
Escaping the SpuriVerse: Can Large Vision-Language Models Generalize Beyond Seen Spurious Correlations?

Yiwei Yang^{1,*}, Chung Peng Lee^{4,*}, Shangbin Feng¹, Dora Zhao², Bingbing Wen¹,
Anthony Z. Liu³, Yulia Tsvetkov¹, Bill Howe¹

¹University of Washington ²Stanford University ³University of Michigan

⁴Princeton University

yanyiwei@uw.edu, cl6486@princeton.edu, shangbin@cs.washington.edu,
dorothy@stanford.edu, bingbw@uw.edu, anthliu@umich.edu,
yuliats@cs.washington.edu, billhowe@uw.edu

Abstract

Spurious correlations occur when models rely on non-essential features that coincidentally co-vary with target labels, leading to incorrect reasoning under distribution shift. We consider spurious correlations in Large Vision Language Models (LVLMs) pretrained on extensive and diverse datasets without explicit task supervision. We develop a benchmark by sourcing GPT-4o errors on real-world visual-question-answering (VQA) benchmarks, then curating a subset through LVLM-human annotation and synthetic counterfactual evaluation to identify errors caused by spurious correlations. This process yields SpuriVerse, a novel benchmark comprised of 124 distinct types of spurious correlations extracted from real-world datasets, each containing 1 realistic and 10 synthetic VQA samples for a total of 1364 multiple choice questions. We evaluate 15 open and closed-source LVLMs on SpuriVerse, finding that even state-of-the-art closed-source models struggle significantly, achieving at best only 35.0% accuracy. Fine-tuning on synthetic examples that emphasize the spurious correlation improves performance to 78.4%, suggesting that training on diverse spurious patterns generalizes to unseen situations: models appear to learn to avoid "shortcuts" and attend to the overall image context.

1 Introduction

Real-world data frequently contains spurious correlations — patterns predictive of the target during training but irrelevant to the true label [Leino et al., 2018]. For example, in ImageNet, butterflies often appear with flowers [Singla and Feizi, 2021], while the Waterbirds dataset was constructed to have waterbird images commonly appear on water backgrounds [Sagawa et al., 2019]. Models trained on such data may learn to rely on these *spurious features* (backgrounds, co-occurring objects, texture, etc.), leading to errors when the correlations no longer hold. An example of such errors is a landbird incorrectly classified as a waterbird simply due to its water background [Sagawa et al., 2019]. This phenomenon is especially problematic in critical domains such as medicine, where models trained to detect pneumonia might rely on unrelated features (e.g., metal tokens) rather than genuine disease indicators [Zech et al., 2018].

*Equal contribution.

In traditional supervised training where a model is finetuned for a downstream application, spurious correlation arises from the association between spurious features and *target labels* captured by the model through standard loss minimization [Sagawa et al., 2019, Kirichenko et al., 2022]. In the regime of pre-training Large Vision-Language Models (LVLMs), not only are the training data more diverse and extensive, but the objective is no longer predicting a task-specific target label. In this generalized regime, we hypothesize that correlations between spurious features and target concepts can still persist and compromise LVLMs' performance. However, modern LVLMs' evaluation suites do not examine whether zero-shot LVLMs rely on spurious features to make incorrect predictions, where fine-tuning for a particular target label is not available. In this work, we **provide a benchmark to evaluate an LVLM's susceptibility to spurious correlations in generalized settings**. In particular, we are interested in spurious correlations that appear outside of contrived tasks where the correlation is a result of non-representative training data such as Gender-to-BlondHair in CelebA [Liu et al., 2015] and Background-to-BirdType in WaterBirds [Sagawa et al., 2019]. Instead of contrived spurious correlations, we focus on situations where a model may overrely on a *dominant* correlation that is not necessarily irrelevant but is not generally reliable (Figure 1).

We address two important challenges: (1) Existing benchmarks often study spurious correlations by collecting a training set with an imbalanced distribution across spurious features and target labels [Sagawa et al., 2019, Lynch et al., 2023, Joshi et al., 2023, Li et al., 2023b, Arjovsky et al., 2019]. These benchmarks are particularly useful for investigating whether a model learns spurious correlations when trained on a skewed distribution. However, when evaluating pre-trained LVLMs, curating an imbalanced dataset is ineffective due to distribution shifts. (2) The presence of spurious correlations often relies on annotations of *spurious features* and *target label*. For example, in CelebA [Liu et al., 2015], the comprehensive list of annotated face attributes allows one to recognize a strong correlation between *gender* and *hair color* in the training set. Observing that the trained model achieves low accuracy on non-blond males then confirms that the spurious correlation is captured by the model. However, this approach only applies to a specific target label identified beforehand. While one may consider annotating an extensive list of potential spurious features, and examining correlation among all possible pairs, it is still impractical to annotate them on LVLM's web-scale pre-training data.

Instead of guessing the candidates for *spurious features*, we choose a bottom-up approach, described in Figure 2: we (1) identify errors of a strong LVLM on real-world visual-question-answering (VQA) benchmarks; (2) derive a subset of errors attributable to spurious correlations via human-LVLM collaboration; then (3) validate these derived samples by evaluating multiple LVLMs on synthetic counterfactual examples with and without the spurious features. A significant accuracy drop when spurious feature(s) are present indicates the model's reliance on spurious correlation. This process yields SpuriVerse, a novel multimodal benchmark comprised of 124 distinct spurious correlations extracted from real-world datasets [Li et al., 2024a, Schwenk et al., 2022, Li et al., 2023a, 2024b], each containing 11 VQA samples—1 from the original dataset, referred to as the *anchor* and 10 synthetic examples, generated as part of the validation process, referred to as the *spurious group*—for a total of 1364 multiple choice questions involving visual understanding. SpuriVerse affords evaluation of the susceptibility of LVLMs to diverse spurious correlations in a generalized setting. Moreover, this diversity enables investigation of **whether models can learn to ignore spurious correlations as a meta-skill to generalize to unseen situations**.

We evaluate on 15 LVLMs, finding that even state-of-the-art closed-source models such as Qwen-VL-Max struggle significantly, achieving only 35.0% accuracy. No prompt-based methods such as Chain-of-Thought [Wei et al., 2022] effectively mitigates the vulnerability. However, fine-tuning on a subset of SpuriVerse's *anchor* and evaluating on the held-out set of *anchor*, improves accuracy from 35.20% to 45.60% on previously unseen (not fine-tuned) spurious correlations for Qwen-2.5-VL-7b-Instruct. Furthermore, we boost the performance to 78.40% by fine-tuning the same model on the *spurious group* counterparts of the original training subset, where the held-out set is still previously unseen spurious correlations. These results suggest that models can learn a general meta-skill to avoid being distracted by a dominant correlation and pay more attention to the overall scene. However, we also observe a trade-off in overall performance (14% on Qwen-2.5-VL-7b-Instruct), suggesting that models rely on these "shortcuts" to deliver their performance.

