Can Vision Language Models Learn from Visual Demonstrations of Ambiguous Spatial Reasoning?

Bowen Zhao
Groundlight
Seattle, WA 98122
bowen@groundlight.ai

Leo Parker Dirac
Groundlight
Seattle, WA 98122
leo@groundlight.ai

Paulina Varshavskaya
Groundlight
Seattle, WA 98122
paulina@groundlight.ai

Abstract

Large vision-language models (VLMs) have become state-of-the-art for many computer vision tasks, with in-context learning (ICL) as a popular adaptation strategy for new ones. But can VLMs learn novel concepts from visual demonstrations with ambiguous text queries, or are they limited to adapting to the output format of ICL examples? We propose a new benchmark we call Spatial Visual Ambiguity Tasks (SVAT) that challenges state-of-the-art VLMs to learn new visuospatial tasks in-context. We find that VLMs fail to do this zero-shot, and sometimes continue to fail after finetuning. However, adding simpler data to the training by curriculum learning leads to improved ICL performance. We release our benchmark generation, training, and evaluation code[1] to facilitate future research.

1 Introduction

Pretrained large vision language models (VLMs) have become essential tools and set new state-of-the-art in many general-purpose vision tasks [8, 18, 20, 34]. Extensive pretraining data allow VLMs to operate in novel domains without fine-tuning, either zero-shot, or with few-shot in-context learning (ICL) [36, 37, 39]. However, as spatial information can be ambiguous in language [33], it remains unclear what it takes to get VLMs to learn a novel visuospatial concept from visual demonstrations.

We focus specifically on the ambiguity of visual referent in the text input to the VLMs, as AI-naive users of computer vision systems in novel domains may assume background knowledge or context that the VLMs would be missing [15]. For example, the word "fiducial" in a novel industrial domain could refer to any number of markings on a piece of equipment to be aligned, and can only be disambiguated with context. Including visual information in the form of labeled ICL examples with images should lead to the desired disambiguation, but only if VLMs are able to correctly analyze the information within the example images. Existing research has demonstrated that large language models only learn the task's expected output format described in the ICL examples [21]. Recent work has also probed VLMs and found them incapable of solving straightforward tasks that specifically require visual information processing, where answers cannot be guessed from text alone [22].

In this paper, we explore how this combination of VLM and ICL limitations prevents quick adaptation of VLMs to novel tasks where the core concept of the task is introduced in the vision modality, and the query text is ambiguous. Specifically, we propose a new benchmark for ambiguous visual-spatial tasks called Spatial Visual Ambiguity Tasks (SVAT). It is a set of tasks of varying degrees of difficulty, where each task is to identify the correct spatial decision boundary in a synthesized image based on very limited ambiguous text and a number of visual demonstrations. Degrees of difficulty are achieved by varying the complexity level of the objects in the foreground and the image background, as well as the number of distracting objects (ambiguous visual referents) present in the image.

[1] https://github.com/groundlight/vlm-visual-demonstrations

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

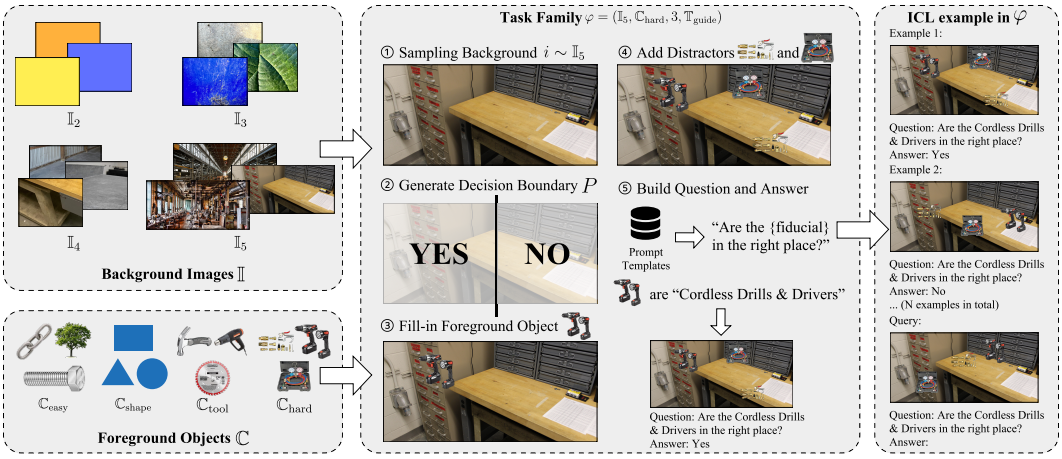


Figure 1: The dataset construction pipeline of SVAT using the task family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ as an example. For this task family, we sample the background from industrial photographs ($\mathbb{I}_5$) where the foreground objects are industrial tools $\mathbb{C}_{\text{hard}}$. Each image contains three objects ($M = 3$), and the question in the prompt mentions the target object's name ($\mathbb{T}_{\text{guide}}$).

We evaluate state-of-the-art VLMs through tasks in SVAT in three settings: zero-shot, directly finetuned, or finetuned through a curriculum learning (CL) [4] approach. Our experiments show that state-of-the-art VLMs fail at tasks in SVAT in the zero-shot setup without finetuning. While simply finetuning VLMs on SVAT can boost their performance by 5.8%-27.3% across different models, we show that curriculum learning SVAT enables VLMs to achieve better accuracy on the most challenging SVAT task, with 14.2% to 34.2% relative accuracy gains compared to direct fine-tuning.

2 SVAT Benchmark

We propose the SVAT benchmark to study the capabilities of VLMs on ambiguous visual-spatial reasoning through ICL. The benchmark consists of a series of classification tasks. Intuitive examples are shown in Fig. 1, where the core task in SVAT is to answer whether a foreground object is present within the image's "correct" location. The unusual challenge is that the "correct" location is not explicitly defined but must be inferred by the model using the in-context demonstrations. Task difficulty is varied by the information provided in the text input, the complexity of the object of interest, the number of distracting objects present in the image, and the complexity of the image background. In detail, Section 2.1 presents the dataset construction process, and Section 2.2 details the curriculum learning (CL) setup we use to improve VLMs' performance on SVAT.

2.1 Generating SVAT Datasets

To achieve the goal of examining whether VLMs can infer whether an object o presents on the "correct" place of the image v, each instance in SVAT dataset $e = (t, v, y) \in \mathbb{E}$ contains a question t paired with image v, while y is the answer (either "Yes" or "No"). The image $v = (i, o_1, ..., o_M) \in \mathbb{V}$ consists of a background image i and several foreground objects o. Among the foreground objects, o_1 is the object of interest while the rest are visual distractors for the model (details can be found in Appendix C). The sampling process of $e \in \mathbb{E}$, especially images $v \in \mathbb{V}$ is not trivial. Since we want to examine VLMs' ambiguous spatial reasoning capabilities at different difficulty levels, SVAT should be built in a manner where the fine-grained complexity of each example is controllable, ranging from one naive shape on a solid background to multiple realistic objects on a complex photograph. Therefore, we parameterize the sampling process by φ, which comprises a set of hyperparameters related to the choice of question, background, objects, and decision boundary. Each specific value of φ defines a task family $\mathbb{E}_\varphi \subset \mathbb{E}$ where each example is of a similar nature and similar difficulty level.

