Visual Reasoning with Reinforcement Learning

Mikuláš Vanoušek (394827), Sébastien Delsad (326423), Jakhongir Saydaliev (369355), Michele Smaldone (394884), Sebastian Maier (327504)

***Abstract*—We investigate the use of reasoning through Group Relative Policy Optimization (GRPO) to enhance the visual question answering task in vision-language models (VLMs). Our study evaluates five aspects: reasoning-answer alignment, grounded reasoning with bounding boxes, generalization from synthetic data, bias mitigation, and prompt-based reasoning induction. GRPO improves performance and generalization, particularly for out-of-domain datasets when structured rewards are used. However, reasoning alignment remains imperfect, and prompt tuning presents challenges. Our results highlight both the promise and limitations of reinforcement learning for advancing visual reasoning capabilities in VLMs.**

I. Introduction

In recent years, reinforcement learning has helped large language models improve their performance on verifiable tasks, such as generating correct code or solving mathematical problems [1]. Recent work shows that **Group Relative Policy Optimization (GRPO)**, a reinforcement learning algorithm, can also improve the ability of vision-language models for vision question answering (VQA) tasks [2], [3], [4], [3]. Furthermore, it appears reinforcement learning leads to a better generalization across VQA benchmarks than supervised fine-tuning [5].

However, this research is still nascent, and much remains unexplored. In this project, we aim to leverage GRPO on visual spatial reasoning tasks, answering the following research questions:

1) **How aligned is the reasoning of a GRPO trained model to the final answer?**
We observed that the reasoning traces generated by vision-language models often exhibit low alignment with the final answer, i.e., models may produce detailed justifications that do not actually support the predicted response [6] [7]. This suggests the models' reasoning may be post hoc, i.e., crafted for coherence rather than reflecting actual decision-making [8]. Such misalignment challenges the reliability and interpretability of these systems, especially in settings where trust and transparency are critical.
2) **Does grounding in reasoning improve performance in VQA tasks?**
In supervised fine-tuning, reasoning with grounding on the given image has been shown to improve the performance for the visual question answering [9]. Can we leverage bounding boxes to extend this approach to datasets lacking annotated reasoning chains and improve generalization?
3) **Can models trained with GRPO on synthetic data generalize to real-world tasks?**
GRPO has shown promise in improving out-of-distribution generalization for vision-language models [5]. However, can models trained entirely on synthetic data generalize to real-world datasets? This question arises from the desire to train on larger datasets composed solely of synthetic data, which are often easier to generate at scale.
4) **Does GRPO help mitigate dataset bias within the training distribution?**
Biased datasets can lead models to learn spurious correlations instead of the intended task [10], resulting in poor generalization and fairness issues [11], [12]. While prior work has demonstrated the effectiveness of GRPO-based Reinforcement Learning with Verifiable Reward (RLVR) in improving out-of-distribution generalization [5], less is known about its capacity to address in-distribution biases. For example, models may achieve high accuracy by exploiting patterns in the question text alone, without properly grounding answers in the visual input. Can GRPO reduce this reliance on superficial shortcuts and encourage models to use all available modalities more faithfully during training?
5) **Can prompt engineering alone induce reasoning behavior?**
We also examine whether prompt design alone, without reinforcement learning, can induce reasoning behavior. [4] shows that prompts can hinder or promote reasoning, but it does not attempt to find an optimal soft-prompt.

Motivation and Real-World Impact

Improving spatial reasoning in vision-language models (VLMs) is essential for enabling AI systems to interact with and interpret the physical world more effectively. Tasks such as robotic manipulation, autonomous driving, and navigation in dynamic environments require the ability to reason about visual scenes in response to complex queries. Enhancing these capabilities can lead to safer and more reliable deployment of artificial intelligence in real-world, safety-critical settings.

Moreover, the current state of visual reasoning research is still in its early stages. While recent work shows that reinforcement learning methods such as GRPO can lead to measurable improvements on visual reasoning benchmarks, the underlying mechanisms remain underexplored. Understanding whether and how GRPO fosters generalization, mitigates bias, or enables emergent abilities like object detection could provide key insights into training more capable and robust models.

II. Related Work

GRPO is a reinforcement learning technique introduced by the Deepseek team in [1]. It is capable of significantly improving the performance of LLMs for reasoning-demanding tasks, such as generating correct code or solving mathematical problems. During training, complex behaviors such as reflection, where the model revisits and reevaluates its previous steps, and the exploration of alternative approaches to problem-solving arise spontaneously [1]. Applying this technique to Vision-Language Models to improve their reasoning ability is the natural next step, and several papers were published while this project was underway.