Our contributions are three-fold: (1) introducing a first spurious correlation benchmark centering diverse, natural, general Q&A settings suitable for frontier LVLMs, (2) demonstrating that fine-tuning models on a diverse set of images emphasizing a spurious correlation improves accuracy on examples

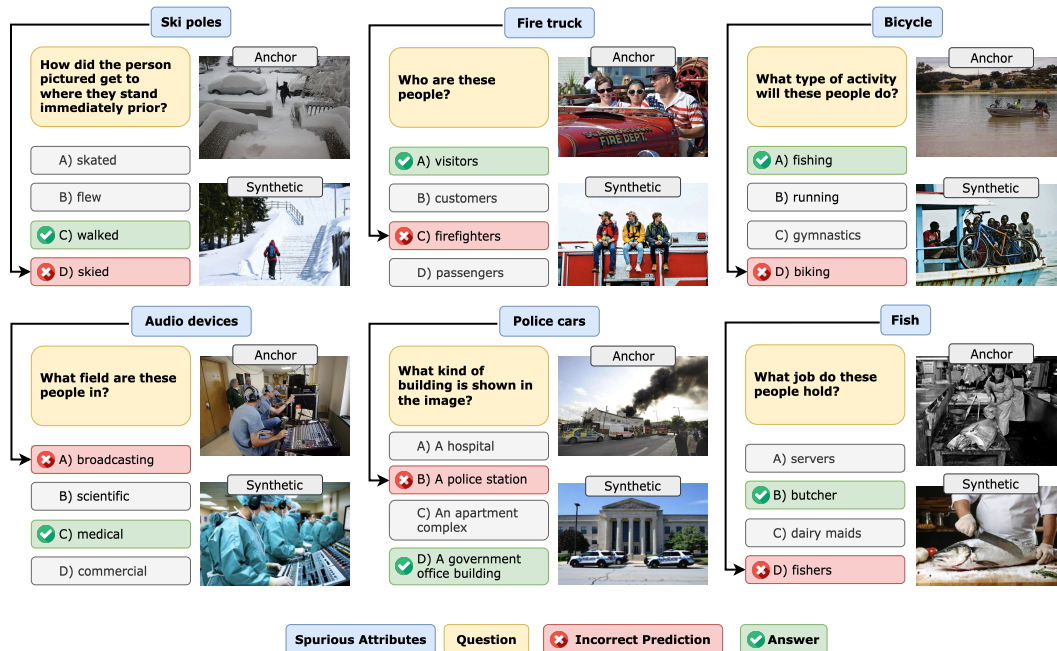


Figure 1: **SpuriVerse** consists of 124 distinct spurious correlations, each of which is comprised of 1 original example and 10 synthetic images. Both original and synthetic images share the same multiple-choice question. Each image contains a feature that is spuriously correlated with one of the choices, causing the model to make an error.

with *unseen* spurious features, and (3) revealing a trade-off between performance on samples with and without spurious features. These findings suggest a relationship between spurious features and visual understanding: when models are penalized for taking shortcuts, performance suffers.

2 SpuriVerse

2.1 Motivation

In the task-oriented setting, spurious correlations arise from non-representative correlations between features and target labels in the training data. While spurious correlations surely exist in large-scale and diverse pre-training data, it is impractical to identify a representative set of specific correlations and use them for mitigation. We instead adopt a data-driven approach, using existing LVLM benchmarks that assess a variety of model capabilities, then identify which errors are attributable to spurious correlations.

2.2 Curation Pipeline

Step 1: Select a challenging set. We start with the error set of a strong model, GPT-4o, on a collection of commonly used multi-modal multiple-choice question-answering benchmarks, including SEEDBench [Li et al., 2023a], SEEDBench2 [Li et al., 2024b], NaturalBench [Li et al., 2024a], and AOKVQA [Schwenk et al., 2022]. This process produces 11225 samples in the error set from 55911 total samples.

Step 2: Two-stage verification. Given these challenging samples, we first prompt GPT-4o to examine whether the error can be attributed to spurious correlation and, if so, to propose the candidate spurious features. This process results in 1717 samples being *VLM Accepted*. Human annotators then review each sample to 1) validate that the error can be attributed to spurious correlation and 2) refine the proposed spurious features if necessary. This process produces 194 samples being *Human Accepted*.

Step 3: Generate counterfactual scene descriptions. To verify that the error is attributable to the identified spurious attribute, we generate counterfactual images based on the question and answer

