Supplementary Material for Visual-RFT: Visual Reinforcement Fine-Tuning

In the appendix, we first provide a detailed overview of all the visual perception tasks' training and testing data involved in our experiments in Sec. A. In Sec. B, we introduce the prompts and formats used for the training data in our experiments. In Sec. C, we provide experimental results and examples of Visual-RFT on several domain-specific and out-of-domain classification and detection datasets. In Sec. D, we provide additional experimental results to complement the length limitations of the main text. These include more shot settings for classification and detection tasks, experimental results of the Qwen2-VL-7B model, and open vocabulary performance on the LVIS dataset. In Sec. E, We present more model reasoning cases across a variety of tasks.

A. Model and Data Sources

In the experimental section of the paper, we evaluated our Visual-RFT on a wide range of visual perception tasks, including fine-grained image classification, few-shot object detection, reasoning grounding, and open-vocabulary object detection. These experiments covered numerous datasets across different tasks, which are listed in Tab. 1.

Flowers102 The Flowers102 [9] Dataset comprises 8,189 images categorized into 102 distinct flower species commonly found in the United Kingdom. Each category includes between 40 to 258 images, capturing a wide variety of poses, lighting conditions, and backgrounds. This diversity introduces significant intra-class variability and inter-class similarity, making the dataset particularly challenging for fine-grained image classification tasks. Researchers frequently utilize this dataset to evaluate algorithms designed for detailed visual distinctions.

Pets37 The Oxford-IIIT Pet Dataset consists of 7,349 images spanning 37 pet breeds, with approximately 200 images per breed. The dataset is evenly divided between cat and dog breeds. Each image is annotated with a breed label, a tight bounding box, and a pixel-level foreground-background segmentation mask. The images exhibit substantial variations in scale, pose, and lighting conditions, providing a robust benchmark for image classification task.

FGVC-Aircraft The FGVC-Aircraft [8] (Fine-Grained Visual Classification of Aircraft) Dataset contains 10,200 images across 102 aircraft model variants. Each image is labeled with the aircraft model variant. The dataset is divided into training, validation, and test sets, each comprising 3,334 images. The high level of detail in the annotations makes this dataset ideal for evaluating models on fine-grained visual classification tasks, particularly in distinguishing between similar aircraft models.

Stanford Cars The Stanford Cars [4] Dataset consists of 16,185 images covering 196 classes of cars, categorized by make, model, and year (e.g., "2012 Tesla Model S"). The dataset is split into 8,144 training images and 8,041 testing images. The images are 360×240 pixels in size and were captured from the rear of the vehicles. This dataset is widely used for fine-grained image classification tasks, especially those focusing on vehicle identification and categorization.

COCO Dataset The COCO [6] dataset is a large-scale benchmark widely used in computer vision for tasks such as object detection, segmentation, keypoint detection, and image captioning. It contains over 200,000 labeled images and covers 80 object categories, providing rich annotations including object bounding boxes, segmentation masks, and keypoints for human pose estimation. COCO is well-known for its challenging settings, featuring complex scenes with multiple objects, diverse backgrounds, and varying lighting conditions. The dataset's image captions are also commonly used to train and evaluate models in image-to-text and visual question answering tasks. The COCO challenge has become a prestigious competition, driving advancements in visual perception models and techniques.

LVIS Dataset The LVIS [1] dataset (Large Vocabulary Instance Segmentation) is designed specifically for large-scale instance segmentation with a focus on long-tail distributions. Unlike COCO, which includes 80 object categories, LVIS features over 1,200 categories, many of which belong to the rare and uncommon class spectrum. It provides pixel-wise segmentation masks, bounding boxes, and detailed annotations, enabling fine-grained recognition and segmentation tasks. The dataset is notable for its balanced representation of both frequent and rare categories, encouraging the development of models that perform well under few-shot and zero-shot learning conditions. LVIS's challenging nature makes it a valuable benchmark for evaluating a model's ability to handle complex real-world scenarios with diverse and nuanced object classes.

Table 1. **Benchmark Sources.** We have included information and links for all the multi-image and single-image benchmarks tested in the paper in the table.

Tasks	Datasets	Evaluation Metric	Val Number	Source
Fine-Grained Classification	Flowers102 [9]	Accuracy	2,463	Flowers102
	Pets37 [10]	Accuracy	3,669	Pets37
	FGVC-Aircraft [8]	Accuracy	3,333	FGVC-Aircraft
	Stanford Cars [4]	Accuracy	8,041	Stanford Cars
	ChestXR	Accuracy	3432	ChestXR
Few-Shot Detection	COCO [6]	mAP	5,000	COCO
	LVIS [1]	mAP	19,809	LVIS
	Monster Girls	mAP	215	MG
Reasoning Grounding	LISA [5]	mIoU, gIOU	200	LISA
Open-Vocabulary Detection	COCO [6]	mAP	5,000	COCO
	LVIS [1]	mAP	19,809	LVIS

Table 2. **Few-shot results on Domain Specific Classification Dadtaset: ChestXR.** We evaluated a complex medical fine-grained classification datasets on ChestXR.