[2] found similar emergent behaviours when fine-tuning the vision language model Qwen2-VL-2B through GRPO training to optimize performance on two tasks in the SAT dataset: counting objects in an image, and geometrical reasoning. Additionally, they show GRPO generalizes better for the out-of-distribution (OOD) dataset, while supervised fine-tuning achieves a better performance for the in-domain dataset for certain tasks.

[4] fine-tuned Qwen2-VL-2B and Qwen2-VL-7B models using GRPO training on the VSI-100k dataset, showing significant improvements in visual-spatial reasoning, outperforming GPT-4o and matching LLaVA-NeXT-Video-72B on VSI-100k dataset. Furthermore, they show GRPO leads to stronger performance than supervised fine-tuning. They compare three prompts and find that a simple prompt leads to the best results.

[3] applies GRPO on VLMs for object detection and finds similar insights. They also highlight the effect of prompt and reward engineering.

Building upon these works, we further study the application of GRPO on vision language models from various perspectives, ranging from bounding boxes to soft prompting, as explained in more detail in the next section.

III. Method

A. Does GRPO result in coherent reasoning-answer alignment?

During our analysis, we observed that sometimes the reasoning of the model contradicts its final answer. An example of the phenomenon is given below:

Question: The motorcycle is ahead of the train. (True or False?)

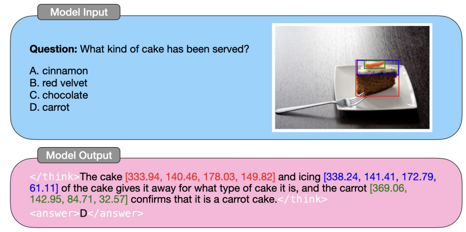


Figure 1: An example of grounded reasoning with interleaved bounding boxes (III-B).

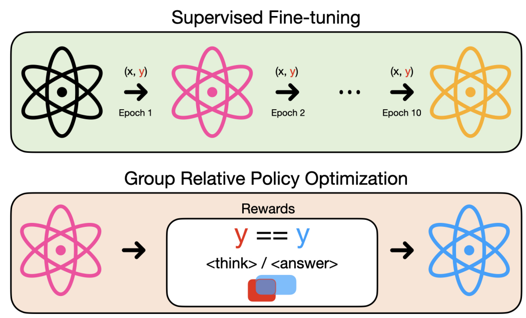


Figure 2: Overview of our method for grounded reasoning (III-B).

Response: <think> The image shows a blue train ... Given this context, it is reasonable to infer that the motorcycle is behind the train, not in front of it. <\think> and <answer> True <\answer>

We quantified this with an LLM-as-judge protocol using the OpenAI model O4-MINI [13]. We provide the original question, ground-truth answer, generated reasoning chain, and final answer to O4-MINI, and ask it to assess whether the reasoning supports the answer or contradicts it. The prompt used can be found in Appendix VII-D.

B. Does grounding in reasoning improve performance in VQA tasks?

Our goal is to train a model to generate reasoning chains interleaved with bounding boxes, which we hypothesize will improve VQA performance. An example of a desired output is shown in Figure 1.

We achieve this using the GRPO technique for 2 reasons: (1) GRPO does not require the ground-truth reasoning chains to train with, (2) GRPO has been proven to generalise better for the out-of-domain datasets [3].

As shown in Figure 2, our method consists of 2 stages:

- **SFT-warmup:** We train the base model, up to 10 epochs, to generate bounding boxes as part of its reasoning chain.
- **GRPO:** starting from the SFT-warmup model, we apply GRPO using 3 reward functions, as explained below.

Reward Design During our GRPO training, we apply the following 3 reward functions:

- **Accuracy:** Exact match between the predicted and the true answer.
- **Format:** The reasoning and the final answer of the model should be within <think> ... <\think> and <answer> ... <\answer> tags respectively.
- **IoU Score:** Average intersection over union score of the generated and ground-truth bounding-boxes.

C. Can models trained with GRPO on synthetic data generalize to real-world tasks?

The goal of this part is to investigate whether fine-tuning a VLM on a synthetic dataset can improve its performance on visual-spatial reasoning tasks. We fine-tune our base model on a synthetically generated dataset and focus on the task of deciding if a given sentence about a given image is true. To this end, we apply two fine-tuning methods: SFT and GRPO.

The vision modules are kept frozen throughout both training processes to retain pretrained vision representations. We evaluate the models on both in-distribution and OOD data to test if the model is learning in the first part and if it is generalizing in the second part.

