

AAAI 2026 Supplementary Material

Appendix

Jiaao Yu, Shenwei Li, Mingjie Han, Yifei Yin, Wenzheng Song, Chenghao Jia, Man Lan

Abstract

This supplementary material contains additional information about the proposed method and experimental results that were omitted from the main manuscript due to space constraints. The appendix is structured as follows:

- **Benchmark Analysis.** First section mainly includes the construction process of the Benchmark and related data analysis.
- **Details of Models and Training.** Second section mainly includes the process of selecting components for parameter sensitivity verification experiments and training strategies through experiments, and more experimental results.
- **Details of Generalization Experiment.** Third section mainly includes the details and analysis of generalization experiments
- **Visualization and Analysis.** Fourth section includes more visualized results, case analyses, and reasoning result analyses.
- **Limitations.** Fifth part discusses limitations.

Benchmark Analysis

Data construction

Data filtering. In the process of constructing the MPCC-Eval benchmark, we first carried out data screening. Specifically, we used the validation set of the COCO dataset (Lin et al. 2014) as the basic data source and conducted the first round of preliminary screening, aiming to eliminate images with weak semantic relevance and those that are illogical. During this screening phase, we adopted the prompt words shown in Table 1 prompt 1 and used the GPT-4o (Hurst et al. 2024) and Qwen2.5-VL-7B (Bai et al. 2025) models to perform the screening operation, finally retaining high-quality data for subsequent screening processes. After the preliminary data screening, we conducted further manual screening to select higher-quality data and classify them by difficulty levels. Specifically, we categorized the difficulty levels based on the following principles: Easy level: If the target object in the image occupies a dominant visual position and the answer can be obtained only through simple reasoning (such as analogical reasoning) based on the image context,

it is classified as easy. Moderate level: If the target object occupies a dominant position in the image, but the answer can only be derived through simple object relationship reasoning or scene commonsense reasoning, it is classified as moderate. Hard level: If the answer requires reasoning based on both image context details and common sense, and involves in-depth reasoning through a multi-step elimination process, it is classified as hard. Through the above process, we finally constructed the sample data of the MPCC-Eval benchmark.

Prompt 1:

An image has been partially occluded by a black mask, with an object of category $\{\text{answer}\}$ located within the occluded region. Based on the contextual cues present in the visible regions of the image, please evaluate the level of difficulty in inferring that the occluded bounding box contains a $\{\text{answer}\}$. Difficulty is defined as follows: 1 (Low Difficulty): The object can be confidently inferred based on strong contextual or semantic cues. 2 (Medium Difficulty): The object can be plausibly inferred, but with moderate ambiguity or reliance on prior knowledge. 3 (High Difficulty): The object is unlikely to be inferred due to weak or ambiguous contextual information. 0 (Impossible): There is no discernible contextual or semantic information to support inference. Please output only one integer: 0, 1, 2, or 3, reflecting the difficulty level of this visual reasoning task.

Prompt 2:

A portion of the image has been occluded with black. Please infer what object is located within the occluded area. Provide your top ten possible guesses enclosed in square brackets, separated by commas.

Prompt 3:

A portion of the image has been occluded with a black mask. Please infer what object is likely hidden in the masked area. You are given the following seven candidate options: A. $\{\text{option1}\}$ B. $\{\text{option2}\}$ C. $\{\text{option3}\}$ D. $\{\text{option4}\}$ E. $\{\text{option5}\}$ F. $\{\text{option6}\}$ G. $\{\text{option7}\}$ Please select the most appropriate option based on visual context and reasoning.

Prompt 4:

A portion of the image has been occluded with a black mask. Please infer what object is likely hidden in the masked area. You are given the following four candidate options: A. $\{\text{option1}\}$ B. $\{\text{option2}\}$ C. $\{\text{option3}\}$ D. $\{\text{option4}\}$ Please select the most appropriate option based on visual context and reasoning.

Table 1: Prompts for the data construction process.