We parameterize the difficulty $\varphi = (\mathbb{I}, \mathbb{C}, M, \mathbb{T})$ with a known set of background images $i \in \mathbb{I}$ and a known set of categories of images to be used as foreground objects $c_j \in \mathbb{C}$, as well as the number of distracting foreground objects M and the set of possible text inputs $\mathbb{T}$. Text can be uninformative

($\mathbb{T}_{\text{none}}$), such as "Is everything okay?", or guiding the VLM ($\mathbb{T}_{\text{guide}}$) by including the name of the target object c_1 in the question. To avoid making SVAT tasks overly challenging, we simplify the decision boundary to be either a horizontal or vertical line on the image v, as shown in Fig. 1.

For all choices of φ, we keep the labels balanced. The in-context examples always include an equal number of YES and NO examples, although the order is random. Also, during training, the query image is equally likely to be from either class. Appendix C.3 describes the input prompt generation procedure in detail. In the experiments shown in this paper we choose φ among five different background image sets ($\mathbb{I}_1$ to $\mathbb{I}_5$) and five foreground object category sets ($\mathbb{C}_{\text{easy}}$, $\mathbb{C}_{\text{shape}}$, $\mathbb{C}_{\text{tshape}}$, $\mathbb{C}_{\text{tool}}$, $\mathbb{C}_{\text{hard}}$), and we set the M in our task families to be either 1 or 3. Thus, we curate $5 \times 5 \times 2 = 50$ task families in SVAT. Each factor ($\mathbb{I}$, $\mathbb{C}$, and M) would influence the difficulty level of the task to be generated. For each task family in SVAT, we generate 1,000 training, 200 validation, and 1,000 testing examples. More details of each task family's characteristics can be found in Appendix D.

2.2 Curriculum Learning on SVAT

The different choices of φ form a set of task families in SVAT with varying levels of difficulty. We will show in Table 1 that state-of-the-art VLMs struggle to tackle complex task families, both in a zero-shot setting and after finetuning. However, progressively increasing task difficulty during CL finetuning increases VLM performance. This section formalizes CL on SVAT.

We define a task family $\mathbb{E}_\varphi$ which is a subset of all possible examples parameterized by φ, thus a curriculum $\mathcal{C}(\varphi) = (\mathbb{E}_{\varphi_1}, ..., \mathbb{E}_{\varphi_{|\mathcal{C}|}})$ is an ordered sequence of task parameterizations. Unless explicitly mentioned, we train VLMs in two stages when using CL, starting with an easier task family $\mathbb{E}_{\varphi_1}$, and then a harder task family $\mathbb{E}_{\varphi_2}$. We design four CL strategies corresponding to the three perspectives in φ that affect the task difficulty, namely $\mathcal{C}^{\mathbb{I}}$ for background complexity, $\mathcal{C}^{\mathbb{C}}$ for object category variety, $\mathcal{C}^{M}$ for the number of distracting objects, and $\mathcal{C}^{\text{all}}$ for all aspects where we start to train VLMs from the simplest task, thus for $\varphi_2 = (\mathbb{I}_i, \mathbb{C}_i, M_i, \mathbb{T}_i)$:

$$\begin{aligned} \mathcal{C}^{\mathbb{I}}(\varphi_2) &= (\mathbb{E}_{(\mathbb{I}_1, \mathbb{C}_i, M, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}), \mathcal{C}^{\mathbb{C}}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_i, \mathbb{C}_{\text{easy}}, M, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}) \\ \mathcal{C}^{M}(\varphi_2) &= (\mathbb{E}_{(\mathbb{I}_i, \mathbb{C}_i, 1, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}), \mathcal{C}^{\text{all}}(\varphi_2) = (\mathbb{E}_{(\mathbb{I}_1, \mathbb{C}_{\text{easy}}, 1, \mathbb{T}_t)}, \mathbb{E}_{\varphi_2}) \end{aligned} \tag{1}$$

3 Experiments

We evaluate the capacity of VLMs to learn in-context novel visuospatial concepts in our SVAT benchmark in this section. We report the performance of several current VLMs in zero-shot, finetuned, and curriculum learning (CL) settings. We leave the discussion and limitation of SVAT in Appendix G.

3.1 Experimental Setup

Backbone VLMs. We evaluate and finetune the following VLMs pretrained on different corpora: LLaVA-1.6-Mistral-7B [20], VILA-1.5-8B [18], Idefics2-8B [16], InternVL2-8B [8], and MiniCPM-V-2.6 [34] from Huggingface. All of these models, except LLaVA-1.6, were either pretrained on image-text-interleaved datasets (VILA and Idefics2), or are known to excel on existing multi-image benchmarks (InternVL2 and MiniCPM-V-2.6). We evaluate only the 7B (or 8B) parameter versions of each backbone for experiment efficiency and comparison fairness.

Task Selection. As SVAT consists of numerous task families with different selections of the parameterization φ, it would be infeasible if we enumerate every task selection throughout SVAT. Therefore, we only consider two main sets of task families ($\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{none}}$) and ($\mathbb{I}_5, \mathbb{C}, 3, \mathbb{T}_{\text{guide}}$) in Table 1, as the former one tests whether a VLM can do spatial reasoning without the help of texts, and the latter one investigates if a VLM can identify the target object with the help of the guiding texts. We also show the performance of VLMs on a simpler task ($\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{guide}}$) in Appendix F, where the question explicitly mentions the target object's category but there is no distractor in the image.

Training. We use ModelScope's *swift* library [38] to finetune VLMs on SVAT. We use LoRA [13] to finetune the VLMs, either on a single task or in stages via CL. When using CL, within each difficulty level $\mathbb{E}_{\varphi_i}$, we shuffle the order of training examples and use the finetuned LoRA parameters to initialize the training for the subsequent difficulty level. After the last and most difficult finetuning step,

Table 1: Main results of VLMs' performance on SVAT. M denotes the number of objects per example, and the second row on the header indicates the foreground object category set $\mathbb{C}$ in task family φ. The complexity of the background images is fixed at level 5 ($\mathbb{I}_5$). Accuracy significantly better than random guessing is in green, and each task's best model's result is in **bold**.