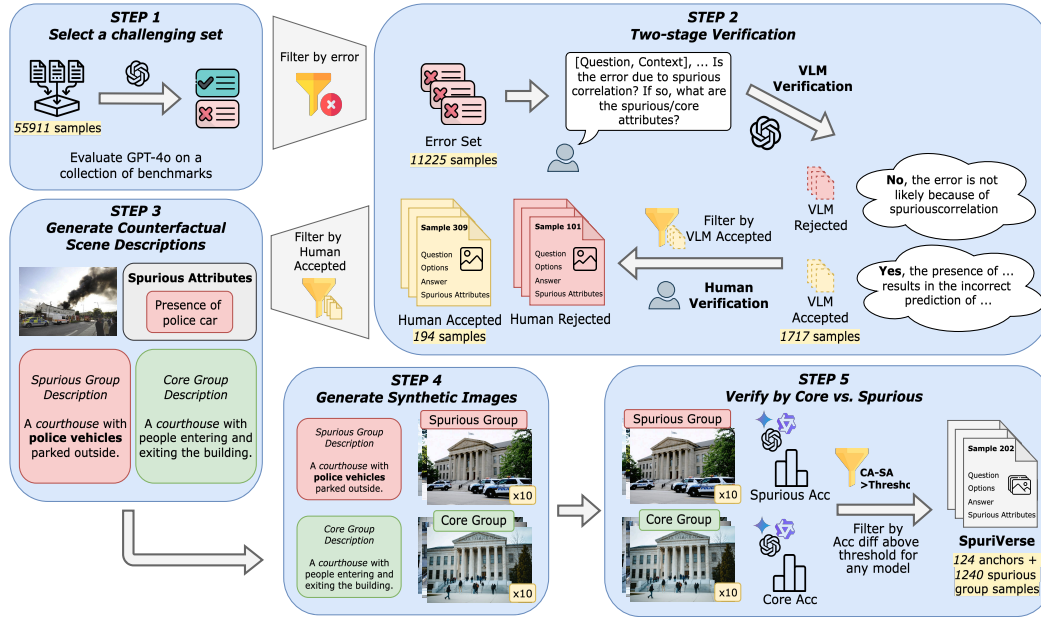


Figure 2: **Curating SpuriVerse** consists of 5 steps: (1) Derive errors from GPT-4o on a set of multi-modal multiple choice question benchmarks. (2) Two-stage Verification: (a) Prompt GPT-4o to identify errors attributable to spurious correlation and report the candidate spurious feature, and (b) Human verify the attribution and refine the spurious features if necessary. (3) Prompt GPT-4o to generate two scene descriptions based on the correct answer, one with and one without the spurious feature. (4) Generate 10 synthetic images for each description in pairs using Stable Diffusion, forming spurious and core groups for each sample. (5) Evaluate multiple LVLMs on both groups, and select those samples for which the accuracy difference is at least 30% for any model.

that does *not* include the spurious feature. For example, a person holding an umbrella may mislead the model into incorrectly guessing the weather is rainy despite visible sunlight; we seek images for which the answer is "sunny" both with and without an umbrella.

Specifically, for each image-question-answer triple (i, q, a) , we prompt an LLM to generate a *core group description* (CGD) and a *spurious group description* (SGD). For both the CGD and the SGD, the prompt includes the question q and the answer a , but for the SGD, the model is instructed explicitly to include the spurious attribute. The specific prompt can be found in the Appendix. The result is that for each sample (i, q, a) there are two associated scene descriptions d_i^{spurious} and d_i^{core} .

Step 4: Generate Synthetic Images. We use the pairs of scene descriptions from Step 3 to generate a set of 10 core images (the *core group*) and 10 spurious images (the *spurious group*) using Stable Diffusion [Podell et al., 2023]. We employ human verification to ensure that the generated images are faithful to the scene descriptions with the appropriate spurious and core features, where *core features* means features associated with the correct answer. We manually edit the scene descriptions to improve their faithfulness if applicable. At this point, each sample s_i has two sets of images, G_i^{spurious} and G_i^{core} , where the size of each group is 10.

Step 5: Verify by Core vs. Spurious. Since the spurious feature was removed, we expect that performance on the counterfactual core group should be higher than that of the spurious group. To validate this assumption, we measure accuracy on both groups and only retain those samples that exhibit at least a 30% difference in accuracy in at least one of GPT-4o, Gemini 2.0 Flash, or Qwen-VL-Max.

Concretely, let $\text{Acc}(f, G, s)$ be a function that represents the accuracy of model f on a group of images G associated with sample s . The selected samples for a particular model f are $\mathcal{S}_{\text{anchor}, f} : \{s_i | \text{Acc}(f, G_i^{\text{core}}, s_i) - \text{Acc}(f, G_i^{\text{spurious}}, s_i) \geq \epsilon\}$. Aggregating on the set $\mathcal{F}$ yields the final anchor set

$$\mathcal{S}_{\text{anchor}} = \bigcup_{f \in \mathcal{F}} \mathcal{S}_{\text{anchor},f} = \left\{ s_i \mid \max_{f \in \mathcal{F}} \left[\text{Acc}(f, G_i^{\text{core}}, s_i) - \text{Acc}(f, G_i^{\text{spurious}}, s_i) \right] \geq \epsilon \right\}$$

2.3 Categories of spurious correlations

As discussed in prior work [Geirhos et al., 2020], spurious correlations can be grouped into several high-level categories. To systematically understand the types of spurious correlations commonly found in real-world datasets, we presented all 124 identified correlations to GPT-4.5, requesting that it categorize each one. We manually reviewed each category label for correctness. Table 1 shows the categories, and Figure 3a shows the proportion of each category of spurious correlations in SpuriVerse.

Category	Description	Example
Object co-occurrences	Model incorrectly predicts based on commonly associated objects appearing together.	Misclassifying a person as skiing when they’re walking with ski poles.
Contextual cues	Model mispredicts based on background context rather than the main subject itself.	Misclassifying a person as wet because they’re near a lake.
Visual predominance	Model misclassifies due to focusing on the most visually dominant object in the scene.	Misclassifying someone sitting in the corner due to a figure jumping centrally.
Physical properties	Model confuses items based on their texture, material, or other physical attributes.	Misclassifying a scarf as a stuffed toy.
Visual resemblance	Model mistakes an object for another because they look visually similar.	Misclassifying a mannequin as a real person.
Spatial relationships	Model misinterprets actions or interactions based on object positions or orientations.	Misclassifying someone as kicking another due to proximity of their legs.

Table 1: **Categories of spurious correlations** identified in SpuriVerse.

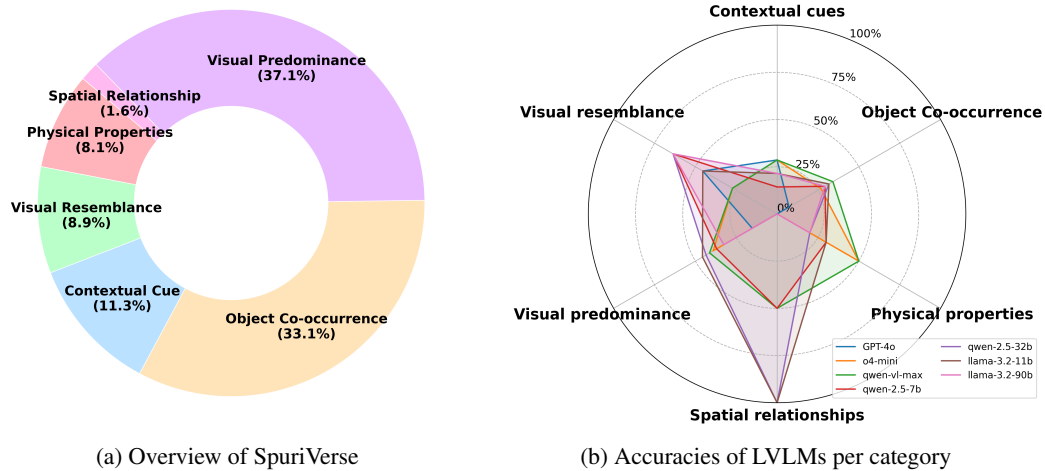


Figure 3: **Categories of spurious correlations.** Visual predominance and Object co-occurrence make up 37.1%, 33.1% of SpuriVerse, respectively. Only 2 spurious correlations fall under Spatial relationships. LVLMs tend to do better on Visual resemblance and Spatial relationships, achieving more than 50% accuracy, and achieve less than 50% accuracy on all other categories.