Construction of confusing items. To conduct a more

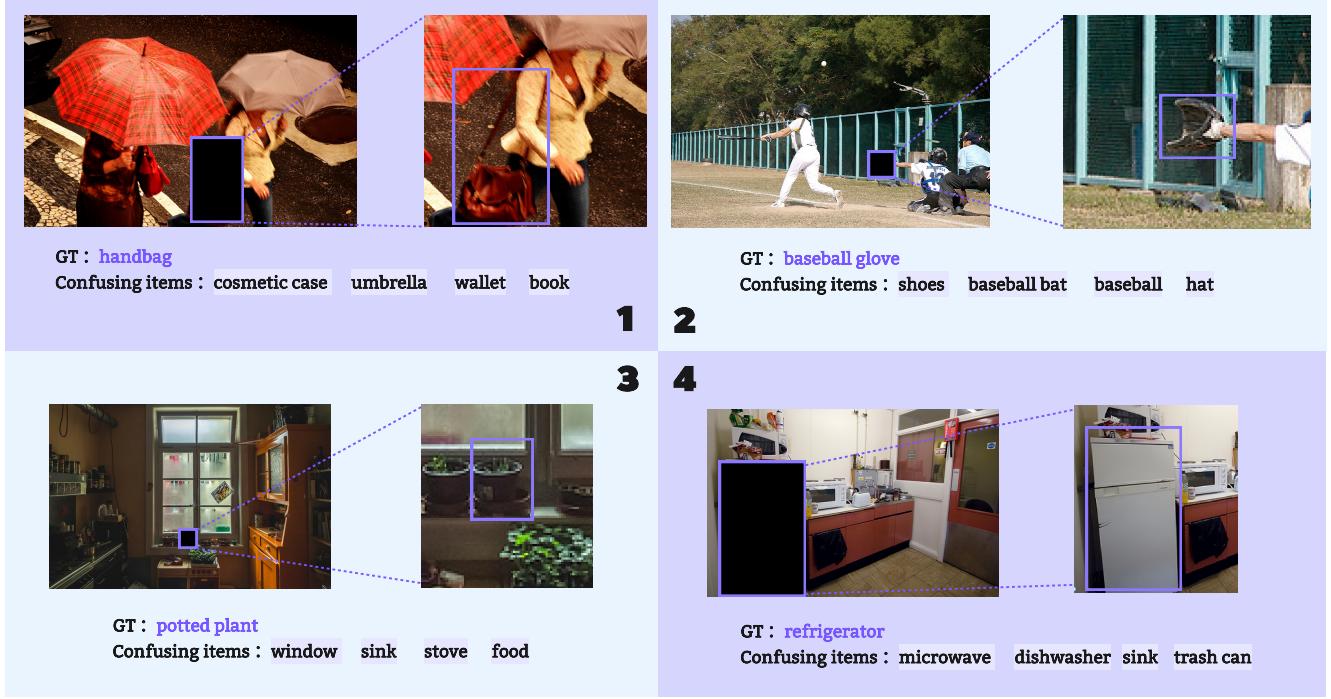


Figure 1: Data examples and confusing item examples of MPCC-Eval.

challenging evaluation, we used multiple models to respond to prompt2 as shown in Table 1. After collecting multiple sets of answers, we further determined the final options through manual screening. The screening criteria are as follows: the selected answers must have a certain correlation with the image context and basic logicality, but their logical rigor is weaker than that of the ground truth (GT). Based on these criteria, we constructed several confusing items and irrelevant items for each target object, and designed two versions of single-choice questions accordingly. The hard version of the single-choice question includes 1 ground truth (GT), 4 confusing items, and 2 irrelevant items, totaling 7 options, requiring the model to select the correct answer from them. The easy version of the single-choice question includes 1 ground truth (GT), 2 confusing items, and 1 irrelevant item, totaling 4 options, also requiring the model to complete a single-choice task. During the evaluation, the two versions of single-choice questions used prompts3 and prompt4 as shown in Table 1, respectively.

Case presentation

To intuitively show how MPCC-Eval tests visual context and commonsense reasoning, Figure 1 presents representative cases. In the first scene, two people with umbrellas form the core context, with a masked "handbag" (GT). Confusing items like "cosmetic case" and "book" lack contextual alignment, while "umbrella" repeats existing elements. Models must use the rainy scenario and commonsense (typical outdoor items) to identify the GT. In the second scene, the masked "baseball glove" (GT, near a catcher) faces distract-

tors like "shoes" (hand-irrelevant) and "baseball bat" (batter gear). Models need sports knowledge and hand position cues to recognize the GT. These cases highlight MPCC-Eval's focus on testing integration of visual details, scene commonsense, and object-relationship reasoning, core capabilities MPCC training activates, by challenging models to reason about masked positions and filter inconsistent distractors.

Mainstream model reasoning results

Building on the benchmark experiments in the main text, we further tested the performance of several mainstream models after reasoning guided by prompts in 1 prompt 1, with the results reported in Table 2. Among them, Qwen-VL-series models showed a significant improvement in accuracy on hard questions, while LLaVA-series models (Liu et al. 2023) exhibited a notable decline. This is largely attributed to differences in the instruction-following capabilities of various models—some models were judged incorrect due to format errors in their responses. Thus, the results of this experiment are for reference only. However, the significant accuracy improvement of Qwen2.5-VL-7B on hard questions after reasoning also confirms that the MPCC task is more suitable for activating the reasoning abilities of models.

