Fine-tuning SmolVLM-256M with SFT and GRPO for Object-State Understanding in SimplerEnv

Anton Koksharov [1]

Abstract

I investigate whether a compact 256M-parameter vision-language model (SmolVLM) can be fine-tuned for object-state reasoning tasks. I target binary yes/no queries such as "Is the carrot on the plate?" and measure performance through accuracy, precision, recall, and F1-score. I evaluated three approaches: prompt technique for model, SFT and GRPO with environment-derived rewards.

1. Introduction

Recent advances in vision-language models (VLMs) demonstrate strong zero-shot reasoning but require billions of parameters. For robotics, where low-latency, low-memory models are critical, compact models are appealing. In this work, I use SmolVLM-256M and evaluate its ability to perform binary state reasoning in the **SimplerEnv** simulator. The task of detecting whether an object is held, placed on a plate, or located on clothin is the focus.
I also tried to teach the model to determine the distance between the actual position of the object and the target position. To do this, I first fine-tuned the model to predict data in the "yes/no, distance" format, and then applied GRPO to improve the accuracy of the predictions.

2. Methods

2.1. prompt technique

To improve the quality of answers to the questions, several approaches to writing prompts were used. In general, the main idea of my approach is to use the color representation of the object instead of its actual name.

[1]EECS Department, BMSTU, Moscow, Russia. Correspondence to: Anton Koksharov <MrAnton07@mail.ru>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2.2. Supervised Fine-Tuning (SFT)

I collected an image-question-answer dataset from SimplerEnv and fine-tuned SmolVLM with LoRA adapters. Prompts were intentionally kept short to reduce token overhead, as long system prompts degraded performance. Training used cross-entropy loss over answer tokens.

2.3. Group Relative Policy Optimization (GRPO)

As a result of the work, two models were obtained, train with GRPO. The first model used only the first reward function and was trained to give binary answers, the second model used both functions during training, and is also able to predict the distance between objects. Also, during GRPO training, a special "SYSTEM PROMPT" was used, which is a sentence of the form "Answer strictly in the format: yes/no" for a model that predicts only a binary label and "Answer strictly in the format: yes/no, ¡number¿ Example: yes, 0.021 Example: no, 0.513" for a model that predicts distance. This prompt is added to the text immediately after the user's question.

The reward functions:

- **Yes/No reward:** reward = 1 if the prediction matches the ground truth, 0 otherwise.
- **Distance reward:** reward = -MSE between predicted and true distances.

3. Evaluation Protocol

I use confusion matrices to derive:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad \text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

4. Results

Metrics were calculated directly from confusion matrices from vanila, SFT, GRPO - tuned model evaluations.

4.1. Zero-shot results: general questions

First, I asked the model general questions and assessed the extent to which the model could perceive semantic connections.

Question: Is the gripper holding an object?

$$\begin{bmatrix} 20 & 0 \\ 1 & 19 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the object is held by the gripper	0.975	0.952	1.000	0.976

4.2. Zero-shot results: Carrot on Plate

To evaluate the zero-shot abilities of the model, I chose a binary question: Is the object in the target position or not.

Question: Is the carrot on the plate?

$$\begin{bmatrix} 17 & 3 \\ 18 & 2 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.475	0.486	0.850	0.618

Question: Is the orange object (carrot) on the yellow plate in this image?

$$\begin{bmatrix} 13 & 7 \\ 11 & 9 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.550	0.542	0.650	0.591

Question: Is the orange object on the yellow object in this image?

$$\begin{bmatrix} 18 & 2 \\ 10 & 10 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.700	0.643	0.900	0.750

Question: Is the carrot on the plate? (Full episode)

$$\begin{bmatrix} 3 & 2 \\ 27 & 2 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.147	0.100	0.600	0.171

4.3. Zero-shot results: Spoon on Clothin

Question: Is the spoon on the clothin?

$$\begin{bmatrix} 1 & 9 \\ 4 & 6 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the spoon are on the clothin	0.350	0.200	0.100	0.133

Question: Is the green/yellow object (spoon) on the blue clothin in this image?

$$\begin{bmatrix} 5 & 5 \\ 2 & 8 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the spoon are on the clothin	0.650	0.714	0.500	0.588

Question:
Is the green/yellow object on the blue object in this image?

$$\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the spoon are on the clothin	1.000	1.000	1.000	1.000

4.4. SFT results: Carrot on Plate

After the SFT, I assessed how much better the model started to cope with the target task, and then checked whether the quality had deteriorated on other tasks.