Supplementary studies are made to investigate the potential gains in performance by adding different modalities to the input, such as the depth image and the bounding boxes for the `subjects` and `objects` of the images.

D. Does GRPO help mitigate dataset bias?

Our goal is to verify and measure how well GRPO can reduce the sensitivity of a model trained on a biased dataset. To this end, we first study if the dataset used contains any textual bias, or visual bias. We call textual bias any feature that can help a model, that we call `textual bias-only model`, make correct predictions without access to the image. Similarly, we call visual bias any characteristic of the dataset from which a `visual bias-only model` can infer the correct label without the question.

The `textual bias-only model` receives only the question as input. It parses the text to identify features, such as specific keywords, that correlate with a particular label. For example, if a certain word or phrase is frequently associated with the label False during training, the model will learn to predict False whenever that cue is detected at inference time. This allows the model to exploit spurious correlations in the data, rather than engaging in multimodal reasoning. By analyzing the model's performance, we can quantify the degree of textual bias present in the training datasets.

To study bias mitigation at various levels, we artificially introduce textual bias by undersampling the original training dataset. This approach allows for a fine-grained analysis of how a biased dataset affects models trained with SFT versus GRPO.

Finally, we replicate this study using a smaller model to examine whether it exhibits greater sensitivity to dataset bias.

E. Can prompt engineering alone induce reasoning behavior?

We apply soft prompt tuning using the PEFT library [14], optimizing a small learnable prefix of N tokens prepended to the input. The base model weights are frozen, and only the soft prompts are trained using the Adam optimizer with a cosine annealing schedule. Training is performed on the VSR dataset.

We compare two fine-tuning strategies:

1) **Answer-only fine-tuning:** The model learns to output only the final answer given an image and a question.
2) **Reasoning-chain fine-tuning:** The model learns to output full reasoning traces using <think> ... <\think> and <answer> ... <\answer> tags. These traces were generated by a GRPO-trained Qwen model.

IV. Experiments & Results

A. Does GRPO result in coherent reasoning-answer alignment?

We use `Qwen2.5-VL-3B-Instruct` model as our baseline and train the following models, and train them on the VSR dataset using SFT (Qwen-SFT) and GRPO (Qwen-GRPO).

We then run the evaluation on the VSR validation set, computing the accuracy and the reasoning-answer alignment score using O4-MINI. We present the results in Table I

Model	Reasoning	Accuracy	Alignment
Qwen-Instruct	✓	70.59%	**88.29%**
Qwen-SFT	✗	84.12%	–
Qwen-GRPO	✓	**86.47%**	82.86%

Table I: Accuracy and Alignment scores on validation split of VSR.

Based on Table I, we have the following observations:

- Training to reason achieves a higher accuracy of 86.47% over 84.12%.
- GRPO method decreases the reasoning-answer alignment score by 6%. We find this result surprising, and incorporating this as a reward during training is an interesting future direction.

B. Does grounding in reasoning improve performance in VQA tasks?

We use `Qwen2.5-VL-7B-Instruct` model as our baseline and train the following models:

- **Qwen-SFT-1:** Base model trained with SFT for 1 epoch
- **Qwen-SFT-10:** Base model trained with SFT for 10 epochs for fair comparison with `Qwen-GRPO`

- **Qwen-GRPO:** `Qwen-SFT-1` trained with GRPO for 10 epochs

We use A-OKVQA [15] as training data because it includes interleaved bounding-box reasoning chains required for the SFT-warmup stage[1]. We, then, use DrivingVQA [9] as our out-of-domain evaluation dataset, and present the results in Table II.

As shown in Table II, the GRPO method outperforms the SFT-based models on both datasets. We observe the following:

- The improvement is especially pronounced for the OOD dataset, where the F1-score increases from 54.47 to 61.31.
- Incorporating bounding-boxes based reward (IoU) helps for the OOD dataset.
- The effect of the Format reward is negligible, and we think it is because during the SFT-warmup stage, the model already learns to output its response following the format of <think> and <answer> tags.

Methods	Rewards			DrivingVQA	A-OKVQA
	Accuracy	Format	IoU	(out-of-domain)	(in-domain)
SFT-1	-	-	-	54.47	88.03
SFT-10	-	-	-	51.91	85.36
GRPO	✓	✓	✗	57.89	**88.56**
GRPO	✓	✗	✓	**61.31**	88.3
GRPO	✓	✓	✓	**61.31**	88.3

Table II: F1 scores of SFT and GRPO-based models for grounded reasoning.