Details of Models and Training

Supplementary experiment of fine-tuning strategy

This section presents in detail the experimental results that could not be fully displayed in the main text due to space constraints. Table 3 shows the experimental results of Qwen2.5-VL-3B (Bai et al. 2025) as the baseline model on

Subset Methods	Hard		Moderate		Easy		Ave.		Sum Score
	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	
Qwen2.5-VL-3B -reasoning	48.14	21.43	63.05	27.57	66.08	41.46	59.09	30.15	267.73
	50.00	25.78	62.76	34.02	66.96	39.91	59.91	33.24	279.43
Qwen2.5-VL-7B -reasoning	58.39	21.12	73.02	33.43	75.83	42.79	69.08	32.45	304.58
	52.80	36.02	69.50	49.26	74.50	63.86	65.60	59.76	376.09
LLaVA1.6-7B -reasoning	41.30	18.63	48.30	22.58	47.01	17.96	45.54	19.72	195.78
	35.09	17.09	43.70	18.48	41.46	23.06	40.08	19.54	178.88
LLaVA1.6-13B -reasoning	46.58	23.60	57.77	26.39	60.09	34.15	54.81	28.05	248.58
	44.10	19.57	56.01	28.45	56.98	27.94	52.36	25.32	233.05
Intern3-VL-1B -reasoning	44.40	25.78	49.56	32.26	58.31	34.37	50.76	30.80	244.68
	38.47	20.57	46.58	21.45	47.72	21.63	44.26	21.22	196.45
Intern3-VL-8B -reasoning	60.56	41.61	70.67	57.18	72.72	65.41	67.98	54.73	368.15
	54.97	43.79	71.85	49.30	70.95	56.78	65.92	49.96	347.64

Table 2: The performance of current mainstream vision-language models on the MPCC-Eval benchmark is presented.

the three difficulty subsets of the MPCC-Eval benchmark, whereas only the average (Ave.) results were presented in the main text.

For further experimental validation, we conducted additional fine-tuning training using Qwen2-VL-2B (Wang et al. 2024) as the baseline model to supplement the experiments on fine-tuning strategies. As shown in Table 4, in the in-distribution performance evaluation, RFT with prior sampling performs more prominently in difficult problems. In experiments with Qwen2-VL-2B as the base model, the conclusions are largely consistent with those of the experiments using Qwen2.5-VL-3B as the base model in the main text.

Strategy component ablation

Reward function ablation. This section presents the reward functions we experimented with. In addition to the format reward, we tested three types of reward functions, the first two of which have been introduced in the main text, namely the matching reward r_{match} and the levenshtein ratio reward $r_{leven..}$. Furthermore, we used the CLIP model (Radford et al. 2021) to calculate the image-text cosine similarity—by matching the cropped masked region image m with the model’s final answer a , and the resulting similarity was adopted as the third reward function, with its specific implementation shown in Eq 1. However, fine-tuning using this reward function did not yield satisfactory results. We believe that using small models to calculate rewards in specific tasks is a promising research direction, but since this is not the focus of this paper, no further optimization was performed, and related explorations will be carried out in future work.

$$r_{clip} = Sim(CLIP_{text}(a), CLIP_{image}(m)) \quad (1)$$

Here, $CLIP_{text}$ denotes the text encoder of the CLIP model, $CLIP_{image}$ denotes its image encoder, and $Sim(\cdot)$ is the cosine similarity calculation function. Finally, we tested these reward functions under both the Reinforcement Fine-Tuning (RFT) strategy and the Reinforcement Fine-Tuning with prior sampling strategy, with the results presented in Table 5.

RFT with prior sampling. In implementing the Reinforcement Fine-Tuning with prior sampling strategy, we adopted two approaches, as illustrated in Figure 3. Among them, approach (a) has been elaborated on in the main text. Additionally, we explored approach (b): specifically, we first utilized the annotated COT (Wei et al. 2022) data, input the reasoning-annotated data into the policy model, and allowed it to continue generating outputs, thereby obtaining results based on the annotated reasoning process. Subsequently, we concatenated this reasoning process with the sampled answer to serve as the prior sampling sample. Table 6 presents the performance results of the two approaches, and their performance is comparable. In line with Occam’s Razor (entities should not be multiplied beyond necessity), we only reported the first implementation approach in the main text. However, these two approaches may play their respective roles in different scenarios in the future, and we will continue to explore this in our future work.