Question:
Is the carrot on the plate?

$$\begin{bmatrix} 20 & 0 \\ 0 & 20 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	1.000	1.000	1.000	1.000

Question:
Is the orange object on the yellow object in this image?

$$\begin{bmatrix} 19 & 1 \\ 0 & 20 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.975	1.000	0.950	0.974

Question:
Is the carrot on the plate?
(Full episode)

$$\begin{bmatrix} 5 & 0 \\ 1 & 28 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.971	0.833	1.000	0.909

4.5. GRPO results: Carrot on Plate

Question:
Is the carrot on the plate?
+ SYSTEM PROMPT

$$\begin{bmatrix} 18 & 2 \\ 0 & 20 \end{bmatrix}$$

Task	Acc	Prec	Recall	F1
Determine if the carrot are on the plate	0.950	1.000	0.900	0.947

5. Discussion

During the work, the following facts were noted:

- In the answers to the question "What objects are in the scene?" carrots were found with a probability of 0.475, while if you ask the model the question "Is there an object that can be eaten?", this probability increases to 0.750. This means that with our questions we can "push" the model into correct reasoning.
- Given the large number of different objects in the world, it is difficult to understand which specific objects are depicted in the picture without fine-tuning the model. But on the other hand, we have 7 standard colors, which can often fully describe an object if we don't need to use its "features". Based on this hypothesis, experiments were conducted that showed that the zero-shot metric of the model is indeed higher if we work with the color representation of the object, rather than the nominal one.
- Fine-tuned models are noticeably better at meeting their targets, while doing no worse and sometimes even better at meeting other initial tasks.
- An attempt to train the model to predict distance yielded the following results: SFT taught the model to give answers in the right format, but at the same time, such a model demonstrates a poor understanding of distance, since most likely it simply memorized the "edge" cases. Fine-tuning of such a model using GRPO will most likely solve this problem, but unfortunately I didn't have enough time to check it.
- In the course of my work, I collected data using the Octo-small model, but its metrics were weak and the "EFFICIENCY" of its data left much to be desired. Therefore, since the main purpose of the work is experiments with VLM, I decided to use the optimal trajectory generator as a data collector.

Overall, SFT provides usable binary classification ability, but robustness varies strongly across tasks.

6. Limitations

- All experiments were conducted in a simulator, which is why transferring results of the same quality to the "real world" is not guaranteed and is most likely impossible.
- Most of the experiments was carried out with the task of "classification", while in real-world robotic tasks, it is most often necessary to solve "regression" problems
- Due to data collection using optimal trajectories, we might have missed interesting "edge" situations, but since we are rather looking at VLM capabilities in our work, this is not so critical.

7. Broader Impact

This work demonstrates that compact VLM can be adapted for simple robotic perception tasks using lightweight fine-tuning methods (LoRA, GRPO).

References

[1] Azzolini, A., team, N., Romero, D. W., and many others. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[2] Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Hausman, K., Ichter, B., Jakubczak, S., Kong, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Wang, L. Y., and Zhilinsky, U. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[3] Johnson, R., Ito, K., and Zhang, P. Question-answering benchmarks for physical reasoning. *arXiv preprint arXiv:2501.16411*, 2025.

[4] Li, S., Brown, E., and Ahmed, F. Group relative policy optimization for finetuning vision-language models. *arXiv preprint arXiv:2405.10292*, 2024.

[5] Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H. R., Fu, C., Lunawat, I., Sieh, I., Kermani, S., Levine, S., Wu, J., Finn, C., Su, H., Vuong, Q., and Xiao, T. Evaluating real-world robot manipulation policies in simulation, 2024. URL https://colab.research.google.com/github/simpler-env/SimplerEnv/blob/main/example.ipynb.

[6] Liu, J., Gao, F., Wei, B., Chen, X., Liao, Q., Wu, Y., Yu, C., and Wang, Y. What can rl bring to vla generalization? an empirical study, 2025. URL https://github.com/gen-robot/RL4VLA.

[7] Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Ben Allal, L., Lozhkov, A., Tazi, N., Srivastav, V., Lochner, J., Larcher, H., Morlon, M., Tunstall, L., von Werra, L., and Wolf, T. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[8] Park, D., Rossi, J., and Chen, M. Online reinforcement learning for vlms with policy optimization. *arXiv preprint arXiv:2402.03300*, 2024.

[9] Team, G. R. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.