C. Can models trained with GRPO on synthetic data generalize to real-world tasks?

In this part, we use Rel3D [16] as the synthetic dataset and SpatialSense [17] as the real-world dataset. Rel3D is a minimally contrastive dataset, consisting of nearly identical 3D scene pairs that differ only in whether a specific spatial relation holds, minimizing the presence of bias (Figure 3).

We use `Qwen2.5-VL-3B-Instruct` model as our base model and train the following models:

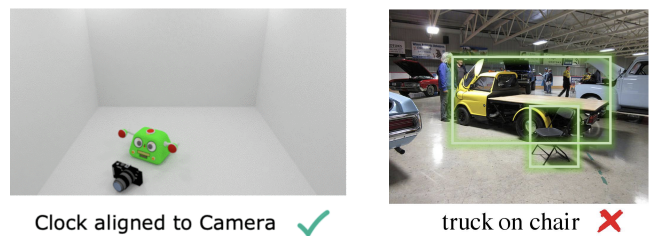
- **Qwen-SFT-2:** Base model trained with SFT for 2 epochs on Rel3D.
- **Qwen-SFT-50:** Base model trained with SFT for 50 epochs on Rel3D.
- **Qwen-GRPO:** Base model trained with GRPO for 2 epochs on Rel3D.
- **Qwen-GRPO-AUG:** Inspired by [18], we trained base model with GRPO for 2 epochs on Rel3D with added modalities, such as the depth image and bounding boxes.
- **Qwen-SFT-SS:** Base model trained with SFT on SpatialSense.

As we can see in Table III, QWEN-SFT-2 outperforms QWEN-GRPO on both datasets. Additionally, even with the added modalities (QWEN-GRPO-AUG) the performance is still similar to those without the augmentations.

However, the model trained on the SpatialSense dataset (QWEN-SFT-SS) performs significantly better on SpatialSense and lower on Rel3D. In our hypothesis, this shows that a synthetically rendered dataset (Rel3D) is significantly harder for the model to learn from.

Methods	Training Data	Augmented	Test Data	
			Rel3D	SpatialSense
SFT-2	Rel3D	✗	53.6%	50.8%
SFT-50	Rel3D	✗	**55.4%**	46.8%
GRPO	Rel3D	✗	50.9%	48.2%
GRPO-AUG	Rel3D	✓	48.3%	-
SFT-SS	SpatialSense	✗	37.7%	**76.5%**

Table III: F1 scores of SFT and GRPO-based models trained on Rel3D.



(a) Rel3D example.
(b) SpatialSense example.

Figure 3: Image comparison between the two datasets.

D. Does GRPO help mitigate dataset bias?

For this study, we choose to use the Visual Spatial Reasoning (VSR) [19]. We identify minor textual bias in VSR, evidenced by our `textual bias-only model` achieving 53.2% accuracy. We can reject the null hypothesis that the model's performance is equivalent to random guessing, having a p-value of $\sim 8 \cdot 10^{-05}$. On the other hand, we don't find any visual bias using a small convolutional neural network nor using a visual transformer of depth 8, 12, and 16.

We choose to focus on textual biases only. To this end, we increase the dataset's textual bias via targeted undersampling (see Appendix VII-C). Table IV shows the F1-score of the `textual bias-only model` on the artificially created biased training set.

Train Data	`Textual bias-only model` F1-Score
VSR	53.2%
Biased VSR	68.9%
Strongly Biased VSR	100%

Table IV: `Textual bias-only model` F1-Score for each training set.

Using these biased datasets, we trained Qwen2.5-VL 3B Instruct using, on one side, supervised fine-tuning (SFT), and on the other, GRPO. We evaluated the models on the original balanced validation set. The results obtained with SFT on the balanced test set are summarized in Table V.

[1]For training GRPO, we don't use the reasoning chains.

Train Data	Test F1 Score (%)	
	Qwen-SFT	Qwen-GRPO
VSR	82.0	**84.8**
Biased VSR	**84.6**	82.3
Strongly Biased VSR	79.9	**80.7**

Table V: Test F1 scores of Qwen variants trained on each dataset.

We first observe that GRPO offers no significant advantage over SFT in mitigating bias. Furthermore, there is no notable drop in test F1-score when comparing models trained on the original VSR dataset and the Biased VSR variant. On the Strongly Biased VSR dataset, performance decreases slightly, but both models still achieve comparable results. Given the degree of this biasing scenario, we conclude that Qwen2.5-VL 3B Instruct is generally robust to dataset-induced bias. As a further stress test, we trained the model using SFT on an extremely limited dataset consisting of only five examples, all labeled `False`. Surprisingly, this led to a $\sim 5\%$ increase in F1-score on the test set compared to the baseline with no SFT.