Parameter selection for updating gradients. To test the effect of fine-tuning, we compared the performance of LoRA (Hu et al. 2021) fine-tuning (with ranks set to 16 and 32) and full-parameter fine-tuning using RFT strategies, with the results presented in Table 7. For a fair comparison, all experiments reported in the main text use full-parameter fine-tuning.

The impact of training steps. Figure 2 illustrates the impact of training steps on performance in reinforcement fine-tuning. We tested the training effects from 100 to 3200 steps and reported the performance using Cho.E single-choice questions on both Hard and Easy evaluation subsets.

Details of Generalization Experiment

Generalization experimental setting

To evaluate the model’s generalization ability, we designed two types of experiments: out-of-distribution (OOD) and cross-task. In the OOD experiments, we divided the 80 categories of the COCO dataset into two disjoint subsets, which were used for training and testing respectively, ensuring that the test categories never appeared during training to verify the model’s reasoning ability on unknown categories. For

Subset	Hard		Moderate		Easy		Ave.		Sum
Methods	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	Score
Qwen2.5-VL-3B	48.14	21.43	63.05	27.57	66.08	41.46	59.09	30.15	267.73
-prompt	50.00	25.78	62.76	34.02	66.96	39.91	59.91	33.24	279.43
-sft	60.25	43.79	78.01	62.46	84.04	70.29	74.10	58.85	398.84
-rft	57.76	27.02	66.57	41.94	74.94	42.57	66.42	37.18	310.80
-sft+rft	59.01	43.48	81.52	59.82	84.26	69.18	74.96	58.18	397.27
-rft+pri.	58.39	36.34	78.00	56.30	81.37	61.86	72.59	51.50	372.26

Table 3: The experimental results of MPCC-Eval fine-tuning using Qwen2.5-VL-3B as the base.

Subset	Hard		Moderate		Easy		Ave.		Sum
Methods	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	Cho.E	Cho.H	Score
Qwen2-VL-2B	31.68	15.84	33.14	20.53	35.48	20.84	33.43	19.07	157.51
-prompt	19.88	11.63	27.27	15.26	29.61	18.65	25.59	15.18	122.30
-sft	46.89	26.71	55.43	30.50	56.76	32.59	53.03	29.93	248.88
-rft	35.78	21.19	43.82	30.91	51.55	31.71	43.72	27.94	214.96
-sft+rft	47.02	26.03	54.89	29.60	58.33	33.96	53.41	29.86	249.83
-rft+pri.	46.03	22.12	49.97	32.09	53.99	35.81	50.00	30.01	240.01

Table 4: The experimental results of MPCC-Eval fine-tuning using Qwen2-VL-2B as the base.

Methods	Hard		Ave.	
	Cho.E	Cho.H	Cho.E	Cho.H
With rft				
- r_{match}	48.45	28.24	61.05	37.55
- $r_{leven.}$	50.93	31.37	60.12	38.86
- $r_{match} + r_{leven.}$	57.76	27.02	66.42	37.18
- $r_{match} + r_{leven.} + r_{clip}$	51.89	25.68	63.64	32.16
With rft+pri.				
- r_{match}	50.93	33.96	65.98	43.48
- $r_{leven.}$	56.52	35.43	72.41	48.83
- $r_{match} + r_{leven.}$	58.39	36.34	72.59	51.50
- $r_{match} + r_{leven.} + r_{clip}$	52.17	32.43	62.11	39.25

Table 5: The ablation experiment of the reward function uses Qwen2.5-VL-3B as the base model.

Methods	Hard		Ave.	
	Cho.E	Cho.H	Cho.E	Cho.H
rft+pri. (a)	59.01	36.34	72.59	51.50
rft+pri. (b)	57.52	37.42	71.26	52.48

Table 6: Comparison of performance of two implementations of RFT with prior sampling uses Qwen2.5-VL-3B as the base model.

Methods	Hard		Ave.	
	Cho.E	Cho.H	Cho.E	Cho.H
LoRA with rank = 16	55.27	26.27	64.81	35.96
LoRA with rank = 32	57.14	28.23	67.16	36.98
full-parameter	57.76	27.02	66.42	37.18

Table 7: Comparison of performance of LoRA fine-tuning and full-parameter fine-tuning.



Figure 2: The impact of training steps on performance in reinforcement fine-tuning.