		$M=1, \mathbb{T}=\mathbb{T}_{\text{none}}$ (no distractors, useless text)					$M=3, \mathbb{T}=\mathbb{T}_{\text{guide}}$ (distractors, text names objects)				
Category	Model	easy	shape	tshape	tool	hard	easy	shape	tshape	tool	hard
Zero-shot	LLaVA-1.6-7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Idefics2-8B	50.4	49.6	49.8	50.7	52.3	51.1	53.7	**52.7**	49.2	49.7
	VILA-1.5-8B	49.3	48.9	49.9	47.6	47.7	49.8	52.4	51.8	**52.3**	48.7
	InternVL2-8B	46.8	49.9	48.2	47.7	46.1	50.2	**54.0**	49.3	49.8	50.1
	MiniCPM-V-2.6	**59.5**	**57.3**	**56.5**	**58.0**	**55.0**	**52.6**	51.9	51.1	50.8	**50.4**
Finetuned (FT)	LLaVA-1.6-7B	52.8	47.9	52.0	49.2	49.3	**80.3**	53.4	51.1	49.3	52.3
	Idefics2-8B	65.6	53.9	51.2	54.6	62.1	49.0	54.1	50.0	49.7	48.6
	VILA-1.5-8B	72.9	49.9	49.9	**77.3**	66.6	49.1	54.5	50.6	49.6	50.6
	InternVL2-8B	70.4	74.7	55.0	52.9	49.8	77.9	**76.9**	52.4	**65.6**	50.9
	MiniCPM-V-2.6	**73.4**	**80.0**	**68.6**	74.2	**71.8**	52.8	72.0	**58.4**	52.2	**62.1**

we merge LoRA parameters with the frozen VLM backbone for evaluation. Across all experiment setups, we finetune VLMs on each task family $\mathbb{E}_{\varphi_i}$ with three epochs unless explicitly mentioned. More details of our finetuning setup, including hyperparameters, can be found in Appendix B.

Evaluation. Because all SVAT tasks are simple yes/no binary tasks with 50-50 class balance, we simply report the exact-match accuracy for all tasks. Additionally we conduct one-sample z-tests on our results to see whether a VLM performs significantly better than random guessing. We set the significance level α as 0.05, so the threshold of any VLM performing significantly better than random guessing on each task's test set with 1,000 examples would be 52.7%.

3.2 Results

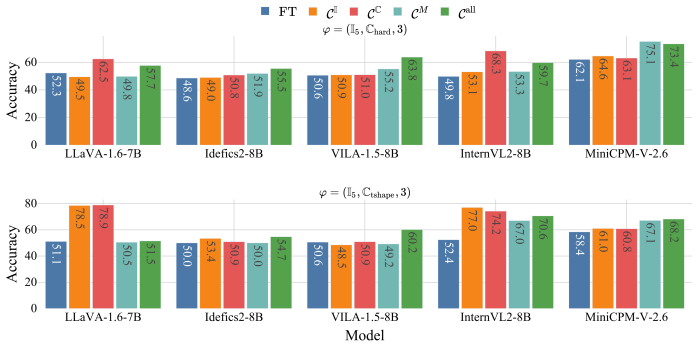


Figure 2: VLMs' performance on $(\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ and $(\mathbb{I}_5, \mathbb{C}_{\text{tshape}}, 3, \mathbb{T}_{\text{guide}})$ using CL and finetuning (FT).

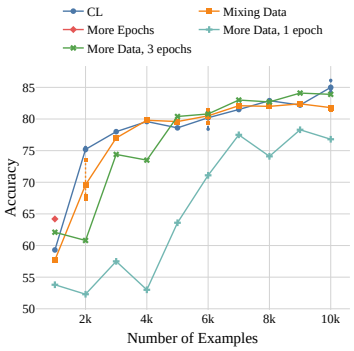


Figure 3: Ablation on MiniCPM for the task $(\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$.

We demonstrate the performance of VLMs in zero-shot and finetuned settings on SVAT in Table 1. In zero-shot settings, pretrained VLMs struggle at ambiguous spatial reasoning regardless of their pretraining and instruction-tuning recipes. Among the evaluated VLMs, MiniCPM performs the best across all tasks with an average accuracy of 54.3%. It is also the only VLM that consistently achieves significantly better than random guessing on $(M=1, \mathbb{T}=\mathbb{T}_{\text{none}})$ under zero-shot settings. In the meantime, some models perform better in $(M=3, \mathbb{T}=\mathbb{T}_{\text{guide}})$, e.g., Idefics2 on $\mathbb{C}_{\text{tshape}}$ and InternVL2 on $\mathbb{C}_{\text{shape}}$. The reason could be that textual prompts are clearer in $M=3, \mathbb{T}=\mathbb{T}_{\text{guide}}$ settings where the target object's category is explicitly mentioned in the prompt. We conclude that these models might be more sensitive to language rather than vision prompts at inference. Furthermore, most models can correctly follow the format of the ICL examples to answer with either "Yes" or "No", except for LLaVA-1.6. We conjecture that LLaVA-1.6 cannot follow multi-image ICL examples due to not being pretrained on image-text-interleaved datasets.

Finetuning VLMs directly on tasks in SVAT improves their performance (bottom section of Table 1), regardless of how the model was pretrained. MiniCPM still performs best with 66.6% accuracy on average after finetuning. Surprisingly, LLaVA-1.6 achieves extremely-high accuracy on $(\mathbb{I}_5, \mathbb{C}_{\text{easy}}, 3, \mathbb{T}_{\text{guide}})$ after finetuning, while performing poorly on $(\mathbb{I}_5, \mathbb{C}_{\text{easy}}, 1, \mathbb{T}_{\text{none}})$. We conjecture that mentioning the target object's category in the query is essential for LLaVA-1.6 to learn the objects' spatial relationship within the images. Similar to this phenomenon, we see that InternVL2 performs better on most $M = 3, \mathbb{T}_{\text{guide}}$ tasks than their $M = 1, \mathbb{T}_{\text{none}}$ counterparts. However, when the foreground object's vocabulary becomes larger (for $\mathbb{C}_{\text{tool}}$ and $\mathbb{C}_{\text{hard}}$), VLMs consistently get worse results on has-distractor settings, whereas only InternVL2 and MiniCPM achieve non-trivial performance on $\mathbb{C}_{\text{tool}}$ and $\mathbb{C}_{\text{hard}}$, respectively.

Although task families in SVAT with a larger object vocabulary, complex background, and some distractors are challenging for VLMs, Fig. 2 shows that applying CL to VLMs effectively improves model performance. Across different models with varied curriculum setups, 34 out of 40 (85%) trained models' performance increases compared to straightforward finetuning. Furthermore, all models can achieve significantly better accuracy after CL than random guessing ($> 52.7\%$). We also notice that different VLMs benefit most from different CL strategies. For example, $\mathcal{C}^{\mathbb{C}}$ increases the performance of LLaVA-1.6 and InternVL the most, whereas MiniCPM barely gains improvements. Our further analysis shows that succeeding in the first task $\mathbb{E}_{\varphi_1}$ after the first-stage finetuning is a necessary factor for the performance improvements using CL. The improvement of the final model can also be reflected after the first training stage. We leave the analysis details in Appendix I.