Setting	Models	Average	Nomal	Covid-19	Pneumonia
-	Qwen2-VL-2B	32.3	98.0	0.3	12.4
1-shot	+ SFT	29.3	100.0	0.3	0.2
	+ Visual-RFT	33.2	97.7	0.0	16.2
4-shot	+ SFT	29.0	41.8	35.9	6.3
	+ Visual-RFT	42.3	90.3	8.0	43.4
8-shot	+ SFT	36.8	44.5	28.8	40.5
	+ Visual-RFT	53.9	95.8	42.3	28.5

LISA Dataset The LISA [5] dataset is a novel benchmark designed to evaluate reasoning grounding tasks, which require generating a segmentation mask based on complex and implicit query texts. Unlike traditional segmentation tasks that rely on explicit instructions or predefined categories, LISA challenges models to comprehend and reason with implicit user intentions, often involving intricate reasoning and world knowledge. The dataset comprises over 1,000 image-instruction-mask samples, providing a diverse set of scenarios to test the model's ability to handle complex visual and textual input. LISA's dataset is annotated with implicit text queries that demand advanced reasoning. These queries range from short phrases to long, descriptive sentences, ensuring a robust assessment of a model's reasoning segmentation capability. The dataset is split into training, validation, and test sets, with 239, 200, and 779 samples respectively.

B. Visual-RFT Data Cases

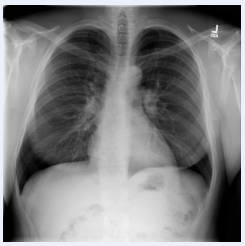
For different visual perception tasks, we trained the model using a small amount of data. To construct this training data, we utilized a set of prompt templates. The prompt templates for different tasks are presented below.

Fine-grained Few-shot Image Classification The prompt we use to guide model for fine-grained few-shot image classification is as follows:

```
This is an image containing a plant.Please identify the species of the plant based on the image.Output the thinking process in <think> </think> and final answer in
```

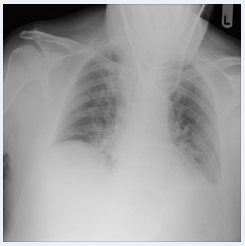
<image> You are a professional medical image analysis AI, specializing in interpreting chest CT images. I will provide a chest X-ray (CT image), and your task is to determine the most appropriate category based on the image content. Here are three possible categories. Please choose the most appropriate one from them:1. pneumonia 2.covid 3. normal
Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.The output answer format should be as follows:
<think> ... </think> <answer>category</answer>
Please strictly follow the format.

Visual RFT



<think> The chest X-ray image shows a normal chest with no obvio us abnormalities such as consolidation, masses, or other pathologies. The lungs appear clear and symmetric. Therefore, the most approp riate category is normal.**</think>**
<answer>normal**</answer>**

Visual RFT



<think> The chest X-ray image shows a patient with a dense, mottl ed appearance in the lungs, which is characteristic of pneumonia. T he presence of a dense, mottled appearance in the lungs is also see n in COVID-19. However, the absence of other typical COVID-19 f eatures such as a diffuse infiltrate or a pleural effusion makes pne umonia more likely. Therefore, the most appropriate category for t his image is pneumonia.**</think>**
<answer>pneumonia**</answer>**

Visual RFT

<think> The chest X-ray image shows a patient with a chest tube a nd a catheter, which suggests that the patient is undergoing a proc edure or treatment. The presence of the chest tube and catheter i ndicates that the patient may have a medical condition that require s monitoring or treatment. Therefore, the most appropriate catego ry for this image is 2. COVID.**</think>**
<answer>covid**</answer>**

Figure 1. **Domain Specific Dataset Classification Case.** We present three inference examples of Visual-RFT on the medical CT image dataset ChestXR in the figure. This dataset is highly specialized and difficult.

<answer> </answer> tags. The output answer format should be as follows: <think> ... </think> <answer>species name</answer> Please strictly follow the format.

Reasoning Grounding The prompt we use to guide model for reasoning grounding follows the template from LISA [5]:

This is an image containing a plant. Please identify the species of the plant based on the image. Output the thinking process in <think> </think> and final answer in <answer> </answer> tags. The output answer format should be as follows: <think> ... </think> <answer>species name</answer> Please strictly follow the format.

Table 3. **Few-shot results on MG dataset of 5 categories.** By introducing out-of-domain data, we increased the difficulty of model recognition and reasoning, further demonstrating the strong generalization ability of Visual-RFT in visual perception tasks.