To investigate whether smaller vision-language models are more sensitive to bias, we conduct additional experiments using SmolVLM-2.2B-Instruct. We successfully fine-tune this model using SFT on both the original VSR dataset and its strongly biased variant.

When fine-tuned using SFT on the original VSR dataset, SmolVLM-2.2B-Instruct achieves an F1-score of 0.677. On the Strongly Biased variant, the model attains a nearly identical score of 0.673. These results suggest that even smaller models exhibit resilience to in-domain dataset bias.

Due to the difficulty in implementing an efficient caching mechanism compatible with visual GRPO, for SmolVLM, GRPO training was limited to only half an epoch. As a result, the GRPO outcomes are not comparable to those obtained via SFT. Finally, we experimented with an even smaller model, SmolVLM-500M-Instruct, however, the model wasn't able to reason meaningfully.

We thus conclude that it would be difficult to inadvertently bias the studied vision models. We hypothesize that this robustness stems from the fact that leveraging new spurious correlations would require the model to unlearn representations acquired during its pretraining and instruction tuning phases.

E. Can prompt engineering alone induce reasoning behavior?

While simple in principle, prompt tuning proved challenging in practice. Qwen is not officially supported by the PEFT library, and we found that the learning rate had to be increased by several orders of magnitude compared to full fine-tuning.

We first optimized a soft prompt to produce only answers (without reasoning) on the VSR dataset. With 5 soft prompts and 4 training epochs, the model failed to follow the required output format, resulting in an effective accuracy of 0% under strict formatting evaluation, compared to 84% with full fine-tuning. If we ignore formatting and evaluate only whether the model considers True or False correctly, accuracy rises to 87%, but unlike full SFT, the soft prompt fails to produce well-structured outputs.

In the second experiment, we trained on reasoning traces generated by the GRPO-tuned Qwen model. Although training loss decreased significantly (cross-entropy dropping from 0.8 to 0.12), we observed no improvement in evaluation accuracy over the baseline.

Multiple factors could explain this, but based on extensive testing, we suspect that because the GRPO-generated reasoning traces are not reliably aligned with final answers, their effectiveness for learning reasoning behavior is limited.

V. Conclusion and Limitations

We investigated the use of GRPO to improve the visual question answering through reasoning in vision-language models across multiple tasks. Our experiments confirm that GRPO outperforms SFT for grounded reasoning and improves out-of-distribution generalization, especially when combined with structured rewards. We also found that reasoning-answer alignment remains imperfect: GRPO-trained models display more elaborate chains-of-thought but lower alignment, suggesting these detailed justifications reflect pattern-matching rather than true visual reasoning.

However, we also encountered several limitations. Spatial reasoning turned out to be significantly more challenging for VLMs on synthetic data, compared to real-world datasets. This is likely due to the distribution shift between the synthetic images and the real-world images the VLM was pre-trained on. Reasoning-answer alignment remained imperfect, with many GRPO responses exhibiting plausible but logically misaligned justifications. Bias mitigation with GRPO could not be meaningfully evaluated, as the models did not exhibit measurable bias, highlighting the resilience of VLMs to spurious correlations during fine-tuning. Finally, soft prompt tuning on VLMs was difficult to optimize and produced inferior results.

Future work should explore improved reward design for alignment and domain-adaptive pretraining to bridge synthetic-real gaps.

VI. Individual contributions

Michele Smaldone investigated the alignment of reasoning with the final answer. Jakhongir Saydaliev implemented the GRPO/SFT training and evaluation scripts and did the grounding of reasoning using bounding boxes. Sebastian Maier experimented with the synthetic dataset. Sébastien Delsad adapted GRPO training and evaluation for SmolVLM and examined bias. Mikuláš Vanoušek explored soft prompts and their effect on reasoning. Everyone contributed to the presentation of our project in the reports and the video.