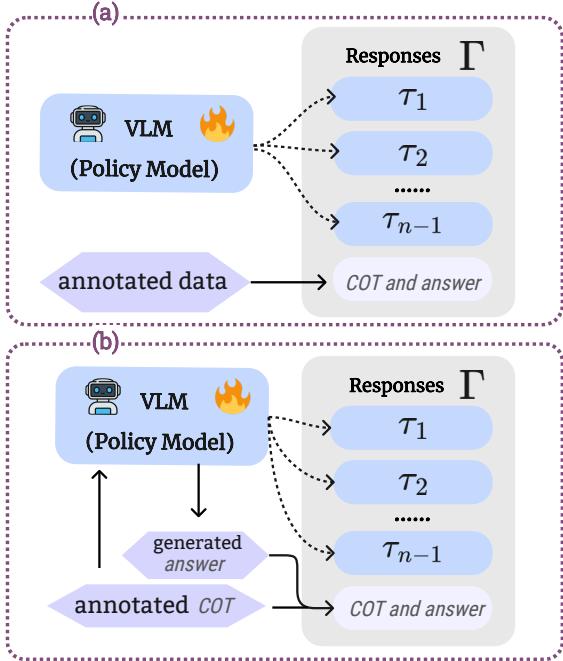


Figure 3: Two implementations of RFT with prior sampling.

the cross-task experiments, we selected four tasks from the BLINK benchmark (Fu et al. 2024): Spatial Relation, Semantic Correspondence, Relative Depth, and Object Localization. We compared the performance of models fine-tuned on the MPCC task with those fine-tuned on other perception tasks to assess the model’s transfer ability across different reasoning tasks. In the main text experiments, we used Qwen2.5-VL-3B as the baseline model for all trials, with training conducted over one epoch on the dataset. Examples of cross-task experiments will be introduced in the following sections.

Supplementary experiment

Case study of cross-task generalization experiment. Case studies of cross-task generalization experiments, as shown in Figure 4, show that models fine-tuned via the MPCC task perform well across multiple tasks: in semantic correspondence, relative depth, and object localization tasks, the models all derive correct results by combining image context and commonsense reasoning; in the spatial relation task, which requires no complex reasoning, the SFT strategy model gives a wrong answer due to incorrect reasoning, while the RFT strategy model with prior sampling provides the correct answer directly without explicit reasoning.

Cross-task generalization experiment based on Qwen2-VL-2B. Cross-task generalization experiments with Qwen2-VL-2B (Wang et al. 2024) as the base (results in Table 8) align with the main text conclusions, reaffirming the strong cross-task generalization of MPCC task fine-tuning.

Methods	Sem.	Rel.	Obj.	Spa.
Fine tune on perceive task				
Qwen2-VL-2B	33.09	56.45	44.26	65.03
PR1-Counting	27.34 <small>-5.75</small>	57.26 <small>+0.81</small>	40.16 <small>-4.10</small>	64.34 <small>-0.69</small>
PR1-Grounding	25.90 <small>-7.19</small>	53.23 <small>-3.22</small>	47.54 <small>-3.28</small>	65.73 <small>+0.70</small>
PR1-OCR	22.30 <small>-10.79</small>	57.26 <small>+0.81</small>	50.00 <small>+5.74</small>	66.43 <small>+1.40</small>
ReasonRFT-VC	32.37 <small>-0.72</small>	59.68 <small>+3.23</small>	40.16 <small>-4.10</small>	64.34 <small>-0.69</small>
Fine tune on MPCC				
sft	33.81 <small>+0.72</small>	58.87 <small>+2.42</small>	46.72 <small>+2.46</small>	64.34 <small>-0.69</small>
rft	35.25 <small>+2.16</small>	61.29 <small>+4.84</small>	50.82 <small>+6.56</small>	67.83 <small>+2.80</small>
sft+rft	31.65 <small>-1.44</small>	59.68 <small>+3.23</small>	45.08 <small>+0.82</small>	62.24 <small>-2.79</small>
rft+pri.	39.57 <small>+6.48</small>	60.48 <small>+4.03</small>	51.64 <small>+7.38</small>	67.13 <small>+2.10</small>

Table 8: The cross-task generalization performance of different fine-tuning strategies and various tasks base on Qwen2-VL-2B.

Notably, RFT with prior sampling performs well, showing improvements across multiple tasks.

Visualization and Analysis

Reasoning analysis

To verify the impact of different fine - tuning strategies on model reasoning performance, we introduce large models such as LLaVA1.6-13B(LLaVA.), Qwen2.5-VL-7B(Qwen.), GPT-4o-mini(mini.), GPT-4o(GPT.) and Intern3-VL-8B(Intern.)(Chen et al. 2024) as evaluation tools to score the outputs of models using strategies. We present the results in Table 8 and showcase the scoring prompts we used in Table 10. All fine-tuning strategies significantly improve the reasoning quality. We find that fine-tuning guided by annotated reasoning paths can lead the model to achieve higher - quality optimization. Among them, the RFT with prior sampling is particularly prominent in exploring the potential for improving the model’s reasoning quality.