Models	mAP	bird	feline-or-canid	serpent	slime	wyvern
Qwen2-VL-2B	20.6	12.9	19.8	25.5	18.4	26.4
4-shot						
+ SFT	26.8	19.5	22.4	26.8	33.5	31.8
+ Visual-RFT	61.8	63.9	53.2	70.2	64.5	57.5
Δ	**+41.2**	**+51.0**	**+33.4**	**+44.7**	**+46.1**	**+31.1**
16-shot						
+ SFT	51.3	42.7	44.4	56.4	61.1	52.0
+ Visual-RFT	63.4	59.9	50.8	76.3	71.7	58.1
Δ	**+42.8**	**+47.0**	**+56.4**	**+50.8**	**+53.3**	**+31.7**

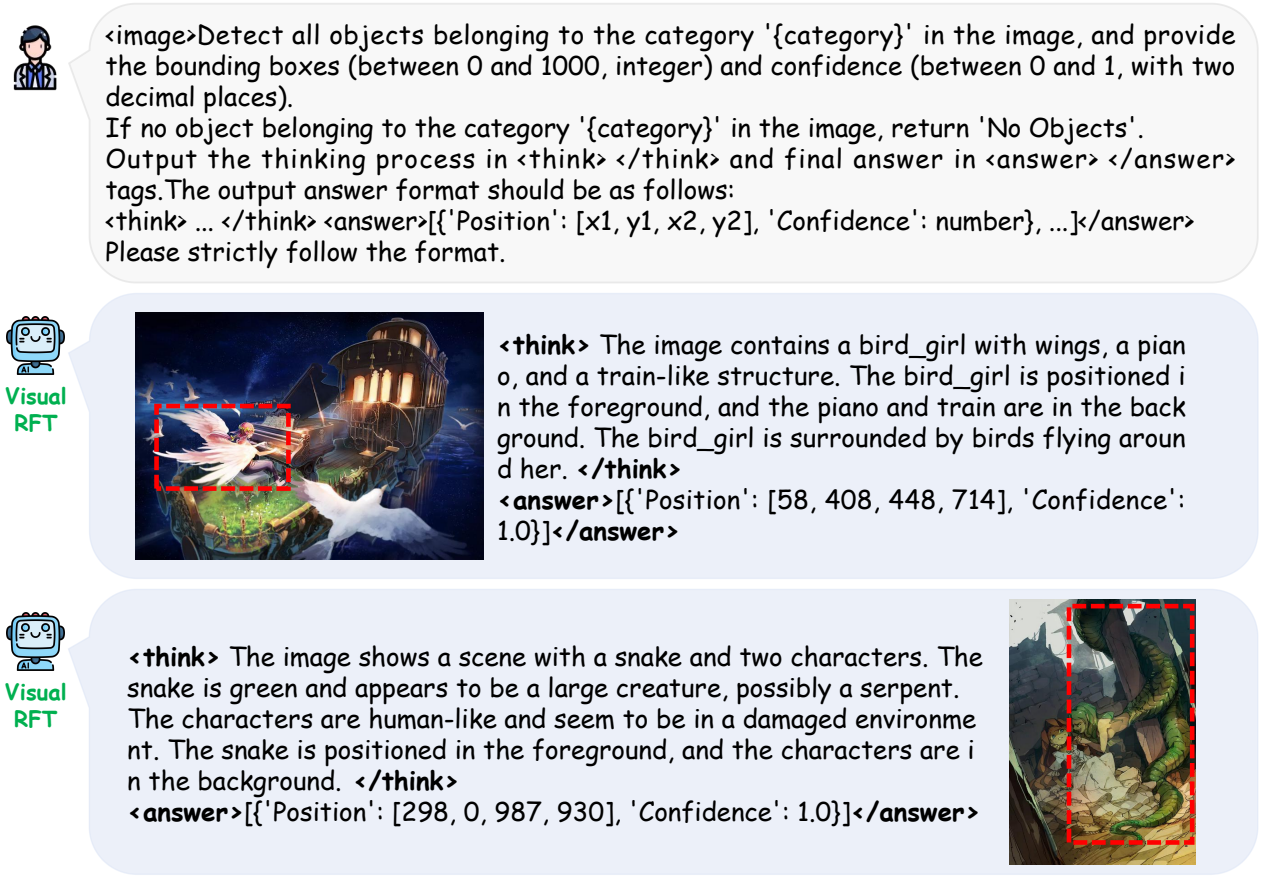


Figure 2. **Domain Specific Dataset Detection Case.** We present three inference examples of Visual-RFT on the MG dataset in the figure. This dataset is out of domain.

Few-shot & **Open Vocabulary Detection** The prompt we use to guide model for few-shot object detection and open vocabulary detection is as follows:

`A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within`

Table 4. **Few-Shot Open Vocabulary Object Detection Results on COCO dataset.** We trained on 65 base categories under 1-shot, 2-shot, 4-shot and tested on 15 novel categories.