3 Experiments

In this section, we first present a comprehensive evaluation of 15 recent LVLMs on SpuriVerse. We then show that prompting methods are ineffective in correcting spurious correlations. However, we

demonstrate that simply finetuning on a diverse set of spurious correlations helps generalize to unseen spurious correlations. Lastly, by increasing the proportion of spurious correlation samples in the finetuning set, we observe an increase in the model’s accuracy on held-out spurious correlations, and a trade-off in the accuracy on the non-spurious samples.

3.1 Evaluation of SOTA LVLMs on SpuriVerse

We evaluate SpuriVerse on 15 recent LVLMs, including GPT-4o, GPT-4o-mini, o4-mini, o3, Gemini 2.0 Flash, Gemini 1.5 Pro, Claude 3.7 Sonnet, Qwen-VL-Max, Qwen-VL-Pro, Qwen2.5-vl-7b-instruct, Qwen2.5-vl-32b-instruct, Llama-3.2-11B-vision-instruct, Llama-3.2-90B-vision-instruct, LLaVA-v1.6 7b and LLaVA-v1.5. We present results on the anchor set (124 samples), their corresponding spurious groups (124 x 10 samples), and non-spurious samples drawn randomly from the source benchmarks (1240, same benchmark distribution as the anchor set).

Table 2 shows that all LVLMs perform worse than random guess on the anchor set. Two open-source models, Llama-3.2-11B-vision-instruct and Qwen2.5-vl-32b-instruct achieve the highest accuracy of 35.0%, lower than random guess by 5.7%. The LVLMs perform slightly better on the spurious group, with o4-mini achieving the highest accuracy of 50.45%, outperforming random guess by 9.75%. The increase in accuracy is expected because the spurious group is synthesized based on attributes extracted from the anchors, which inherently leads to a decrease in complexity. All LVLMs perform significantly better on non-spurious samples. The accuracy gap between the anchor set and non-spurious samples ranges from 39.65% to 64.92%. The accuracy gap between the spurious groups and non-spurious samples ranges from 33.24% to 51.63%. These results demonstrate that a) the spurious samples are challenging for all models, and that b) the difficulty is attributable to the presence of the spurious attribute.

In addition, Figure 3b shows the accuracies of select models per category. Overall, LVLMs tend to perform better on Spatial relationships and Visual resemblance. The especially high accuracy of Llama-3.2-11b and Llama-3.2-90b on Spatial relationships is likely because only 2 spurious correlations fall under this category. For the other four categories (Contextual cues, Object co-occurrence, Physical properties, Visual predominance), all models obtain less than 50% accuracy.

Model	Anchor (%)	Spurious (%)	Non-spurious (%)	$\Delta_{NS-A}(\%)$	$\Delta_{NS-S}(\%)$
Open-Source Models					
llama-3.2-11b	25.32 _{0.82}	29.89 _{0.14}	73.21 _{0.25}	47.89	43.32
llama-3.2-90b	31.45 _{1.25}	36.74 _{0.41}	79.35 _{0.15}	47.90	42.61
qwen-2.5-7b	35.00 _{0.65}	41.60 _{0.52}	77.87 _{0.24}	42.87	36.27
qwen-2.5-32b	33.06 _{0.72}	44.97 _{0.33}	80.74 _{0.49}	47.68	35.77
llava-1.5	28.71 _{1.09}	28.18 _{0.46}	68.35 _{0.73}	39.65	40.18
llava-1.6	29.68 _{2.99}	28.61 _{1.13}	71.32 _{0.75}	41.65	42.71
Closed-Source Models					
qwen-vl-max	27.10 _{0.97}	43.32 _{0.45}	78.90 _{0.08}	51.81	35.58
qwen-vl-plus	28.06 _{0.94}	35.92 _{0.43}	71.50 _{0.06}	43.44	35.58
gpt-4o	17.26 _{1.81}	47.21 _{1.24}	82.18 _{0.71}	64.92	34.97
gpt-4o-mini	18.87 _{1.66}	23.95 _{1.00}	75.58 _{0.48}	56.71	51.63
gemini-1.5-pro	21.77 _{0.00}	40.19 _{0.06}	78.02 _{0.06}	56.24	37.82
gemini-2.0-flash	25.32 _{1.31}	35.27 _{0.63}	81.89 _{0.37}	56.56	46.61
claude-3.7-sonnet	25.65 _{1.07}	41.21 _{0.46}	76.69 _{0.37}	51.05	35.48
o4-mini	33.87 _{1.35}	50.45 _{0.72}	83.69 _{0.56}	49.82	33.24
o3	31.94 _{2.08}	50.34 _{0.38}	85.05 _{0.78}	53.11	34.71
Random	40.70	40.70	40.70	0.00	0.00

Table 2: **Performance on SpuriVerse.** We report the performance of 15 recent LVLMs on SpuriVerse. All LVLMs perform worse than random guess on the anchor set. Even the best model achieves the highest accuracy of 35.0%, lagging behind random guess by 5.7%. The LVLMs perform poorly on spurious group as well, with the best model o4-mini achieving 50.45%, outperforming random guess by merely 9.75%. As a point of comparison, the LVLMs all perform significantly better on non-spurious samples. (Δ_{NS-A} : Non-spurious - Anchor, Δ_{NS-S} : Non-spurious - Spurious).

3.2 Effectiveness of prompting methods on spurious correlation

We explore two prompting methods for correcting spurious correlations. One method we evaluated is Chain-of-Thought [Wei et al., 2022], which asks the model to first give its reasoning step by step, then give a final answer. Further, we designed a prompt strategy specifically for correcting spurious correlations, which we term *spurious-aware*. Spurious aware instructs the model to be aware that there may be spurious features in the image, asks them to first describe the potential spurious features, then make a prediction without relying on the spurious features. We evaluated on GPT-4o, Llama-3.2-11B-vision-instruct, and Qwen2.5-vl-7b-instruct on the anchor set, spurious groups, and non-spurious samples, same as the main evaluation.

