- Image Path: 5d411d97c0f033c7.jpg
- Query: What is the only visible word on the flag with a blue background?
- GT-Boundingbox: [0.893, 0.573, 0.975, 0.638]
- VisCoT-Boundingbox: [0.739, 0.391, 0.789, 0.420]
- Ours-Boundingbox: [0.894, 0.566, 0.982, 0.641]
- GT-Answer: COMMAND
- VisCoT-Answer: The only visible word on the flag with a blue background is "air".
- LLava-Answer: The only visible word on the flag with a blue background is "air".
- Ours-Answer: COMMAND

Figure 6. More visualization results of LLaVa-1.5 vs. VisCoT vs. VisRL (based on LLaVa-1.5). Ground truth (GT) bounding boxes are shown in blue, VisCoT-generated bounding boxes are shown in red, while Ours-generated bounding boxes are in green.

and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2

[28] 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. 3

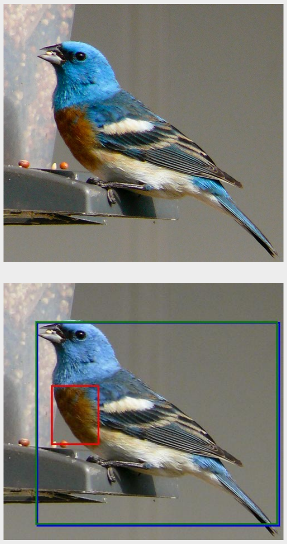
[29] Jinjin Xu, Liwu Xu, Yuzhe Yang, Xiang Li, Fanyi Wang, Yanchun Xie, Yi-Jie Huang, and Yaqian Li. u-llava: Unifying multi-modal tasks via large language model. arXiv preprint arXiv:2311.05348, 2023. 3

[30] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15325–15336, 2023. 1, 3

[31] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. 2

[32] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen

- Dataset: cub
- Image Path: Lazuli_Bunting_0042_14820.jpg
- Query: Does the bird in the picture have brown throat and grey underparts?
- GT-Boundingbox: [0.118, 0.154, 0.986, 0.946]
- VisCoT-Boundingbox: [0.170, 0.398, 0.340, 0.630]
- Ours-Boundingbox: [0.114, 0.150, 0.978, 0.940]
- GT-Answer: No
- VisCoT-Answer: Yes, the bird in the picture has a brown throat and grey underparts.
- LLava-Answer: Yes, the bird in the picture has a brown throat and grey underparts.
- Ours-Answer: No



- Dataset: infographicsvqa
- Image Path: 37029.jpeg
- Query: Which social media platform has 47% mobile only users?
- GT-Boundingbox: [0.269, 0.067, 0.760, 0.361]
- VisCoT-Boundingbox: [0.242, 0.420, 0.349, 0.439]
- Ours-Boundingbox: [0.230, 0.071, 0.768, 0.333]
- GT-Answer: facebook
- VisCoT-Answer: Twitter
- LLava-Answer: Twitter
- Ours-Answer: Facebook

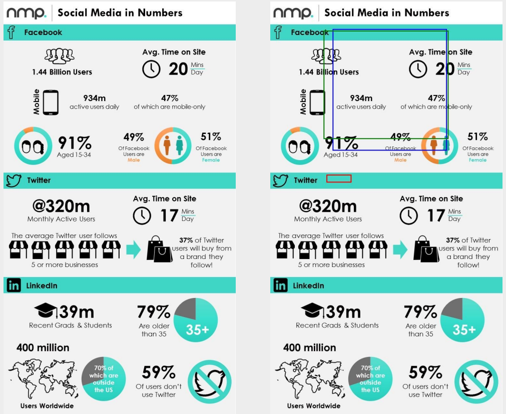

nmp. Social Media in Numbers
Facebook
1.44 Billion Users
Avg. Time on Site
20 Mins Day
Mobile
934m active users daily
47% of which are mobile-only
91% Aged 15-34
49% Of Facebook Users are Male
51% Of Facebook Users are Female
Twitter
@320m Monthly Active Users
Avg. Time on Site
17 Mins Day
The average Twitter user follows 5 or more businesses
37% of Twitter users will buy from a brand they follow!
LinkedIn
39m Recent Grads & Students
79% Are older than 35
35+
400 million
70% of which are outside the US
Users Worldwide
59% Of users don't use Twitter

Figure 7. More visualization results of LLaVa-1.5 vs. VisCoT vs. VisRL (based on LLaVa-1.5). Ground truth (GT) bounding boxes are shown in blue, VisCoT-generated bounding boxes are shown in red, while Ours-generated bounding boxes are in green.

Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704, 2023. 3

[33] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652, 2024. 2

[34] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4995–5004, 2016. 1, 2