Methods	LlaVA.	Qwen.	mini.	GPT.	Intern.	Ave.
prompt	4.21	3.71	5.75	4.95	5.61	4.85
sft	6.72	7.10	7.93	6.89	6.92	7.11
rft	5.10	6.04	6.86	6.49	6.43	6.18
sft+rft	6.63	7.54	7.76	5.66	8.45	7.21
rft+pri.	6.57	7.29	7.53	7.12	8.66	7.43

Table 9: Evaluate the reasoning quality different fine-tuning strategies using large models.

Case study

Figure 5 shows examples of model outputs from various fine-tuning strategies on MPCC task. In the first scenario, a black mask covers the sink-adjacent area. The base model misclassifies the masked object as soap, failing scene - aware reasoning. The SFT model infers it as a toothbrush by leveraging “near-sink” context and the common sense that “toothbrushes, used for oral care, are typically placed by sinks”. The RFT model initially over - associates (e.g., linking to

Please act as an expert in multimodal reasoning quality assessment and conduct a 0-10 quantitative scoring (10 being the best) on the model’s output reasoning process and conclusion based on the following dimensions. Provide an overall score and the rationale for the scoring:

Visual context relevance: Does the reasoning strictly rely on visual cues in the image (e.g., object positions, scene features, information around the masked area)?

Is there any subjective speculation divorced from the image content?

Effectiveness of commonsense reasoning: Does it accurately apply commonsense related to the scene?

Is the combination of commonsense and visual cues reasonable?

Logical coherence: Are the reasoning steps coherently connected?

Is there any logical gap or contradiction (e.g., first reasoning ‘small items’ and then identifying ‘large furniture’)?

Conclusion consistency: Is the final answer reasonably derived from the reasoning process?

Does it accurately respond to the task requirements (e.g., ‘predict masked objects’ ‘judge spatial relationships’)?

This task is to predict the masked object by reasoning through the image context and common sense given an image with a masked part. Below is an example of reasoning and answering based on a given image, which ground truth is gt. Please score and evaluate it:

Reasoning process: {*think*}

Answer result: {*answer*}

Table 10: Prompts for reasoning analysis experiment.

“work preparation”) but corrects via “bathroom item layout” reasoning. With prior sampling, RFT simplifies inference using “sink proximity” and “daily hygiene habits”, quickly identifying the toothbrush. In the second scenario, the mask covers an urban area. The base model misidentifies the masked object as a dog, neglecting urban scene functions. The SFT model infers it as a fire hydrant by combining “sidewalk/street facility” context and the common sense of “fire hydrants’ cylindrical shape and fire fighting water supply role”. The RFT model first misassociates with “transportation (trains, bus stops)” but corrects via “spatial layout and infrastructure needs” reasoning. RFT with prior sampling simplifies judgment using “urban fire-fighting needs, cylindrical, water releasing features”, ensuring accuracy.

Limitations

While MPCC-Eval excels at probing fine-grained visual commonsense in static, single-scene contexts, its scenario design has two inherent boundaries: cultural and long-tail coverage gaps. The benchmark prioritizes universal, high-frequency scenarios to ensure clear reasoning signals—a strength for standardized evaluation. However, this focus sidelines low-frequency, culturally nuanced contexts (e.g., regional festivals with unique attire, specialized tools in traditional crafts). Such long-tail cases, though rare, are critical for real-world robustness but remain underrepresented. MPCC-Eval’s single-image, short-context tasks (e.g., masking in a single frame) effectively improve visual reasoning.

Yet, they stop short of capturing sequential multimodal challenges (e.g., tracking object relationships across a 10-frame video, inferring narratives from a photo series). These dynamic, long-context tasks demand higher-order reasoning (e.g., memory, temporal coherence) that the current benchmark doesn’t fully address, even as its static-scene tests lay a strong foundation. Additionally, the reinforcement fine-tuning method with prior sampling proposed in this paper shows great potential and can perform effectively in scenarios with low-quality reasoning annotations. However, due to the focus of this study, we have not yet quantified the capabilities of this strategy through experiments. Future work will involve more in-depth investigations.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.
- Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; and Clark, J. 2021. Learning Transferable Visual Models From Natural Language Supervision.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Semantic Correspondence

Question:

Humans can find corresponding points for different objects in the same category. For instance, if there are images of two different cats, then the left ear tip of one cat corresponds to the left ear tip of the other cat, and the right front paw of one cat corresponds to the right front paw of the other cat.\nGiven the following two images, a reference point is annotated on the first image, labeled with REF. You are given multiple red-circled points on the second image, choices of 'A, B, C, D' are drawn beside each circle. Select between the choices on the second image and find the corresponding point for the reference point. Which point is corresponding to the reference point?\nSelect from the following choices.