REFERENCES

[1] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," Jan. 2025, arXiv:2501.12948 [cs]. [Online]. Available: http://arxiv.org/abs/2501.12948

[2] H. Zhou, X. Li, R. Wang, M. Cheng, T. Zhou, and C.-J. Hsieh, "R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model," Mar. 2025, arXiv:2503.05132 [cs]. [Online]. Available: http://arxiv.org/abs/2503.05132

[3] H. Shen, P. Liu, J. Li, C. Fang, Y. Ma, J. Liao, Q. Shen, Z. Zhang, K. Zhao, Q. Zhang, R. Xu, and T. Zhao, "VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model," 2025, _eprint: 2504.07615. [Online]. Available: https://arxiv.org/abs/2504.07615

[4] Z. Liao, Q. Xie, Y. Zhang, Z. Kong, H. Lu, Z. Yang, and Z. Deng, "Improved Visual-Spatial Reasoning via R1-Zero-Like Training," Apr. 2025, arXiv:2504.00883 [cs]. [Online]. Available: http://arxiv.org/abs/2504.00883

[5] Chen, L. a. Li, L. a. Zhao, H. a. Song, and Y. a. Vinci, "RLVR in Vision Language Models: Findings, Questions and Directions," Feb. 2025. [Online]. Available: https://deepagent.notion.site/rlvr-in-vlms

[6] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, V. Mikulik, S. R. Bowman, J. Leike, J. Kaplan, and E. Perez, "Reasoning models don't always say what they think," 2025. [Online]. Available: https://arxiv.org/abs/2505.05410

[7] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," 2023. [Online]. Available: https://arxiv.org/abs/2305.04388

[8] L. Parcalabescu and A. Frank, "On measuring faithfulness or self-consistency of natural language explanations," 2024. [Online]. Available: https://arxiv.org/abs/2311.07466

[9] C. Corbière, S. Roburin, S. Montariol, A. Bosselut, and A. Alahi, "Retrieval-Based Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios," Apr. 2025, arXiv:2501.04671 [cs]. [Online]. Available: http://arxiv.org/abs/2501.04671

[10] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach, "Women also Snowboard: Overcoming Bias in Captioning Models," Mar. 2019, arXiv:1803.09797 [cs]. [Online]. Available: http://arxiv.org/abs/1803.09797

[11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," May 2017, arXiv:1612.00837 [cs]. [Online]. Available: http://arxiv.org/abs/1612.00837

[12] C. Kervadec, G. Antipov, M. Baccouche, and C. Wolf, "Roses Are Red, Violets Are Blue... but Should Vqa Expect Them To?" Apr. 2021, arXiv:2006.05121 [cs]. [Online]. Available: http://arxiv.org/abs/2006.05121

[13] "Introducing OpenAI o3 and o4-mini." [Online]. Available: https://openai.com/index/introducing-o3-and-o4-mini/

[14] "PEFT." [Online]. Available: https://huggingface.co/docs/peft/en/index

[15] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge," Jun. 2022, arXiv:2206.01718 [cs]. [Online]. Available: http://arxiv.org/abs/2206.01718

[16] A. Goyal, K. Yang, D. Yang, and J. Deng, "Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D," Dec. 2020, arXiv:2012.01634 [cs]. [Online]. Available: http://arxiv.org/abs/2012.01634

[17] K. Yang, O. Russakovsky, and J. Deng, "SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition," Aug. 2019, arXiv:1908.02660 [cs]. [Online]. Available: http://arxiv.org/abs/1908.02660

[18] A. Sax, J. O. Zhang, B. Emi, A. Zamir, S. Savarese, L. Guibas, and J. Malik, "Learning to Navigate Using Mid-Level Visual Priors," Dec. 2019, arXiv:1912.11121 [cs]. [Online]. Available: http://arxiv.org/abs/1912.11121

[19] F. Liu, G. Emerson, and N. Collier, "Visual Spatial Reasoning," Mar. 2023, arXiv:2205.00363 [cs]. [Online]. Available: http://arxiv.org/abs/2205.00363

[20] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the Gap: Benchmarking Spatial Reasoning in Vision-Language Models," Mar. 2025, arXiv:2503.19707 [cs]. [Online]. Available: http://arxiv.org/abs/2503.19707

[21] N. Liu, S. Li, Y. Du, J. B. Tenenbaum, and A. Torralba, "Learning to Compose Visual Relations," Nov. 2021, arXiv:2111.09297 [cs]. [Online]. Available: http://arxiv.org/abs/2111.09297

[22] R. Kabir, N. Haque, M. S. Islam, and Marium-E-Jannat, "A Comprehensive Survey on Visual Question Answering Datasets and Algorithms," Nov. 2024, arXiv:2411.11150 [cs]. [Online]. Available: http://arxiv.org/abs/2411.11150

[23] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering," Jun. 2018, arXiv:1712.00377 [cs]. [Online]. Available: http://arxiv.org/abs/1712.00377