3.3 Ablation Study

As the improvements in VLM performance can be due not (or not only) to CL but to a greater diversity of data during finetuning or a larger number of training steps, we examine these possibilities through ablations. We only apply ablations to MiniCPM for simplicity, as it is one of the most efficient VLM for training and inference among the models listed above. We run ablations on the task family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ with curriculum $\mathcal{C}^M$, where the model gains the most performance through CL. We conduct the ablation study based on the following three strategies:

1) Mixing Data. We naively combine and randomly shuffle the data from all the datasets in CL.

2) More Epochs. Simply training the VLM with six epochs to match the total training steps in CL.

2) More Data. As SVAT is a synthetic dataset, we generate more training data for the task family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$. We finetune the VLM with one epoch to ensure there are no repeated examples in training to eliminate overfitting and also three epochs to match CL.

Fig. 3 indicates simply training VLMs with more steps on the same examples (#examples = 1,000) does not improve the model performance. However, increasing the quantity of novel training data might help, yet the performance cannot match CL unless all examples are unique in training (#examples = 6,000), mainly because more combinations of the target object's spatial information in demonstration and query examples are presented to the model. Meanwhile, mixing the data from easy and complex tasks can help, yet the trained VLMs' performance is slightly worse than CL. Therefore, we conclude that CL is essential for empowering VLMs on ambiguous spatial reasoning, especially in data-limited scenarios where the target complex task's training example quantity is low.

4 Conclusion

We introduce a benchmark of ambiguous visual-spatial reasoning tasks, namely Spatial Visual Ambiguity Tasks (SVAT), and use it to evaluate a set of current VLMs on their ability to learn novel visuospatial concepts via in-context learning. We find that current VLMs cannot solve these tasks exclusively in context without specific training, and some still fail to learn by finetuning the tasks directly. However, they can learn the more difficult tasks from in-context visual demonstrations if they have previously been finetuned on easier tasks through a curriculum learning approach. Our analysis shows that curriculum learning presents a data-efficient and more robust way of training VLMs on SVAT. These results demonstrate more evidence of the power of curriculum learning to adapt large models. Despite the performance gained from curriculum learning, state-of-the-art VLMs require further development to reliably solve ambiguous tasks with vision prompts or demonstrations. We hope our work will facilitate future research in this direction.

References

[1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] F. B. Baldassini, M. Shukor, M. Cord, L. Soulier, and B. Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1539–1550, 2024.

[3] P. Banerjee, T. Gokhale, Y. Yang, and C. Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1908–1918, 2021.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM, 2009. doi: 10.1145/1553374.1553380.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Proc. of NeurIPS*, 2020.

[6] D. Campos. Curriculum learning for language modeling. *ArXiv preprint*, abs/2108.02170, 2021.

[7] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024.

[8] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *ArXiv preprint*, abs/2404.16821, 2024.

[9] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *ArXiv preprint*, abs/2406.01584, 2024.

[10] D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.247.

[11] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proc. of NeurIPS*, volume 35, pages 30583–30598. Curran Associates, Inc., 2022.

[12] J. Guo, X. Tan, L. Xu, T. Qin, E. Chen, and T.-Y. Liu. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7839–7846, Apr. 2020. doi: 10.1609/aaai.v34i05.6289.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proc. of ICLR*. OpenReview.net, 2022.

[14] A. Kamath, J. Hessel, and K.-W. Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In H. Bouamor, J. Pino, and K. Bali, editors, *Proc. of EMNLP*, pages 9161–9175, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.568.

[15] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon. Building and better understanding vision-language models: insights and future directions. *ArXiv preprint*, abs/2408.12637, 2024.

[16] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models?, 2024.

[17] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[18] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699, 2024.

[19] F. Liu, G. Emerson, and N. Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

[20] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[21] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proc. of EMNLP*, pages 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759.

[22] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind. *ArXiv preprint*, abs/2407.06581, 2024.

[23] K. Ranasinghe, S. N. Shukla, O. Poursaeed, M. S. Ryoo, and T.-Y. Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987, 2024.

[24] T. Srinivasan, X. Ren, and J. Thomason. Curriculum learning for data-efficient vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5619–5624, 2023.

[25] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, 2024.

[26] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, and N. Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *ArXiv preprint*, abs/2406.14852, 2024.

[27] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[28] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, et al. Larger language models do in-context learning differently. *ArXiv preprint*, abs/2303.03846, 2023.

[29] W. Wu, S. Mao, Y. Zhang, Y. Xia, L. Dong, L. Cui, and F. Wei. Visualization-of-thought elicits spatial reasoning in large language models. *ArXiv preprint*, abs/2404.03622, 2024.

[30] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. In *Proc. of ICLR*, 2022.

[31] B. Xu, L. Zhang, Z. Mao, Q. Wang, H. Xie, and Y. Zhang. Curriculum learning for natural language understanding. In *Proc. of ACL*, pages 6095–6104, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.542.

[32] C. Yang, R. Xu, Y. Guo, P. Huang, Y. Chen, W. Ding, Z. Wang, and H. Zhou. Improving vision-and-language reasoning via spatial relations modeling. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 758–767, 2024. doi: 10.1109/WACV57701.2024.00082.

[33] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12527–12536, 2021.

[34] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *ArXiv preprint*, abs/2408.01800, 2024.

[35] J. Zhang, z. wei, J. Fan, and J. Peng. Curriculum learning for vision-and-language navigation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Proc. of NeurIPS*, volume 34, pages 13328–13339. Curran Associates, Inc., 2021.

[36] Y. Zhang, K. Zhou, and Z. Liu. What makes good examples for visual in-context learning? In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Proc. of NeurIPS*, volume 36, pages 17773–17794. Curran Associates, Inc., 2023.

[37] H. Zhao, Z. Cai, S. Si, X. Ma, K. An, L. Chen, Z. Liu, S. Wang, W. Han, and B. Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *Proc. of ICLR*, 2024.

[38] Y. Zhao, J. Huang, J. Hu, X. Wang, Y. Mao, D. Zhang, Z. Jiang, Z. Wu, B. Ai, A. Wang, W. Zhou, and Y. Chen. Swift:a scalable lightweight infrastructure for fine-tuning. *ArXiv preprint*, abs/2408.05517, 2024.

[39] Y. Zhou, X. Li, Q. Wang, and J. Shen. Visual in-context learning for large vision-language models. *ArXiv preprint*, abs/2402.11574, 2024.

[40] Y. Zong, O. Bohdal, and T. Hospedales. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *ArXiv preprint*, abs/2403.13164, 2024.