- (A) Point A (B) Point B (C) Point C (D) Point D



Answer:
(B) Point B

Model output (RFT with prior sampling on MPCC)

<think> In the first image, we have a pug standing on a wooden floor. The reference point (REF) is marked on the dog's face, specifically around its nose area. In the second image, we see a hairless dog wearing an orange shirt lying on a couch. We need to identify which part of this dog corresponds to the REF point from the first image. Looking at the second image, we notice that the hairless dog has a similar shape and size to the pug in the first image. The REF point on the pug seems to be located where the dog's nose would be. Comparing the positions, we can see that Point B on the second image aligns well with the REF point on the first image, as both are positioned near the dog's nose area. Therefore, based on the similarity in shape and size, and the alignment of the REF point with Point B, we can conclude that Point B is the corresponding point to the REF point in the first image.</think><answer>B</answer>

Relative Depth

Question:

Two points are circled on the image, labeled by A and B beside each circle. Which point is closer to the camera?\nSelect from the following choices.

- (A) A is closer (B) B is closer



Answer:
(B) B is closer

Model output (SFT+RFT on MPCC)

<think> The image shows a papercraft figure standing among stacked tires. Points A and B are marked next to circles indicating where these points might be located relative to the camera's perspective. Given the spatial relationship between the figure and the tires, point B appears to be positioned slightly lower than point A, suggesting it's closer to the ground or the base of the scene. Additionally, the angle at which the figure is viewed makes it appear larger compared to the tires behind it, further supporting the idea that point B is nearer to the viewer. Therefore, based on these observations, point B seems to be closer to the camera than point A.</think><answer>B</answer>

Object Localization

Question:

A bounding box is an annotated rectangle surrounding an object. The edges of bounding boxes should touch the outermost pixels of the object that is being labeled. Given the two bounding boxes on the image, labeled by A and B, which bounding box more accurately localizes and encloses the flower arrangement? Select from the following options.

- (A) Box A (B) Box B



Answer:
(B) Box B

Model output (RFT on MPCC)

<think> To determine which bounding box more accurately localizes and encloses the flower arrangement, we need to consider the size and shape of the flower arrangement relative to the bounding boxes.\n\nBox A appears to be larger than Box B, covering a wider area around the vase and flowers. However, Box B seems to be more precise in its positioning, as it captures the entire flower arrangement without any significant gaps or overlaps outside the actual flowers.\n\nGiven these observations, Box B appears to be the better choice for accurately localizing and enclosing the flower arrangement because it maintains a closer fit to the actual shape and size of the flowers while avoiding unnecessary space.\n\nTherefore, the correct answer is (B) Box B.</think><answer>(B) Box B</answer>

Spatial Relation

Question:

Is the umbrella on top of the cat?\nSelect from the following choices.

- (A) yes (B) no



Answer:
(B) No

Model output (SFT on MPCC)

<think> The image shows a cat lying inside an open suitcase. There is indeed an umbrella placed on top of the cat's back. The umbrella appears to be partially covering the cat, suggesting it was placed there intentionally. Therefore, based on the visual evidence provided by the image, the umbrella is on top of the cat.</think><answer>(A) yes</answer>

Model output (RFT with prior sampling on MPCC)

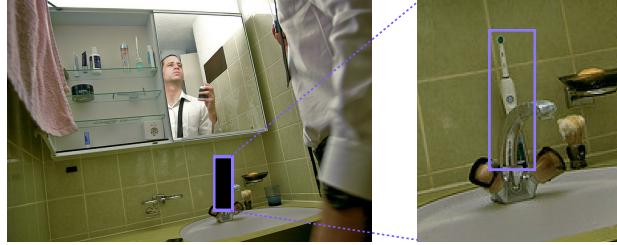
<answer>(B) no</answer>

Figure 4: Case study of cross-task generalization experiment.