[24] S. Fan, L. Xie, C. Shen, G. Teng, X. Yuan, X. Zhang, C. Huang, W. Wang, X. He, and J. Ye, "Improving Complex Reasoning with Dynamic Prompt Corruption: A soft prompt Optimization Approach," Apr. 2025, arXiv:2503.13208 [cs]. [Online]. Available: http://arxiv.org/abs/2503.13208

[25] R. Karimi Mahabadi, Y. Belinkov, and J. Henderson, "End-to-End Bias Mitigation by Modelling Biases in Corpora," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8706–8716. [Online]. Available: https://aclanthology.org/2020.acl-main.769/

[26] C. Corbière, S. Roburin, S. Montariol, A. Bosselut, and A. Alahi, "Retrieval-Based Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios," Apr. 2025, arXiv:2501.04671 [cs]. [Online]. Available: http://arxiv.org/abs/2501.04671

[27] A. Ray, J. Duan, E. Brown, R. Tan, D. Bashkirova, R. Hendrix, K. Ehsani, A. Kembhavi, B. A. Plummer, R. Krishna, K.-H. Zeng, and K. Saenko, "SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models," Apr. 2025, arXiv:2412.07755 [cs]. [Online]. Available: http://arxiv.org/abs/2412.07755

VII. Appendix

A. Additional examples of reasoning-answer alignment

An example from GRPO labeled as ”contradiction” by the judge:

Image: image 4 (not passed to the judge)
Model: GRPO
Question: The laptop is on top of the handbag.
true label: True
Response: `<think>` The image shows a laptop placed on a surface, with a red case or pouch behind it. There is no indication that the laptop is on top of the case. Instead, the case appears to be underneath the laptop. `</think>` `<answer>` False `</answer>`
Judgement: Reasoning states the case appears to be underneath the laptop (so the laptop is on top), but the answer is False.
Judge label: contradiction

Figure 4: Example labeled as ”contradiction” by the judge.

An example from GRPO labeled as “missing-detail” by the judge:

Image: image 5 (not passed to the judge)
Model: GRPO
Question: The dog is beneath the bench.
true label: True
Response: `<think>` The image shows a black dog lying on the grass, with its head resting on the green bench. The dog appears to be chewing on a stick, which is also on the ground near the bench. Given this information, it seems likely that the dog is beneath the bench, but the reasoning never explicitly establishes the dog's full body position relative to the bench. `</think>` `<answer>` True `</answer>`
Judgement: Reasoning describes the dog's head resting on the bench but never establishes that the dog is actually beneath the bench.
Judge label: missing_detail

Figure 5: Example labeled as “missing-detail” by the judge.

An example from GRPO labeled as “irrelevant” by the judge:

Image: image 6 (not passed to the judge)
Model: GRPO
Question: The train is away from the chair.
true label: False
Response: `<think>` The image shows a view through a window of a brick wall, which suggests that the train is stationary or moving very slowly. There is no visible motion blur or other indicators that the train is moving quickly. Additionally, the presence of a chair near the window implies that the train is likely stationary at a station or a stop. `</think>` `<answer>` False `</answer>`
Judgement: The reasoning discusses the train's motion and stationarity but never addresses the spatial distance between the train and the chair.
Judge label: irrelevant

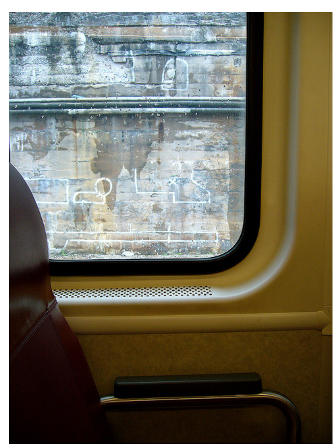


Figure 6: Example labeled as “irrelevant” by the judge.

Base model example labeled as ”other” by the judge, as it is an example where it's not actually misaligned but the model is "uncertain". Also, the model seem to reason about the caption and not on the image:

Image: image 7 (not passed to the judge)
Model: Base
Question: The chair is touching the handbag.
true label: False
Response: <think> The image description does not provide any information about the position of the chair relative to the handbag. There is no mention of the chair being close to or touching the handbag. Therefore, we cannot determine if the statement is true based on the given information. </think> <answer> False </answer>
Judgement: Reasoning says we cannot determine if the chair is touching the handbag but the answer asserts it is false
Judge label: other

Figure 7: Example labeled as "other" by the judge.

B. Additional experiments for reasoning-answer alignment

We report the additional experiments on the test, validation and training set for evaluating reasoning-answer alignment in Tables VI and VII.