Models	mAP_n	mAP_b	mAP_{all}
Qwen2-VL-2B	9.8	6.0	6.7
+ SFT (*1-shot*)	17.1	11.1	12.2
+ Visual-RFT	25.1	15.8	17.5
Δ	**+15.3**	**+9.8**	**+10.8**
+ SFT (*2-shot*)	18.9	12.4	13.6
+ Visual-RFT	29.9	18.9	21.0
Δ	**+20.1**	**+12.9**	**+14.3**
+ SFT (*4-shot*)	19.4	13.2	14.4
+ Visual-RFT	32.0	20.8	22.9
Δ	**+22.2**	**+14.8**	**+16.2**

Table 5. **Open Vocabulary Object Detection Results on LVIS dataset.** We trained on the 65 base categories of the COCO dataset and tested on the 13 rare categories of the LVIS dataset.

Models	mAP	casserole	die	egg roll	futon	garbage	handsaw	hippopotamus	kitchen table	mallet	omelet	shot glass	stepladder	sugar bowl
GroudingDINO-B [7]	23.9	17.1	0.0	2.4	47.5	27.7	13.4	15.2	92.5	0.0	26.6	16.0	41.0	10.7
Qwen2-VL-2B	2.7	1.6	1.2	0.0	2.4	0.0	10.0	0.0	13.4	0.2	4.7	2.1	0.0	0.0
+ SFT	7.6	3.9	21.2	0.0	0.0	10.7	9.0	11.6	19.4	0.0	11.7	6.3	0.0	5.2
+ Visual-RFT	20.7	24.5	23.4	2.0	16.0	27.7	20.2	14.4	45.8	11.1	22.7	6.0	6.0	40.2
Δ	**+18.0**	**+22.9**	**+22.2**	**+2.0**	**+13.6**	**+27.7**	**+10.2**	**+14.4**	**+32.4**	**+10.9**	**+18.0**	**+3.9**	**+6.0**	**+40.2**
Qwen2-VL-7B	15.7	3.7	21.9	0.7	24.5	15.3	19.2	13.1	14.5	11.9	18.1	27.9	0.0	33.8
+ SFT	24.0	20.8	25.4	0.6	41.8	12.2	19.2	18.8	42.5	11.9	15.3	27.9	28.1	47.8
+ Visual-RFT	30.4	19.7	27.8	4.3	41.8	17.4	35.1	20.0	70.6	16.7	23.5	29.8	29.3	59.8
Δ	**+14.7**	**+16.0**	**+5.9**	**+3.6**	**+17.3**	**+2.1**	**+15.9**	**+6.9**	**+56.1**	**+4.8**	**+5.4**	**+1.9**	**+29.3**	**+26.0**

`<think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think><answer> answer here </answer> Question Output the thinking process in <think> </think> and your grouding box.Following "<think> thinking process </think> <answer>`$(x1, y1), (x2, y2)$`</answer>)" format.`

C. Visual-RFT on Domain Specific Datasets

C.1. Classification Domain Specific Datasets

Traditional Visual Instruction Tuning (Supervised Fine-Tuning, SFT) methods typically rely on large-scale datasets to fine-tune models through supervised learning. However, these methods demonstrate limited performance improvements when data availability is restricted, particularly in specialized domains where data collection is challenging and annotation costs are high. Fields such as medicine, military applications, and industrial inspection often face significant obstacles in gathering sufficient data samples, leading to suboptimal model performance. Consequently, finding an effective fine-tuning strategy under low-data scenarios has become a critical research question.

We propose Visual Reinforcement Fine-Tuning (Visual-RFT) to address this challenge. Visual-RFT introduces reinforcement learning mechanisms that enable strong few-shot learning capabilities and significantly enhance the model's general-

Table 6. **Few-shot results on Fine-grained Classification dataset.** We evaluated four fine-grained image classification datasets. Baseline results from InPK [13] under 4-shot settting.

Models	Average	Flower102	Pets37	FGVC	Cars196
		4-shot			
CoOp [11]	62.7	70.7	89.4	24.9	65.7
CoCoOp [12]	68.9	82.6	93.0	30.9	69.1
PromptSRC [3]	72.3	91.3	93.2	32.8	71.8
MaPLe [2]	67.7	80.8	92.1	29.0	68.7
		Qwen2-VL-2B			
Baseline	56.0	54.8	66.4	45.9	56.8
		1-shot			
+ SFT	51.7	56.6	54.7	65.3	30.0
+ Visual-RFT	80.3	70.8	84.1	72.5	93.8
Δ	**+24.3**	**+16.0**	**+17.7**	**+26.6**	**+37.0**
		2-shot			
+ SFT	58.8	60.3	65.6	68.9	40.2
+ Visual-RFT	83.5	75.8	87.5	75.3	95.4
Δ	**+27.5**	**+21.0**	**+21.1**	**+29.4**	**+38.6**
		4-shot			
+ SFT	55.6	58.5	55.5	67.9	40.5
+ Visual-RFT	81.9	71.4	86.1	74.8	95.3
Δ	**+25.9**	**+16.6**	**+19.7**	**+28.9**	**+38.5**
		8-shot			
+ SFT	60.3	59.6	71.4	69.2	40.9
+ Visual-RFT	85.1	77.7	90.2	75.9	96.5
Δ	**+29.1**	**+22.9**	**+23.8**	**+30.0**	**+39.7**
		16-shot			
+ SFT	64.0	66.8	71.6	76.1	41.5
+ Visual-RFT	85.3	79.2	87.1	79.4	95.3
Δ	**+29.3**	**+24.4**	**+20.7**	**+33.5**	**+38.5**
		Qwen2-VL-7B			
Baseline	57.4	49.8	61.8	41.0	76.8
		4-shot			
+ SFT	67.3	63.7	76.9	76.4	52.2
+ Visual-RFT	85.6	76.8	91.3	79.5	94.7
Δ	**+28.2**	**+27.0**	**+29.5**	**+38.5**	**+17.9**