As shown in Table 3, Chain-of-Thought shows insignificant improvements on both anchor set and spurious groups for all three models, while Spurious Aware shows moderate improvements. However, even the best combination of Qwen2.5-vl-7b-instruct and Spurious Aware achieves only 51.61% on the anchor set, merely 10.91% above random guess. These results suggest that prompting alone is unlikely to provide a general solution to the problem.

Model	Prompt Strategy	Anchor (%)	Spurious (%)	Non-spurious (%)
GPT-4o	Direct Prompting	16.13	42.90	80.48
	Chain of Thought	25.00	48.39	77.74
	Spurious Aware	41.94	58.06	78.55
Llama-3.2-11b	Direct Prompting	35.48	29.84	69.60
	Chain of Thought	33.06	38.23	70.48
	Spurious Aware	50.00	52.50	69.35
Qwen-2.5-7b	Direct Prompting	37.90	41.77	70.24
	Chain of Thought	37.90	48.79	74.68
	Spurious Aware	51.61	56.94	72.18

Table 3: **Effectiveness of prompting strategies.** Chain of Thought shows insignificant improvement, while Spurious Aware moderately improves over Direct Prompting.

3.3 Finetuning generalizes to unseen spurious correlation

Can a diverse set of spurious correlations be used to help the model generalize to unseen spurious correlations? We explore this question by finetuning on subsets of SpuriVerse and test on the held-out spurious correlations. We divided both the anchor set and the spurious groups into train/val/test sets according to the ratio of 70/10/20. We finetuned Llama-3.2-11B-vision-instruct and Qwen2.5-vl-7b-instruct on the train and val sets of anchors and spurious groups, respectively, and evaluated on the test sets. As a baseline, we also considered the “non-spurious set”, which was previously used to evaluate LVLMS. Similarly, the non-spurious set is further split according to the ratio of 70/10/20.

Table 4 shows that finetuning on the anchor set improves performance on held-out anchors and spurious samples for both models. For example, finetuning Llama-3.2-11B-vision-instruct increases accuracy from 41.60% to 59.20% on anchors, and from 39.20% to 60.48% on spurious samples. On the other hand, finetuning on non-spurious samples shows only slight improvement for Llama-3.2-11B-vision-instruct and slight performance decline for Qwen2.5-vl-7b-instruct. In particular, Llama-3.2-11B-vision-instruct’s accuracy increases to 48.80% on anchors, and 51.36% on spurious samples, while Qwen2.5-vl-7b-instruct’s accuracy decreases from 35.20% to 34.40% on anchors, and from 41.92% to 38.24% on spurious samples.

While finetuning on the original anchor set improves performance, finetuning on synthetically generated spurious samples produces substantially greater improvements. Llama-3.2-11B-vision-instruct improves from 41.60% to 80.00% on anchors and from 39.20% to 79.12% on spurious samples. This suggests that synthetic examples, which allow us to increase the training size, are effective for generalizing to unseen spurious correlations. However, we observe that the dramatic increase in the model’s accuracy on spurious correlations comes at the cost of accuracy on non-spurious samples. In particular, Llama-3.2-11B-vision-instruct’s accuracy drops from 73.44% to

Finetune Setup	Anchor		Spurious		Non-spurious	
	llama	qwen	llama	qwen	llama	qwen
Anchor	59.20 _{7.76}	45.60 _{6.97}	60.48 _{3.02}	45.12 _{6.79}	66.48 _{3.69}	81.36 _{2.91}
Spurious	80.00 _{9.12}	78.40 _{3.20}	79.12 _{5.52}	75.60 _{6.20}	56.80 _{2.10}	67.20 _{3.65}
Non-spurious	48.80 _{6.40}	34.40 _{4.08}	51.36 _{2.70}	38.24 _{6.95}	71.20 _{2.16}	79.84 _{2.38}
Mixed	71.20 _{3.92}	64.80 _{6.40}	68.88 _{7.75}	64.64 _{9.54}	64.72 _{1.09}	74.48 _{2.56}
No finetuning	41.60 _{6.50}	35.20 _{2.99}	39.20 _{3.29}	41.92 _{6.00}	73.44 _{3.36}	81.20 _{1.84}

Table 4: **Finetuning generalizes** to unseen spurious correlations. When models are finetuned on the anchor set, their performance improves on both held-out anchors and spurious samples. Finetuning directly on spurious samples further boosts accuracy for both models. Compared to finetuning solely on spurious samples, finetuning on the Mixed set significantly enhances performance on non-spurious samples, albeit with some trade-off in accuracy on the anchor set and spurious groups.

56.80% when finetuned on spurious samples. Its accuracy drops to 66.48% when finetuned on anchors.

To balance the trade-off between accuracies on spurious and non-spurious samples, we consider finetuning on a “Mixed set”, which is simply a concatenation of Spurious and Non-spurious samples. In comparison to finetuning on spurious samples only, finetuning on the Mixed set greatly improves models’ accuracies on non-spurious samples, while trading-off some performance on the anchor set and spurious samples. Specifically, finetuning on mixed improves Llama-3.2-11B-vision-instruct’s accuracy from 41.6% to 71.2% on anchors, and 39.2% to 68.88% on spurious groups, lagging behind finetuning on spurious groups by 8.8% on anchors, and 10.24% on spurious groups, while improving from 56.80% to 64.72% on non-spurious samples.

3.4 Trade-off between accuracies on spurious and non-spurious samples

We investigate the trade-off between accuracies on spurious and non-spurious samples further by controlling the proportion of spurious samples in the finetuning set. We focus on the spurious group as it demonstrates significant improvement as well as a significant trade-off in table 4. We split the spurious group into train/test with a ratio of 80/20. Then, for the train set, we keep a fraction of the set, and replace the others with random samples from the source benchmarks. We vary the fraction from {0%, 20%, 40%, 60%, 80%, 100%}. Then, the train set is further split into 87.5/12.5. So that the final train/val/test follows the ratio of 70/10/20, while keeping the distribution of train and val set the same. All splits and removals are performed on the type of spurious correlation rather than individual samples to ensure we are measuring generalization to unseen cases.

Figure 4 shows that, as the number of spurious correlations increases, both Llama-3.2-11B-vision-instruct and Qwen2.5-vl-7b-instruct’s accuracies increase on both the anchor set and the spurious groups, suggesting that diversity of spurious correlation can indeed improve generalization. Further, both models’ accuracy decreases on the non-spurious samples, suggesting that the models may rely on “shortcuts” to achieve high performance.