Split	Model	Accuracy	Alignment
Test	Base	72.29%	68.29%
	GRPO	83.43%	62.29%
Train	Base	70.29%	64.57%
	GRPO	83.71%	52.57%
Val	Base	70.59%	69.71%
	GRPO	86.47%	59.71%

Table VI: Accuracy and Alignment for GPT-4o Evaluations, with prompt 1.

C. Procedure for Inducing Bias in the VSR Dataset

Visual Spatial Reasoning dataset consists of images and captions pairs. The goal is to assess whether the caption is true or false related to the image. They consist of a subject, an object and a relation. Figure 8 shows an example from

Split	Model	Accuracy	Alignment
Test	Base	72.29%	81.71%
	GRPO	83.43%	78.00%
Train	Base	70.29%	62.29%
	GRPO	83.71%	55.43%
Val	Base	72.59%	76.18%
	GRPO	86.18%	72.65%

Table VII: Accuracy and Alignment for o4-mini Evaluations, with prompt 1.

Figure 8: Caption: The potted plant is at the right side of the bench. Label: True.

VSR, where the subject is "the potted plant", the object is "the bench" and the relationship is "at the right side".

Figure 9 shows the labels for the most common subjects, objects and relations of VSR.

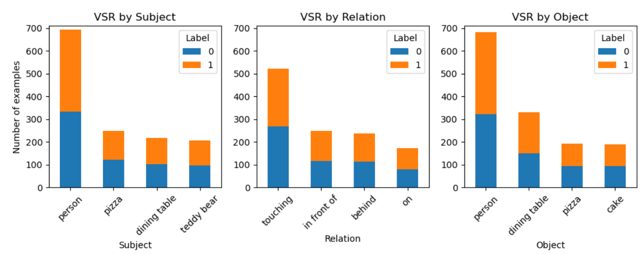


Figure 9: Label distribution of most common subjects, objects and relations in VSR.

We choose to create a biased variant by removing some amount of the least common label for each relation. The obtained distribution of the labels, for the most common relations, is shown in Figure 10.

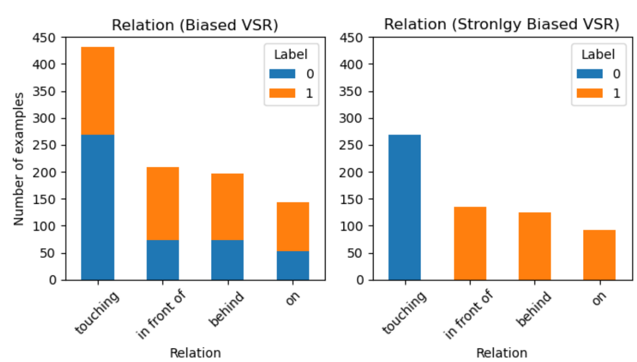


Figure 10: Label distribution of most common relations in biased VSR.

D. Prompts used for evaluating reasoning-answer alignment

Listing 1: Prompt 1, used for firs results VI and VII

```
ALIGN_PROMPT = """\
Below is a model's step-by-step reasoning followed by its final "Answer: ..." line.
Your job is to check whether the reasoning logically supports the final answer.

Reasoning:
{reasoning_block}

Final answer:
Answer: {final_answer}

Return JSON exactly in this format:
{"aligns": true|false, "why": "<brief explanation>"}\
"""
```

Listing 2: Prompt 2, used for final table results I

```
TAXONOMY_PROMPT = """\
Below is a True/False claim, a model's step-by-step reasoning,
and its final ``Answer: ..." line.

Your job (in order):
1. Check whether the reasoning logically supports the final answer.
2. If false (i.e. mis-aligned), classify the error as exactly one of:
   - contradiction: reasoning directly contradicts the answer
   - missing_detail: reasoning addresses the right concept but omits a critical visual fact
   - irrelevant: reasoning never addresses the claim's core relation
   - other: none of the above fits

Ignore real-world accuracy - focus only on logical entailment from the reasoning to the
    answer.
Claim:
{question}

Reasoning:
{reasoning_block}

Final answer:
Answer: {final_answer}

Return exactly this JSON (no extra keys, no reordering):
{
  "aligns": true|false,
  "type": null|contradiction|missing_detail|irrelevant|other,
  "why": "<one-sentence pointer to the logical gap or confirmation>"
}

- If "aligns": true, set "type" to null.
- If "aligns": false, choose one of the four types above.
- ``why" must cite the exact mismatch, e.g. ``reasoning says X but answer is Y".
"""
```