ization. Unlike traditional instruction fine-tuning, Visual-RFT leverages limited data more effectively by integrating reward-based learning, guiding the model to autonomously explore optimal solutions. This approach not only reduces dependency on vast labeled datasets but also excels in complex tasks, particularly where sophisticated reasoning, cross-domain adaptation, or handling long-tail data distributions are required.

To evaluate the generalization capability of Visual-RFT, we selected a challenging medical classification dataset, ChestXR, for our experiments. The dataset contains three classes of CT chest images: COVID-19, Pneumonia, and Normal (healthy). These images are highly specialized and complex, making it difficult even for non-expert doctors to distinguish between the categories visually. Additionally, medical data often come with strict privacy and security constraints, making large-scale data collection impractical. Therefore, we adopted a few-shot learning approach, fine-tuning the model with a minimal amount of training samples.

As shown in Tab. 2, due to the high difficulty of the data, the results of SFT and Visual-RFT under the 1-shot setting are quite similar and close to the baseline. In this scenario, the model tends to classify all CT images as "Normal." When the setting is increased to 4-shot, SFT continues to perform poorly, while Visual-RFT demonstrates a slight improvement, particularly in distinguishing Pneumonia and COVID-19 cases. At the 8-shot level, both Visual-RFT and SFT show performance gains over the baseline, but Visual-RFT outperforms SFT. The results, shown in Tab. 2, demonstrate that Visual-RFT maintains strong performance and generalization even in this low-data, high-difficulty scenario. It significantly outperforms

Table 7. **Few-Shot results on COCO dataset of 8 categories.** We conducted one-shot, 2-shot, 4-shot, 8-shot, and 16-shot experiments on 8 categories from the COCO dataset.

Models	mAP	bus	train	fire hydrant	stop sign	cat	dog	bed	toilet
Qwen2-VL-2B									
Baseline	19.6	19.0	15.8	25.8	18.4	29.9	23.2	14.6	9.8
1-shot									
+ SFT	19.5	18.3	17.4	23.1	18.2	28.0	23.4	17.3	10.4
+ Visual-RFT	33.6	23.4	35.7	39.1	23.8	54.3	42.5	19.5	30.8
Δ	**+14.0**	**+4.4**	**+19.9**	**+13.3**	**+5.4**	**+24.4**	**+19.3**	**+4.9**	**+21.0**
2-shot									
+ SFT	21.0	22.1	15.8	29.8	19.0	28.9	26.5	15.5	10.2
+ Visual-RFT	41.5	28.8	39.6	38.2	48.0	63.8	52.7	25.9	34.9
Δ	**+21.9**	**+9.8**	**+23.8**	**+12.4**	**+29.6**	**+33.9**	**+29.5**	**+11.3**	**+25.1**
4-shot									
+ SFT	25.2	25.4	23.6	26.6	21.5	35.6	30.6	18.4	19.9
+ Visual-RFT	40.6	30.0	40.6	45.7	35.0	60.9	44.9	24.6	43.1
Δ	**+21.0**	**+11.0**	**+24.8**	**+19.9**	**+16.6**	**+31.0**	**+21.7**	**+10.0**	**+33.3**
8-shot									
+ SFT	30.2	25.8	35.1	29.4	21.9	44.5	39.0	22.6	23.5
+ Visual-RFT	47.4	36.2	47.9	50.4	47.7	65.2	57.0	30.4	44.0
Δ	**+27.8**	**+17.2**	**+32.1**	**+24.6**	**+29.3**	**+35.3**	**+33.8**	**+15.8**	**+34.2**
16-shot									
+ SFT	31.3	24.0	35.9	32.0	23.6	39.8	40.6	27.5	26.8
+ Visual-RFT	46.8	36.2	44.4	51.4	48.5	66.6	56.2	27.6	43.4
Δ	**+27.2**	**+17.2**	**+28.6**	**+25.6**	**+30.1**	**+36.7**	**+33.0**	**+13.0**	**+33.6**
Qwen2-VL-7B									
Baseline	43.0	35.0	43.3	37.1	36.7	57.3	50.3	37.4	47.1
4-shot									
+ SFT	44.1	41.4	51.7	35.6	30.8	60.5	52.7	38.5	41.5
+ Visual-RFT	54.3	44.3	59.8	52.0	46.0	72.7	62.8	41.9	55.0
Δ	**+11.3**	**+9.3**	**+16.5**	**+14.9**	**+9.3**	**+15.4**	**+12.5**	**+4.5**	**+7.9**

traditional instruction fine-tuning methods, highlighting the practical value of our approach in data-scarce environments. We also present some inference cases in Fig. 1.