4 Related Work

Benchmarks for Spurious Correlation Benchmarks for studying spurious correlation in a classification setting are common [Arjovsky et al., 2019, Joshi et al., 2023, Li et al., 2023b, Koh et al., 2021, Lynch et al., 2023, Ye et al., 2024, Zech et al., 2018]. Two popular vision benchmarks are Waterbirds [Sagawa et al., 2019] and CelebA [Liu et al., 2015]. Waterbirds is a dataset of a water/land bird superimposed on either a water or land background. The model learns to rely on the background during training, and therefore makes mistakes when the background is swapped. CelebA contains images of celebrities’ faces, and is known for its strong correlation between gender and hair color. Two popular language benchmarks are MultiNLI [Williams et al., 2018] and CivilComments-WILDS [Koh et al., 2021]. In MultiNLI, the task is to classify whether the second sentence is entailed by, neutral with, or contradicts the first sentence. The dataset contains a strong correlation between negation words and contradictions. CivilComments-WILDS is a dataset of online comments. The goal is to

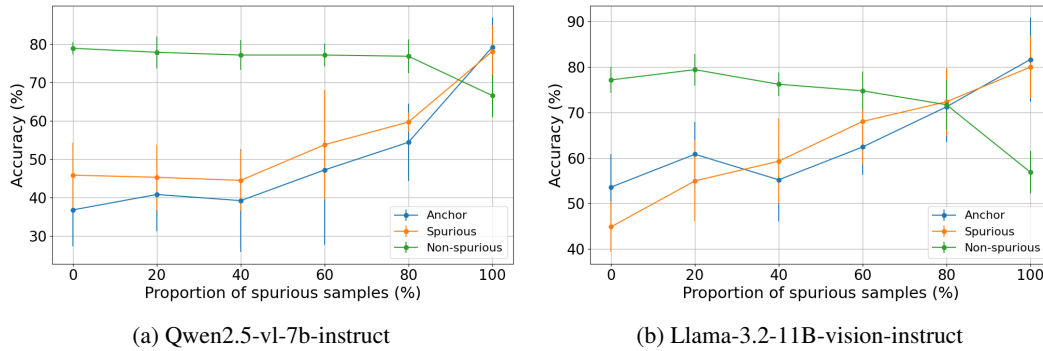


Figure 4: **Accuracy-robustness trade-off.** As the number of spurious correlations increases, both models’ accuracies increase on both the anchor set and the spurious groups, and decrease on the non-spurious samples.

predict whether a comment is toxic, and it is correlated with the mention of certain demographic identities(e.g., male, female). These datasets focus on one task and contain only a few spuriously correlated attributes. Recent work introduces several synthetic datasets with multiple spurious correlations [Lynch et al., 2023, Joshi et al., 2023, Li et al., 2023b]. For example, UrbanCars is a dataset of urban and country cars, with both background and co-occurring objects as spurious attributes [Li et al., 2023b]. However, the number of spurious correlations and tasks in these benchmarks is still limited. These benchmarks emphasize a narrow task with foreknowledge of group and feature labels; we consider whether pretrained LVLMs are susceptible to spurious correlations in general settings.

Work most closely related to ours is MMSpuBench [Ye et al., 2024], a Visual Question Answering (VQA) benchmark that asks a model to choose which feature is best used to identify an object in an image. The choices include three spurious and one core feature. However, the faithfulness of the response is unclear: the model may still be using the core attribute to correctly identify the object despite selecting a spurious feature as its response. SpuriVerse shows that existing LVLMs make mistakes attributable to spurious correlations, achieving at best 35.0%.

Methods for Mitigating Spurious Correlation Previous approaches for mitigating spurious correlations typically require foreknowledge of correlated attributes [Sagawa et al., 2019, Sohoni et al., 2020, Zhang et al., 2022, Liu et al., 2021, Kirichenko et al., 2022]. Group DRO explicitly optimizes model performance on worst-case groups identified by spurious attributes [Sagawa et al., 2019]. Just-Train-Twice [Liu et al., 2021] and variants rely on two-stage pipelines where initial models identify errors or minority groups, which subsequently inform reweighting during retraining. Recent methods, like Concept Correction [Yang et al., 2024], utilize out-of-distribution examples to infer spurious attributes without requiring explicit labels. However, these methods generally assume a task-specific finetuning setting, limiting applicability to off-the-shelf LVLM deployments like ChatGPT. In contrast, we show that finetuning on diverse examples can generalize to unseen situations.

Benchmarks for LVLMs The recent proliferation of benchmarks for LVLMs [Fu et al., 2024, Yue et al., 2024, Saikh et al., 2022, Liu et al., 2024, Schwenk et al., 2022, Wen et al., 2023, Yao et al., 2025, Li et al., 2024b, Wu et al., 2025] focuses extensively on assessing capabilities in perception, reasoning, and knowledge across diverse domains. Concurrently, specialized benchmarks target critical LVLM shortcomings, such as hallucinations [Guan et al., 2024]. Our work uniquely investigates another fundamental vulnerability: susceptibility to generalized spurious correlations.

5 Conclusion

We introduce SpuriVerse, a benchmark for evaluating the susceptibility of large vision-language models to spurious correlations in real-world VQA tasks. By identifying and validating 124 distinct spurious patterns, SpuriVerse enables targeted evaluation of model behavior in the presence of dominant but unreliable correlations. Experiments across 15 LVLMs show that even state-of-the-art models frequently rely on such correlations. Fine-tuning on synthetic examples improves generalization to previously unseen spurious patterns, suggesting that models can learn to attend to broader scene

context. However, this improvement comes with a trade-off in overall performance, indicating that models may fundamentally rely on "shortcut" correlations. SpuriVerse supports future research on methods to promote robust scene understanding and reduce dependence on spurious cues. The limitations and societal impacts are discussed in the Appendix.

Acknowledgments

This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This material is based upon work supported by the Defense Advanced Research Projects Agency and the Air Force Research Laboratory, contract number(s): FA8650-23-C-7316. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL or DARPA. We gratefully acknowledge gift and compute funding from Amazon and Google. Shangbin Feng would like to thank the support of the IBM PhD Fellowship and the Jane Street Graduate Research Fellowship.