GT: toothbrush

Base: soap 😕



Prompt: \n<think>\nThe context suggests that the black rectangular mask is placed over an area where something might be hiding or obscured. Given the bathroom setting, it's plausible to assume that the mask could be concealing a hairbrush or another grooming tool typically found near the sink.\n</think>\n<answer>\nHairbrush



SFT: \n<think>\nThe masked area appears to be positioned where one would typically place a toothbrush for use at the sink. The presence of other toiletries like toothpaste and mouthwash suggests that this space is used for personal hygiene routines such as brushing teeth. Given the proximity to the sink and the typical placement of these items, it's reasonable to infer that the covered object could indeed be a toothbrush. There aren't any other common objects in this setting that fit the described location and function so well. Therefore, based on the context provided by the rest of the scene, the most logical conclusion is that the masked item is a toothbrush.</think><answer>toothbrush</answer>



RFT: \n<think>\nThe context suggests that the person is getting ready for work or an important event. Given the formal attire (white shirt) and the presence of grooming items like toothbrushes and toiletries, it's reasonable to assume they might be preparing for a professional setting such as an office. Therefore, the most likely object hidden under the mask could be a tie.\n</think>\n<answer>\ntie</answer>



SFT + RFT: \n<think>\nThe masked area appears to be positioned where one would expect to find a toothbrush holder or similar item typically found next to a sink for convenience during grooming routines like brushing teeth. The space directly below the mirror aligns well with typical placement for such items. Given the context of a bathroom setting where personal care products are commonly stored, it's reasonable to infer that the masked object could indeed be a toothbrush holder due to its proximity to other toiletries and the common use of these spaces for storage. There aren't any other objects in the immediate vicinity that suggest another function for the masked area, further supporting the conclusion that it is likely a toothbrush holder.</think><answer>toothbrush holder</answer>

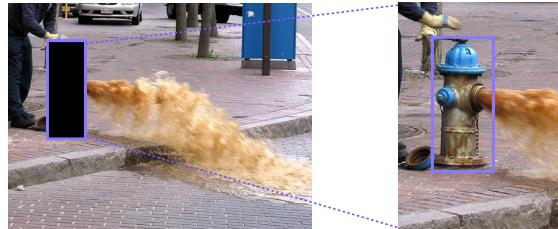


RFT with prior sampling: \n<think>\nThe masked area appears to be positioned directly above what looks like a sink or countertop where various toiletries would typically be placed. Given the context of a bathroom setting, common items such as toothbrushes, toothpaste, and other personal care products are often stored nearby. The presence of a mirror reflecting the person's reflection suggests they might be grooming themselves, which aligns with typical activities involving these objects.\n\nConsidering all these factors, it is reasonable to infer that the masked section contains a toothbrush due to its proximity to the sink and the usual placement of such items in bathrooms.</think>\n<answer>toothbrush</answer>



GT: fire hydrant

Base: Dog 😕



Prompt: \n<think>\nThe context suggests that the black rectangular mask is concealing something important or valuable. Given the urban setting and the presence of water flowing out from under it, it's plausible that the object could be related to plumbing or utilities.\n</think>\n<answer>\nWater main break or utility pipe</answer>



SFT: \n<think>\nThe masked area appears to be positioned where one would expect a fire hydrant to be located next to a sidewalk and close enough for someone to interact with it. The person's hand holding what looks like a tool suggests they might be working on or near something related to water management. Given these observations, along with the typical placement of fire hydrants alongside streets and sidewalks, it is reasonable to infer that the masked object is indeed a fire hydrant.</think><answer>fire hydrant</answer>



RFT: \n<think>\nThe context suggests that this is an urban setting, possibly near a street or sidewalk. Given the shape of the mask and its placement, it could be related to public transportation like buses or trains. Considering the water spilling out, it might indicate a malfunctioning system such as a bus stop or train station. Therefore, the most likely object hidden beneath the mask is a bus.\n</think>\n<answer>\nbus</answer>



SFT + RFT: \n<think>\nThe masked area appears to be positioned where one would expect a fire hydrant to be located given its proximity to the sidewalk and the presence of another similar object nearby. The water flowing from what seems like a spigot suggests it could indeed be a fire hydrant. Commonly, fire hydrants have a cylindrical shape and are typically placed along streets for easy access during emergencies. The size and position relative to other urban elements further support this conclusion. Therefore, based on these observations and typical urban infrastructure practices, it is reasonable to infer that the masked object is a fire hydrant.</think><answer>fire hydrant</answer>



RFT with prior sampling: \n<think>\nThe masked area appears to be positioned where one would expect a fire hydrant to be located on a sidewalk next to a street. Given its size relative to other objects like cars or pedestrians, it aligns well with what we typically see under such conditions. Additionally, the presence of water spilling out suggests that whatever is underneath could indeed be a fire hydrant due to its function in releasing water for firefighting purposes.\n</think><answer>fire hydrant</answer>



Figure 5: Case study of MPCC.