C.2. Detection Domain Specific Datasets

We further test on some abstract out-of-domain datasets. We selected the MG (Monster Girls) dataset, which contains different types of anime-style monster girls. By using out-of-domain data, we increased the difficulty of both model recognition and reasoning, and conducted experiments under 4-shot and 16-shot settings. The results, shown in Tab. 3, indicate that reinforcement fine-tuning achieved a significant performance improvement, surpassing supervised fine-tuning (SFT). We also present some inference cases in Fig. 2.

D. More Experiments Results

D.1. Open Vocabulary Detection

OV on COCO Dataset In the main text, we present a comprehensive evaluation of our model's performance under the open vocabulary setting using the COCO dataset. Specifically, our model was trained on 48 base categories and tested across 17 unseen categories.

To further investigate the generalization capability of Visual-RFT, particularly in low-data scenarios, and to provide a

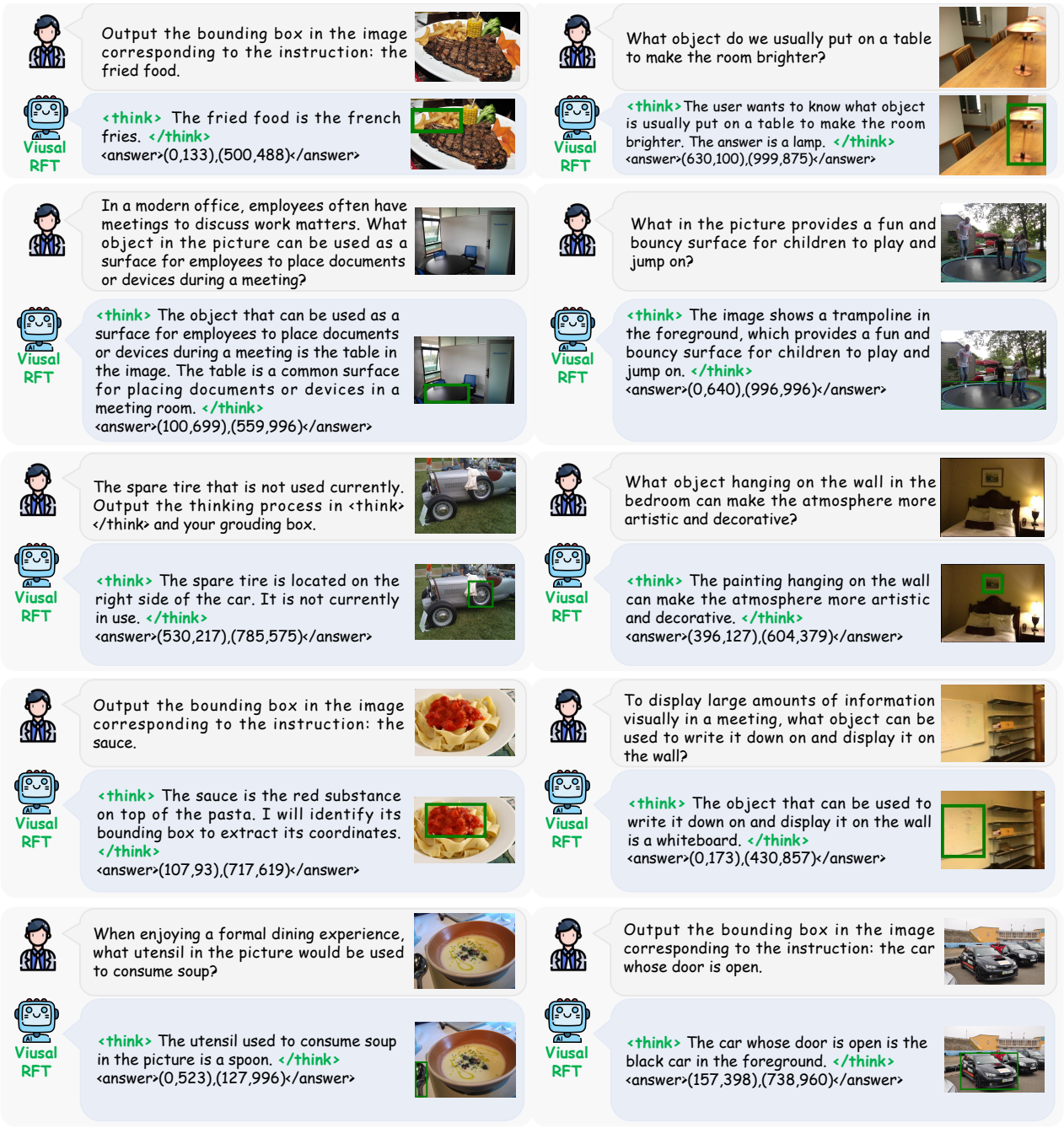


Figure 3. **More qualitative results of reasoning grounding on LISA [5] dataset.**