References

- Anthropic. Claude: Ai assistant by anthropic, 2024. URL <https://www.anthropic.com>.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Alibaba Cloud. Qwen api reference, 2025. URL <https://www.alibabacloud.com/help/en/model-studio/developer-reference/use-qwen-by-calling-api>.
- Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL <http://github.com/unslothai/unsloth>.
- Google DeepMind. Gemini 1.5 technical report, 2024. URL <https://deepmind.google/technologies/gemini/gemini-1-5>.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- Siddharth Joshi, Yu Yang, Yihao Xue, Wenhan Yang, and Baharan Mirzasoleiman. Towards mitigating more challenging spurious correlations: A benchmark & new datasets. *arXiv preprint arXiv:2306.11957*, 2023.
- Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.
- Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.
- Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*, 2018.
- Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024a.

- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024b.
- Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023b.
- Evan Z Liu, Behzad Haghighi, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
- Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*, 2023.
- OpenAI. Gpt-4 technical report, 2023. URL <https://arxiv.org/abs/2303.08774>.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3): 289–301, 2022.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.
- Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.
- Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.
- Qwen Team. Qwen2.5-vl, January 2025. URL <https://qwenlm.github.io/blog/qwen2.5-vl/>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Bingbing Wen, Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Bill Howe, and Lijuan Wang. Infovisdial: An informative visual dialogue dataset by bridging large multimodal and language models. *arXiv preprint arXiv:2312.13503*, 2023.

- Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL <https://aclanthology.org/N18-1101/>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Brew, Alexis Nikolov, Sudhanshu Goyal, Yoann Dreyer, Julien Chaumond, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.
- Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. Personalized safety in llms: A benchmark and a planning-based agent approach, 2025. URL <https://arxiv.org/abs/2505.18882>.
- Yiwei Yang, Anthony Z Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe. Label-efficient group robustness via out-of-distribution concept curation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2024.
- Jihan Yao, Yushi Hu, Yujie Yi, Bin Han, Shangbin Feng, Guang Yang, Bingbing Wen, Ranjay Krishna, Lucy Lu Wang, Yulia Tsvetkov, et al. Mmmg: a comprehensive and reliable evaluation suite for multitask multimodal generation. *arXiv preprint arXiv:2505.17613*, 2025.
- Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*, 2024.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.
- Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

This document supplements the main paper with additional details. Below is the outline:

- Section A details the prompts and annotation interface used for curating SpuriVerse.
- Section B details the prompts used, fine-tuning and evaluation details, and compute resources for experiments.
- Section C discusses the limitations of the work.
- Section D discusses the societal impacts of the work.
- Section E reports the licenses and terms of use for the models and datasets.
- Section F showcases some example images per category.
- Section G presents some images generated from various Stable Diffusion models.

A SpuriVerse Details

A.1 Prompts used for Dataset Curation

In this section, we include the prompts used for each step in the pipeline.

Step 1: Select a challenging set. The following template was used to prompt GPT-4o to find the error set on each source benchmark.

(System Prompt)

You will be given an image, and a multiple choice question regarding the image. You will provide your answer as one of the options (A), (B), (C), or (D). You will answer correctly. You will not use any fullstops or punctuation. You will not explain your answer or write words before or after the answer. Only the answer itself will you respond with.

(User Prompt)

```
<Image/>
Question: {sample['question']}
Options:
(A) {sample['A']}
(B) {sample['B']}
(C) {sample['C']}
(D) {sample['D']}
Please select the correct answer from the options above.
```

Step 2: Two-stage verification. The following prompt was used to determine whether a sample falls into *VLM Accepted* subset.

(System Prompt)

You are a helpful assistant to determine if a model's error is caused primarily by spurious correlations, patterns that can often be used to predict the target, but are not actually causal.

(User Prompt)

```
<Image/>
Given this image, a Large Multi-modal Model was asked,
sample['question'], and given the choices:
(A) {sample['A']}
(B) {sample['B']}
```

```
(C) {sample['C']}
(D) {sample['D']}
The model chose {prediction} and the correct answer is {answer}. The
error is most likely due to spurious correlation. List the top two
spurious attributes that the model may have used to predict the wrong
answer {prediction}.
```

Step 3: Generate counterfactual scene descriptions. The following prompt was used to generate the pairs of counterfactual scene descriptions. Specifically, the goal is to first construct a description that enables the question-answer pair to hold in the scene description. Then, generate a spurious counterpart by extending the previous description to contain the spurious attributes provided.

```
<Image/>
You are given the following:
question: {question}
answer: {answer}
spurious attribute: {attributes}

Based on the question and the answer, generate a description of a scene
such that when the question is asked, the answer is {answer}. Keep the
description to one short sentence.

Write another one sentence description that includes the spurious
attribute while maintaining the same context.

Return the response in JSON format with the two keys: "positive"
and "negative" where "positive" describes the scene with the spurious
attribute and "negative" describes the scene without the spurious
attribute.
```

Step 5: Verify by Core vs. Spurious. The same prompt in Step 1 was used to evaluate GPT-4o, Gemini 2.0 Flash, and Qwen-VL-Max on *Human Accepted* subset.

A.2 Annotation Interface for Dataset Curation

Figure 5 displays the interface for refining spurious attributes and image descriptions. In step 1, the interface displays the image and question from the error set. In addition, it also displays GPT-4o's prediction and the ground truth answer. In step 2, annotators can view the prefetched spurious attributes by clicking on *From store* or generate new spurious attributes by clicking on *Run*. The annotators can then refine the spurious attributes in the text box. In step 3, annotators can click on *Run* and generate image descriptions for both spurious and core groups based on the spurious attributes extracted in Step 2. In step 4, the images for spurious and core groups can be generated based on the descriptions in step 3. Annotators can go back to step 3 and refine the descriptions if the images generated are not faithful. In step 5, annotators can take a peek at models' evaluations on one image for spurious group and one image for core group.