A Related Works

A.1 VLMs for Spatial Reasoning

Despite the rapid development of vision language models (VLMs) [8, 18, 20, 34], today's best VLMs are very limited in their ability to solve seemingly simple spatial reasoning tasks [22, 25]. Challenging benchmarks have been proposed to examine and improve VLMs' performance in spatial reasoning, including but not limited to spatial relationship detection [7, 9, 14, 19], object localization [23], navigation [29], distance measuring [9], etc. However, such tasks are delicately defined with engineered prompts so that VLMs can understand the question, yet SVAT focuses on tasks that are ambiguous in words but can be properly defined by visual demonstrations. Moreover, existing methods in tackling spatial reasoning tasks with VLMs often rely on prompt engineering [26, 29] and explicit spatial modeling [3, 32], but our analysis with SVAT finds that curriculum learning can be a more efficient way of enabling VLMs' ability in spatial reasoning.

A.2 In-Context Learning

In-context Learning (ICL) was first introduced in Brown et al. [5], which found that pretrained large language models can be adapted to novel tasks given several demonstration examples at inference time, rather than using them to update the model's parameters. Research found that such ICL learning process can be seen as linear regression [11], Bayesian models [30], gradient descent [10], etc. However, research also pointed out that ICL might only help language models shift to a new input and output distribution rather than deeper reasoning capabilities [21, 28].

Besides language models, recent advancements in multi-image multimodal learning addressed the fact that VLMs can also learn novel tasks through ICL [18]. Many benchmarks have been developed to specifically examine existing VLMs' ICL capabilities with multi-image inputs [37, 40], while

recent research also stressed that ICL can be adapted to VLMs to tackle visual-related reasoning tasks [39]. Nonetheless, such performance improvements highly rely on the text modality of the task rather than the image modality [2], leaving the ICL's effect in vision-oriented tasks under-explored.

A.3 Curriculum Learning

Curriculum learning (CL) was first proposed in Bengio et al. [4], suggesting training machine learning models from easier to harder task examples could achieve better model convergence and robustness. In the area of language modeling, CL also shows its effectiveness in both pertaining [6] and finetuning [12, 31] stages. In multimodal learning, especially with VLMs, research also demonstrated that CL can improve model performance in navigation [35] and vision-language alignment [24].

B Experiment Details

We set the learning rate as 1e-4 for finetuning, while LoRA r and alpha are set as 8 and 32, respectively. The full hyperparameters we use for all VLMs finetuning are shown in Table 2.

Table 2: Hyperparameters used in finetuning VLMs on SVAT

Hyperparameter	Value
Learning rate	1e-4
Batch size	16
#Epochs	3
Warmup ratio	0.05
Weight decay	0.1
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.95
Adam ϵ	1e-8
Gradient clipping	1.0
LoRA r	8
LoRA α	32
LoRA dropout	0.1

When training and evaluating with Idefics2 models, we set `do_image_splitting` to True to reach the full potential of the model's capabilities. For the InternVL2 model, we leave the input image size as `(448, 448)` by default and set the maximum number of crops generated from its processor to 12. For MiniCPM-V-2.6, we set the `max_slice_nums` to None in both training and evaluation stages. At inference, as tasks in SVAT are all binary question answering problems that expect the VLM to respond with either "Yes" or "No", we set the `max_new_tokens` as 5. We do not set it to 1 because we want to see whether a pretrained VLM can directly follow the ICL demonstrations' output format in zero-shot settings (details introduced in Appendix F). Meanwhile, we do not use sampling or beam search at inference time.

C SVAT Data Generation Details

C.1 Background: Visual In-context Learning

We first formulate the problem of visual ICL. Formally, under the vision question-answering (VQA) setting, we define an input prompt x that consists of a set of in-context examples together with a new question and image:

$$x = (E, t^q, v^q), \text{ where } E = \{e_i | e_i = (t_i^d, v_i^d, y_i^d) \in \mathbb{E}, t_i^d \in \mathbb{T}, v_i^d \in \mathbb{V}\}_{i=1}^N, t^q \in \mathbb{T}, v^q \in \mathbb{V} \quad (2)$$

where $\mathbb{T}$ is a finite set of textual questions the VLM should answer, $\mathbb{V}$ is the set consisting of all possible images given a specific task, and y_i^d is the ground-truth label for the image-question pair $y_i^d = l(t_i^d, v_i^d)$. Thus, a VLM is expected to tackle the task that $y^q = \text{VLM}(E, t^q, v^q) = l(v^q, t^q)$.

C.2 SVAT Problem Formulation

SVAT fits into the visual ICL formulation in Eq. (2) by specifying a generation process for the text query t, image v and label y. As shown in Fig. 1, each demonstration or query image $v = (i, o_1, ..., o_M)$ consists of a background image i and M foreground objects $o_1, ..., o_M$, while all examples share the same text question t. Moreover, each foreground object $o_j = (c_j, \xi_j)$ is defined by c, the category of the object, and a low-dimensional vector ξ that defines the pose of the object (position, size, orientation, etc.). Within each input x and sharing across examples E and e^q, we have a decision function P that maps an orientation of an object ξ to a label: $P : \xi \to \{0, 1\}$. When there is more than one object, i.e., $M > 1$, only the first object, o_1 is needed to find the example's label, while the rest are left as visual distractors. Overall, an SVAT dataset $\mathcal{D}_{\text{SVAT}}$ is defined as:

$$\mathcal{D}_{\text{SVAT}} = \{(x, P, y^q) | x = (E, t, v^q), y^q = l_{\text{SVAT}}(v^q, t, P) = P(\xi_1)\} \tag{3}$$

where P is the decision boundary that must be inferred by the VLM from ICL examples. Note that no visual or textual clues in the image and question show P. The same question t and decision boundary P are shared across demonstration examples and the query example.

C.3 Input Prompt Generation in SVAT

We demonstrate an example of full prompt in SVAT under the task family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ in Table 3. For a detailed algorithmic description of the input prompt and image generation process, see Algorithm 1. We also demonstrate some of the sampled data from different task families in SVAT on Fig. 4.

```
Please answer the following question based on the provided examples.

Example 1:
<image>
Question:  Is the Heat Guns in the right position?
Answer:  Yes

Example 2:
<image>
Question:  Is the Heat Guns in the right position?
Answer:  No

Example 3:
<image>
Question:  Is the Heat Guns in the right position?
Answer:  No

Example 4:
<image>
Question:  Is the Heat Guns in the right position?
Answer:  Yes

Query:
<image>
Question:  Is the Heat Guns in the right position?
Answer:
```

Table 3: Sampled prompt from the task family $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ in SVAT.