detailed comparison with traditional Supervised Fine-Tuning (SFT), we conducted a series of few-shot experiments. These experiments included 1-shot, 2-shot and 4-shot training settings exclusively on the 65 base categories, and test on 15 unseen categories. Our goal was to assess how well the model could adapt to unseen classes and maintain robust performance when trained with minimal data.

The results, as shown in Tab. 4, demonstrate how Visual-RFT outperforms SFT, particularly in scenarios with limited data availability. These findings highlight Visual-RFT 's strength in few-shot learning and its ability to generalize effectively to novel categories, offering a promising approach for applications where data collection is challenging or expensive.

Figure 4. **More qualitative results of object detection on COCO [6] dataset.**

OV on LVIS Dataset In Tab. 5, we evaluate the transferability of the model trained on COCO base categories by testing it on the LVIS dataset. To increase the difficulty and accelerate testing, we selected 13 rare categories from LVIS for evaluation. The results show that the Qwen2-VL-2B and Qwen2-VL-7B models improved mAP by 18.0 and 14.7 points over the baseline, respectively.

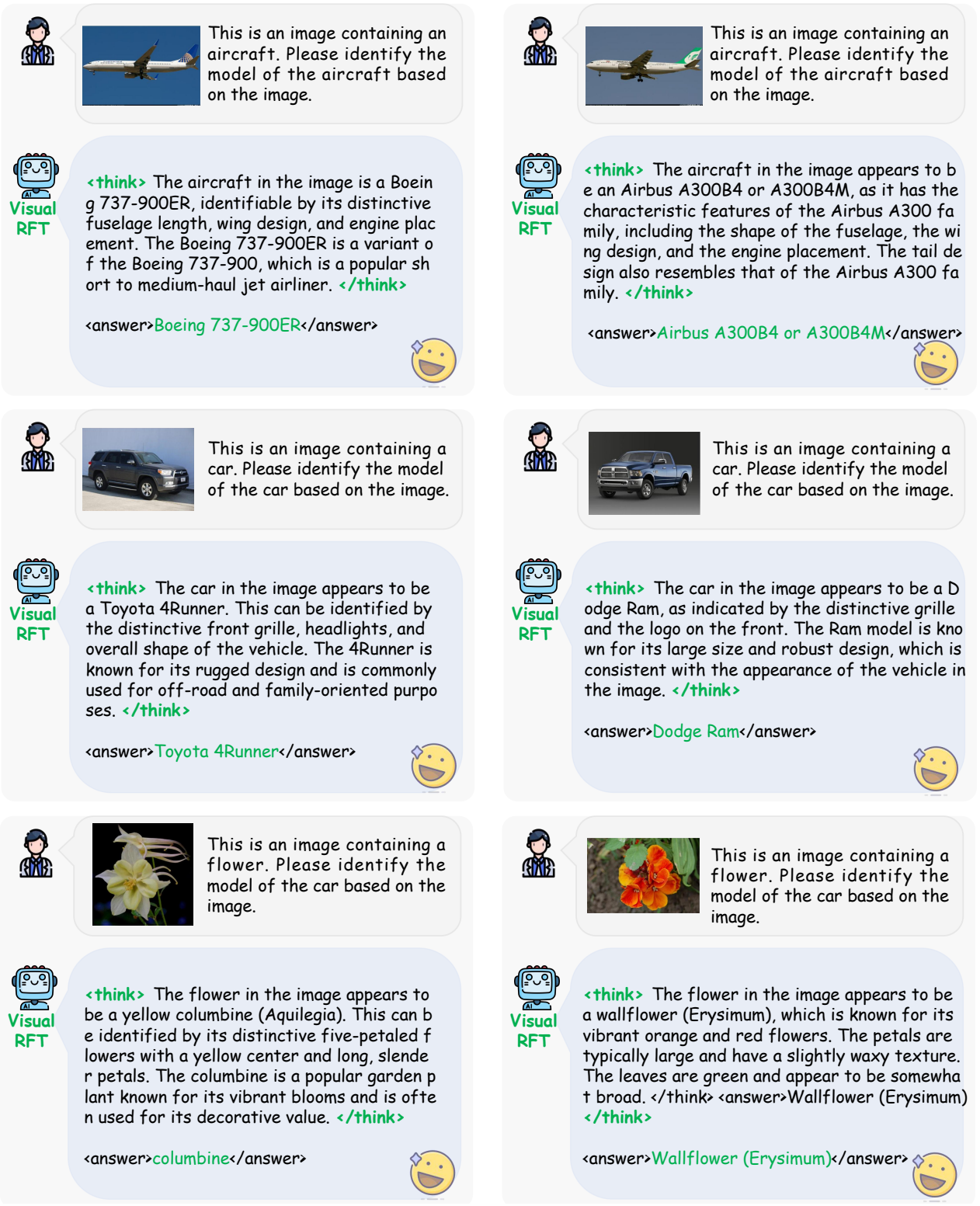


Figure 5. **More qualitative results of fine-grained image classification.**

Table 8. Ablation between CoT and RFT.

	COCO		LISA	
	w/ CoT	w/o CoT	w/ CoT	w/o CoT
Qwen2-VL-2B	19.5	28.6	21.3	26.9
+ SFT	26.4	30.2	29.0	28.3
+ Visual-RFT	42.6	44.3	37.6	35.9

Visual-RFT not only transfers its detection capabilities from the COCO base categories to new COCO categories but also achieves remarkable performance gains on the more challenging rare categories of LVIS. Notably, as shown in Tab. 5, the original and SFT-trained models fail to recognize certain rare LVIS categories, resulting in an average precision (AP) of 0. However, after reinforcement fine-tuning, the model exhibits a qualitative leap from 0 to 1 in its ability to recognize previously unidentifiable categories, such as "egg roll" and "futon."

These results highlight the significant impact of Visual-RFT in enhancing the performance and generalization ability of large vision-language models (LVLMs) in visual recognition tasks, particularly under challenging scenarios with rare and unseen categories.

D.2. Fine-Grained Few-shot Image Classification

In Tab. 6, we present classification results under additional shot settings, including the performance of the Qwen2-VL-7B model in the 4-shot setting. These results provide a broader perspective on how the model performs with varying amounts of training data, demonstrating its adaptability and generalization capability across different scenarios.

D.3. Few-shot Detectoin on COCO

In the main text, we present the test results of the model trained using Visual-RFT in a few-shot setting on eight COCO categories. Additionally, in Tab. 7, we provide more extensive results under different shot settings, including the performance of the Qwen2-VL-7B model. These tables offer a comprehensive view of the model's effectiveness and generalization ability when trained with limited data.