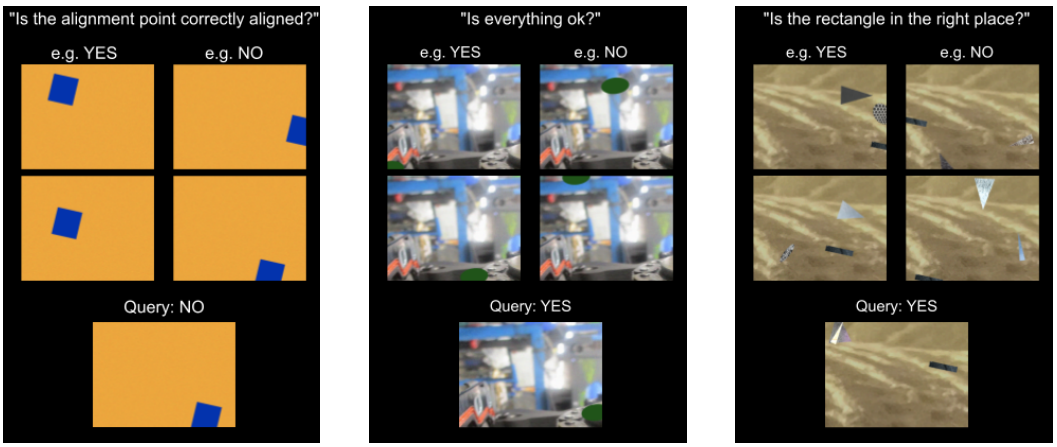
D Dataset Characteristics Details

As described in Section 2.1, the images in SVAT are synthesized based on different sets of background images and foreground objects which makes the difficulty of each task family $\varphi = (\mathbb{I}, \mathbb{C}, M, \mathbb{T})$ controllable.

In detail, we have five different complexity level defined for the background images, ranging from $\mathbb{I}_1$ to $\mathbb{I}_5$, including

Algorithm 1 Input Prompt Generation Algorithm

Input: $\varphi = (\mathbb{I}, \mathbb{C}, M, \mathbb{T})$
Input: N ▷ Number of examples, including query. We use $N = 5$
Input: ε ▷ Difficulty threshold. We use $\varepsilon = 0.05$
$D \leftarrow \dim(\xi)$ ▷ Dimensionality of the pose vector. We use $D = 2$
$\delta \sim \text{Uniform}\{1, \ldots, D\}$ ▷ Pick a dimension for decision boundary
$\tau \sim \text{Uniform}[2\varepsilon, 1 - 2\varepsilon]$ ▷ Threshold for decision boundary
$s \sim \text{Uniform}\{-1, 1\}$ ▷ Direction of decision boundary
$P(\xi) := I[s(\xi_\delta - \tau) > 0]$ ▷ Define decision boundary function
$i \sim \mathbb{I}$ ▷ Shared background image for all examples
$t^q \sim \mathbb{T}$ ▷ Sample text query (may be $\mathbb{T}_{none}$ or $\mathbb{T}_{guide}$)
$c^* \sim \mathbb{C}$ ▷ Sample a target object class
$Y_{\text{init}} \leftarrow [0, 1] \times \lfloor N/2 \rfloor$ ▷ Initialize balanced labels
$Y_{\text{query}} \sim \text{Uniform}\{0, 1\}$ ▷ Sample final label
$Y \leftarrow \text{Shuffle}(Y_{\text{init}}) \cup Y_{\text{query}}$ ▷ Shuffle ICL examples
$E \leftarrow []$ ▷ Initialize list of examples
for $j = 1$ to N **do** ▷ Create N examples
  $y \leftarrow Y[j]$ ▷ Use pre-generated label
  $O \leftarrow \emptyset$ ▷ Initialize set of objects for this example
  for $k = 1$ to M **do** ▷ Create M objects per example
    repeat
      $\xi \sim \text{Uniform}[0, 1]^D$ ▷ Sample pose
    until $P(\xi) = y$ **and** $|\xi_\delta - \tau| > \varepsilon$ ▷ Check label and difficulty
    if $k = 1$ **then**
      $c \leftarrow c^*$ ▷ Use target class for first object
    else
      repeat
        $c \sim \mathbb{C}$ ▷ Sample class for distractor objects
      until $c \neq c^*$
    end if
    $O \leftarrow O \cup \{(c, \xi)\}$ ▷ Add object to example
  end for
  $V \leftarrow (i, O)$ ▷ Build the image with a background and foreground objects
  if $j = N$ **then**
    $v^q \leftarrow V$ ▷ Assign the query example's image
  else
    $E \leftarrow E \cup [(t^q, V, y)]$ ▷ Add example to the demonstration list
  end if
end for
$x \leftarrow (E, t^q, v^q)$
return (x, P, Y_{query}) ▷ Finish constructing an instance in $\mathcal{D}_{\text{SVAT}}$



(a) Trivial background **(b)** Complex background **(c)** Distracting objects

Figure 4: Examples of SVAT tasks where the object of interest is a simple shape. In (a) the colors and textures are trivial with $\varphi = (\mathbb{I}_2, \mathbb{C}_{\text{shape}}, 1, \mathbb{T}_{\text{none}})$, while in (b) there is more visual complexity with $\varphi = (\mathbb{I}_5, \mathbb{C}_{\text{shape}}, 1, \mathbb{T}_{\text{none}})$. In (c) there are distractor shapes, and the model must identify the object of interest using the text of the query, with $\varphi = (\mathbb{I}_3, \mathbb{C}_{\text{tshape}}, 3, \mathbb{T}_{\text{guide}})$

- $\mathbb{I}_1$: empty (solid white) background;
- $\mathbb{I}_2$: solid background but with varied RGB colors randomly sampled from `(0, 0, 0)` to `(255, 255, 255)`;
- $\mathbb{I}_3$: simple, realistic textured images, like grass field, snow, wood, sheet, etc.;
- $\mathbb{I}_4$: simple photographs taken consisting of few objects, e.g., a desk, ceiling, wall, etc.;
- $\mathbb{I}_5$: complex images that contain multiple realistic objects from industrial scenes.

As for the foreground objects, we have defined the following sets:

- $\mathbb{C}_{\text{easy}}$: contains five objects, including a bolt, a chain, a hardhat, a pickup truck, and a tree;
- $\mathbb{C}_{\text{shape}}$: consisting of five naive shapes, namely circle, pentagon, rectangle, square, and triangle. Each object is filled with a solid RGB color randomly sampled from `(0, 0, 0)` to `(255, 255, 255)`;
- $\mathbb{C}_{\text{tshape}}$: same shapes in $\mathbb{C}_{\text{shape}}$, but filled with random textures from $\mathbb{I}_3$;
- $\mathbb{C}_{\text{tool}}$: a set of 87 tools commonly seen in industrial scenes, like hammer, saw, carpet knife, drill, heat gun, etc., where each category of tool has only one image;
- $\mathbb{C}_{\text{hard}}$: 3,437 industrial tool images from 328 categories in total.

Finally, $\mathbb{T}$ in φ controls the construction or sampling process of questions t in SVAT datasets based on the formulation in Eq. (2). t is built based on the following templates shown in Table 4. The `{fiducial}` in the template is randomly replaced with a set of synonyms (including the word "fiducial") if $\mathbb{T} = \mathbb{T}_{\text{none}}$, like "marker", "landmark", "beacon", etc. When $\mathbb{T} = \mathbb{T}_{\text{guide}}$, `{fiducial}` is replaced with the target object's category name c_1. The variable `{description}` in the template is randomly replaced with a set of adjectives and phrases representing the status of "Yes" or "No", like "aligned", "in position", "out of place", etc. Since the same question t is consistent within each input x's demonstration examples and query example, and all questions in SVAT task families are binary, the actual choice of the variable `{description}` here does not affect the ground truth label of the query example, as long as the query example's decision boundary is consistent with the demonstration examples.

E Additional Results: More Backbone Models

We introduce the experimental results of the following models here: Phi-3.5 [1], LLaVA-OneVision [17], and Qwen2-VL [27]. We do not add these models' results to the main table

Table 4: Question templates in SVAT

Templates
`Is the {fiducial} {description}?`
`Are the {fiducial} {description}?`
`Are the {fiducial} {description}?`
`Can you see if the {fiducial} is {description}?`
`Is there a problem with the {fiducial}?`
`Look at the {fiducial}.  Is it {description}?`
`Find the {fiducial}.  Is it {description}?`
`Can you see the {fiducial}?  Is it {description}?`
`Is the {fiducial} properly positioned?`
`Is the {fiducial} correctly aligned?`
`Is the {fiducial} in the correct position?`
`Can you see if the {fiducial} is in the correct position?`
`Is the {fiducial} in the right place?`
`Find the {fiducial}.  Is it in the right place?`
`Can you see the {fiducial}?  Is it in the right place?`
`Is the {fiducial} in the right position?`

Table 5: Zero-shot and finetuned VLMs' performance on $\varphi = (\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{none}})$ and $\varphi = (\mathbb{I}_5, \mathbb{C}, 3, \mathbb{T}_{\text{guide}})$ for additional models. Accuracy significantly better than random guessing is in green.

		$M=1, \mathbb{T}=\mathbb{T}_{\text{none}}$ no distractors, useless text					$M=3, \mathbb{T}=\mathbb{T}_{\text{guide}}$ distractors, text names objects				
Category	Model	easy	shape	tshape	tool	hard	easy	shape	tshape	tool	hard
Zero-shot	Qwen2-VL-2B	48.4	50.2	50.1	51.0	48.0	49.3	52.7	51.3	49.9	50.5
	Qwen2-VL-7B	56.3	52.7	53.8	57.1	55.6	53.4	52.9	52.6	50.8	54.0
	LLaVA-OneVision	49.2	50.5	47.9	48.0	48.9	50.7	53.3	46.8	49.0	49.2
	Phi-3.5	49.9	50.1	50.2	48.2	48.7	51.7	50.9	50.3	50.9	47.7
Finetuned (FT)	Qwen2-VL-2B	50.6	52.2	51.6	48.6	55.1	73.2	63.9	52.3	64.0	51.8
	Qwen2-VL-7B	74.6	61.4	63.0	72.7	71.4	74.4	75.2	60.3	71.0	55.0
	LLaVA-OneVision	75.3	52.3	51.2	60.5	63.9	73.9	75.5	56.7	72.5	55.1
	Phi-3.5	67.0	72.2	63.3	66.4	60.6	50.6	49.8	54.3	49.7	54.0

since some of them are recently released, while the scales of Phi-3.5 and Qwen2-VL-2B models are smaller than the 7-8B scaled models we show in Table 1. Results show that all models perform better after finetuning compared to zero-shot inference. Moreover, Qwen2-VL-7B models achieve better performance than the rest of the models. In the meantime, we notice that Qwen2-VL-2B series models do not get significant performance improvement after finetuning on the $\varphi = (\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{none}})$ task. We assume that smaller-scaled models cannot capture visual features without the guidance of textual prompts.

F Additional Results: Guided Texts without Distractors

We demonstrate VLMs' performance on the task family $\varphi = (\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{guide}})$ under zero-shot and finetuned settings in Table 6. Despite this task should be empirically simpler than the ones shown in Table 1, we still find that VLMs struggle at tackling it under zero-shot settings, where only the MiniCPM model shows its performance significantly better than random guessing on $\mathbb{C}_{\text{easy}}$, $\mathbb{C}_{\text{shape}}$, $\mathbb{C}_{\text{tool}}$, and $\mathbb{C}_{\text{hard}}$. Besides, VILA's performance is much worse than random guessing because it does not follow the output format given in the ICL demonstrations, i.e., answering with either "Yes" or "No". 2,907 out of 5,000 answers (58.1%) from VILA fail to follow the ICL output format.

At the same time, after finetuning, we see that in 18 out of 25 (72%) settings, the model performs significantly better than random guessing. MiniCPM, again, performs the best across most of the settings except for $\mathbb{C}_{\text{tshape}}$. The averaged accuracy across all models on all foreground object selection $\mathbb{C}$ after finetuning achieves 64.5, which is better than that of $(\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{none}})$ (61.0) and

Table 6: Zero-shot and finetuned VLMs' performance on $\varphi = (\mathbb{I}_5, \mathbb{C}, 1, \mathbb{T}_{\text{guide}})$. Accuracy significantly better than random guessing is in green, and each task's best model's result is in **bold**.

		M = 1, $\mathbb{T} = \mathbb{T}_{\text{guide}}$ (no distractors, text names objects)				
Category	Model	easy	shape	tshape	tool	hard
Zero-shot	LLaVA-1.6-7B	0.0	0.0	0.0	0.0	0.0
	VILA-1.5-8B	14.7	28.4	26.3	14.2	20.2
	Idefics2-8B	46.5	49.6	51.8	50.3	48.4
	InternVL2-8B	48.8	51.0	50.3	49.2	49.4
	MiniCPM-V-2.6	**55.4**	**56.7**	**52.3**	**56.0**	**53.7**
Finetuned	LLaVA-1.6-7B	46.8	76.8	50.9	50.7	50.2
	VILA-1.5-8B	70.5	53.5	52.5	68.3	67.5
	Idefics2-8B	61.3	61.8	51.0	49.9	57.3
	InternVL2-8B	79.3	77.8	**74.8**	72.3	55.0
	MiniCPM-V-2.6	**81.7**	**81.1**	73.1	**77.5**	**70.8**

$(\mathbb{I}_5, \mathbb{C}, 3, \mathbb{T}_{\text{guide}})$ (56.5) shown in Table 1, indicating that prompting with VLMs with objects' category names make the task easier even if there is no distractor in the image.