E. Visual-RFT Reasoning Cases

We provide a variety of examples for reasoning grounding, detection, and classification in Fig. 3, Fig. 4, Fig. 5, respectively.

F. Other Experiments

We have added some experiments, and we hope they are helpful.

F.1. Necessity of CoT.

Introducing a reasoning chain is necessary for certain tasks that involve inference, such as reasoning grounding (LISA) and GUI understanding. Using CoT leads to significant improvements in overall performance, demonstrating a trade-off between latency and performance, despite increasing inference time (e.g., about 25% more on 5k images (Tab. 8).

F.2. Ablation between CoT and RFT.

We ablate the effect of CoT by comparing two variants—with and without CoT reasoning. Both variants substantially outperform the baseline and SFT (Tab. 8). CoT improves Visual-RFT's performance on reasoning-heavy tasks like LISA (mAP35.9→37.6), but slightly reduces accuracy on simpler detection tasks (mAP44.3→42.6). These results suggest that Visual-RFT is the primary source of pronounced performance gains, regardless of CoT usage, while CoT provides additional benefits for reasoning-intensive scenarios.

F.3. Extended to more tasks.

We applied our method to the Nuscene-QA. Trained on 2k samples with exact match rewards, our model achieved **41.05** % accuracy, significantly outperforming the Qwen2-VL-2B baseline (**36.11**%), and effectively handled diverse question types such as existence, counting, and comparison.

F.4. Compare to PPO.

We compare GRPO and PPO on Qwen2.5-VL-3B using the same setting. GRPO achieves a higher mAP of 51.98, outperforming PPO (50.12) and the baseline (46.52). In addition, PPO requires careful hyperparameter tuning, while GRPO is simpler and more stable.

F.5. Tracking.

We choose the OTB100 tracking dataset, train the model on just 80 two-frame samples with the IoU reward, and evaluate on every frame of the test videos. Significant improvements over baseline: AUC 28.87 (vs. 15.54), Prec@20 35.61 (vs. 17.41), mAP 30.94 (vs. 11.18).

References

[1] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 2

[2] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2022. 6

[3] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Siddique Khan, Ming Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15144–15154, 2023. 6

[4] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 1, 2

[5] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 3, 8

[6] 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. 1, 2, 9

[7] 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. 5

[8] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 2

[9] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 1, 2

[10] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 2

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348, 2021. 6

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022. 6

[13] Shuchang Zhou. Inpk: Infusing prior knowledge into prompt for vision-language models. 2025. 6