G Discussion and Limitations

Even though our experiment demonstrated in Section 3 has covered various settings in SVAT, we cannot enumerate every possible combination regarding the task parameterization φ to examine VLMs' ambiguous spatial reasoning abilities in extreme details. However, our code can easily be adapted to include tasks with more combinations or choices of φ. Furthermore, the core components in SVAT, namely the decision boundary P, the background image i, foreground objects o, and even natural language questions t can be easily extended based on our released code. We would like to leave the exploration of applying more challenging tasks on VLMs as future work.

Besides extending task variety and difficulty, this paper only examines VLMs with a scale of 7B to 8B due to the limitation of our computational capabilities. In theory, larger models have more potential for conducting visuospatial reasoning, especially under in-context learning setups. Nonetheless, we argue that models can already perform relatively well by applying curriculum learning to VLMs at the 7-8B parameter scale, with the accuracy reaching about 75%. Therefore, a larger parameter scale might not be necessary for VLMs to do ambiguous spatial reasoning with decent accuracy.

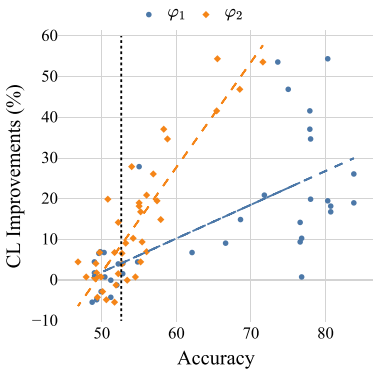


Figure 5: Correlations between the accuracy improvements after CL to the VLMs' performance on $\mathbb{E}_{\varphi_1}$ and $\mathbb{E}_{\varphi_2}$ after training with $\mathbb{E}_{\varphi_1}$.

Finally, as foundation models get more powerful, they are increasingly good at solving real-world tasks zero-shot, without any specific training. However, for many real-world tasks, the specific nature of the goal is somewhat ambiguous, and humans struggle to clearly articulate the exact criteria necessary to define a desired outcome. Oftentimes, it is easier for a person to give examples showing "this is good" and "this is bad" than to explicitly list the exact characteristics of each example that make one good or bad. SVAT only considers ambiguous spatial reasoning tasks with synthetic data, yet no realistic data for training or evaluation is considered. We want to leave the research of combining SVAT and ambiguous, realistic multimodal data as a future direction. What knowledge can be transferred between synthetic datasets like SVAT and real-world datasets and benchmarks for VLMs remains under-explored.

H Impact Statement

Since SVAT proposes a novel paradigm to prompt VLMs with visual demonstrations in a tuning-free manner, it can lead to more efficient leverage of large VLMs. Specifically, this can lead to positive environmental impacts, resource-saving, and democratization of the usage of VLMs. In the meantime, SVAT does not introduce new ethical concerns. However, the use of SVAT VLMs can inherit existing issues in pretrained VLMs, including but not limited to bias in decision-making, unfair response, etc. Misusing harmful vision demonstrations at inference time on SVAT-finetuned models could also result in unsafe outputs of existing VLMs.

I Details of Improvement Analysis in Curriculum Learning

In Fig. 5, we show that the VLM's performance on both $\mathbb{E}_{\varphi_1}$ and $\mathbb{E}_{\varphi_2}$[2] after the first-stage finetuning substantially affects the model's final performance on $\mathbb{E}_{\varphi_2}$ after CL. We see that all models that do not achieve significantly better accuracy on $\mathbb{E}_{\varphi_1}$ cannot improve their performance through CL. Meanwhile, a positive correlation with $R^2 = 0.77$ exists between the VLMs' $\mathbb{E}_{\varphi_2}$ performance after the first and second-stage training. We conclude that learning to tackle $\mathbb{E}_{\varphi_1}$ with spatial reasoning capabilities is necessary for succeeding on $\mathbb{E}_{\varphi_2}$ throughout CL, while the final VLM's performance of CL is predictable based on the intermediate models' performance.

J VLM Training and Evaluation Efficiency on SVAT

Table 7: Training time and peak memory consumption on the task $(\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ in SVAT.

Model	Training Time (s)	Training Memory (GiB)
LLaVA-Next	7,602.1	43.8
Idefics2	3,113.9	53.1
VILA	1,379.4	46.3
InternVL2	12,208.4	75.4
MiniCPM-V-2.6	1,138.7	34.3
Qwen2-VL	1,448.3	49.0

We use ModelScope's *swift* library to finetune and evaluate the following models: LLaVA-Next, InternVL2, and MiniCPM-V-2.6. We implement the training and evaluation pipeline of Idefics2 and VILA by ourselves as *swift* lacks the support for these models. We run our training with two sets of setups, either with 2 NVIDIA A100 80GB GPUs or 4 NVIDIA RTX 3090 GPUs. As for inference, VILA, MiniCPM, and LLaVA-Next can be fit on a single RTX 3090 or RTX 4090 GPU. Idefics2 models require more memory at inference time, so we evaluate them on a single A100 80GB GPU. For InternVL2, we use four RTX 3090 GPUs for inference. The detailed training time and memory consumption for Table 1 is demonstrated in Table 7. Since different task families in SVAT share similar text length, image quantity, and resolutions, we only report the time and memory consumption for the task family $(\mathbb{I}_5, \mathbb{C}_{\text{hard}}, 3, \mathbb{T}_{\text{guide}})$ in Table 7 for simplicity.

[2]We use $\mathbb{E}_{\varphi_1}$ and $\mathbb{E}_{\varphi_2}$ to represent the two tasks used in CL as defined in Eq. (1), i.e., $\mathcal{C}(\varphi_2) = (\varphi_1, \varphi_2)$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We conclude that state-of-the-art VLMs perform poorly on ambiguous spatial reasoning tasks, and curriculum learning can help VLMs achieve success on such tasks. Our main experimental results and ablation studies support the conclusion.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations and potential future directions of our work in Appendix G.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We do not have theoretical results in our paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We will release our code to construct the SVAT dataset, together with VLM training and evaluation pipelines. The detailed hyperparameters we use are also included in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release our code and data after the double-blind review stage.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The hyperparameters details are mentioned in Table 2, and the details of dataset construction and split are mentioned in Section 2.1, Appendix C, and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have done statistical significance tests for the main results, as mentioned in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computational resources of training and evaluating VLMs on SVAT are mentioned in Appendix J, together with the training time and memory consumption recorded in Table 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms with the NuerIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact in Appendix H, introducing that SVAT proposes a lightweight paradigm of using VLMs and does not bring new ethical conerns, yet it might inherit existing negative concerns from VLMs.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks, as the data to be released are synthesized with safe abstract or realistic images.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original assets (tools like ms-swift and pretrained VLMs) are properly cited in this paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets (code and data) introduced in this paper are well-documented and will be released after the double-blind review period.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.