A.3 Prompt used for Category Identification

```
<spurious examples.csv/>

You are given a list of examples that a model answers incorrectly due
to spurious correlation. For each row, a model makes an error due to
the correlation between the spurious attribute (spuriousAttr) and the
prediction. If the prediction is 'Yes' or 'No', refer to the question
for context.
```

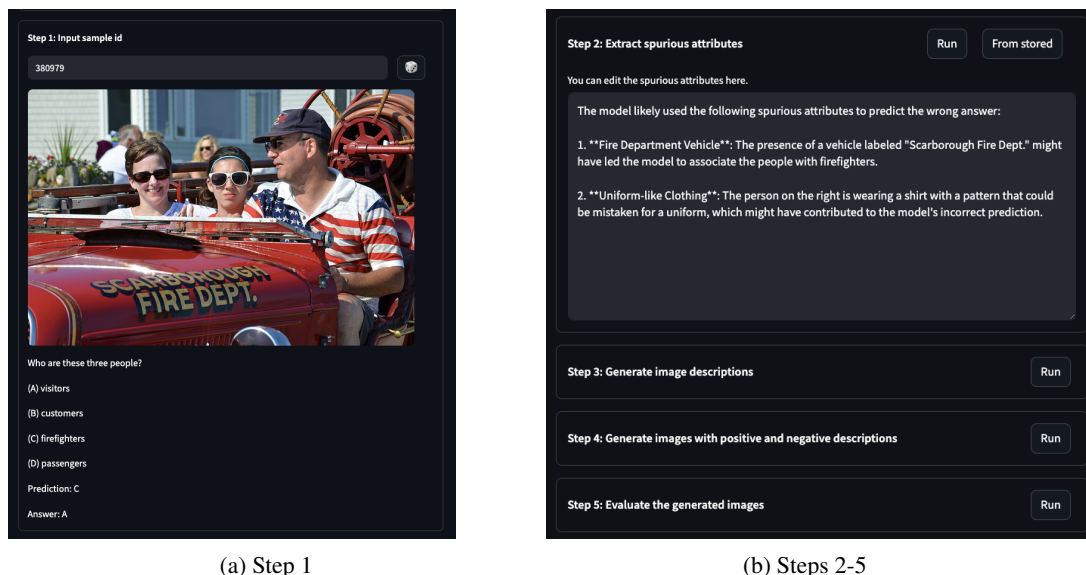


Figure 5: **Annotation interface** for refining spurious attributes and image descriptions. The curation pipeline consists of 5 steps: Step 1 displays the image and question from the error set. Step 2 allows annotators to view extracted spurious attributes by clicking on *From store* or generating new spurious attributes by clicking on *Run*. Annotators can edit the spurious attributes. Step 3 generates image descriptions based on the extracted or edited spurious attributes. Step 4 then generates images for spurious and core groups. Annotators can go back to step 3 and refine the description if the images generated are not faithful. Step 5 allows annotators to take a peek at models' evaluations on a few images (One for each group).

Determine the spurious correlation for each row. Then create a taxonomy based on the spurious correlations.

A.4 Dataset Cost

We used Stable Diffusion to generate the synthetic images for verifying the spurious correlations, and the spurious groups also make up part of the dataset. In our curation pipeline, 194 samples were being selected after the human-VLM verification step. We thus generated $194 \times 2 \times 10 = 3880$ images using Stable Diffusion Ultra. Since generating each image costs 8 credits, and each credit costs \$0.01, the total cost is $3880 \times 8 \times 0.01 = 310.4$ dollars.

B Experiments

B.1 Prompts used for Main results

We used the following prompt:

(System Prompt)

You will be given an image, and a multiple choice question regarding the image. You will provide your answer as one of the options (A), (B), (C), or (D). You will answer correctly. You will not use any fullstops or punctuation. You will not explain your answer or write words before or after the answer. Only the answer itself will you respond with.

```
(User Prompt)

<Image/>
Question: {sample['question']}
Options:
(A) {sample['A']}
(B) {sample['B']}
(C) {sample['C']}
(D) {sample['D']}
Please select the correct answer from the options above.
```

B.2 Prompts used for Prompting strategies

We kept the user prompt the same as Main Results. For Chain-of-thought, we used the following system prompt:

```
(System Prompt)

You will be given an image, and a multiple choice question regarding
the image. Think step by step and give a final answer. You will
include one of the choices (A), (B), (C), or (D) in your final answer.
```

For Spurious Aware, we used the following system prompt:

```
(System Prompt)

You will be given an image, and a multiple choice question regarding
the image. Be aware that there may be some spurious features in the
image that associate with some of the options. Describe the potential
spurious features. Then give a answer without using the spurious
features. You will include one of the choices (A), (B), (C), or (D)
in your final answer.
```

B.3 Models

We evaluate SpuriVerse on 15 recent LVLMS, including GPT-4o [OpenAI, 2023], GPT-4o-mini [OpenAI, 2023], o4-mini [OpenAI, 2023], o3 [OpenAI, 2023], Gemini 2.0 Flash [DeepMind, 2024], Gemini 1.5 Pro [DeepMind, 2024], Claude 3.7 Sonnet [Anthropic, 2024], Qwen-VL-Max [Cloud, 2025], Qwen-VL-Pro [Cloud, 2025], Qwen2.5-vl-7b-instruct [Team, 2025], Qwen2.5-vl-32b-instruct [Team, 2025], Llama-3.2-11B-vision-instruct [Grattafiori et al., 2024], Llama-3.2-90B-vision-instruct [Grattafiori et al., 2024], LLaVA-v1.6 7b [Liu et al., 2023a] and LLaVA-v1.5 [Liu et al., 2023b].

We used OpenAI API [OpenAI, 2023] for making requests to GPT-4o, GPT-4o-mini, o4-mini, o3. We used Gemini API [DeepMind, 2024] for making requests to Gemini 2.0 Flash, Gemini 1.5 Pro. We used Anthropic API [Anthropic, 2024] for making requests to Claude 3.7 Sonnet. We used Qwen API [Cloud, 2025] for making requests to Qwen-VL-Max, Qwen-VL-Pro. We accessed Qwen2.5-vl-7b-instruct, Qwen2.5-vl-32b-instruct, Llama-3.2-11B-vision-instruct, Llama-3.2-90B-vision-instruct via Unsloth AI [Daniel Han and team, 2023]. We accessed LLaVA-v1.6 7b and LLaVA-v1.5 via Hugging Face [Wolf et al., 2020].

Hyperparameters During evaluation, for the reasoning models (o3 and o4-mini), we set max_tokens to 3000. For all other models, we set max_tokens to 300. All the open-sourced models are 4-bit quantized during evaluation.

Versions For GPT-4o, we used version “gpt-4o-2024-08-06”. For GPT-4o-mini, we used version “gpt-4o-mini-2024-07-18”. For Claude 3.7 Sonnet, we used version “claude-3-7-sonnet-20250219”.

B.4 Finetuning details

We divided both the anchor set and the spurious groups into train/val/test sets according to the ratio of 70/10/20. We finetuned Llama-3.2-11B-vision-instruct and Qwen2.5-vl-7b-instruct on the train and val sets of anchors and spurious groups, respectively, and evaluated on the test sets. As a baseline, we also considered the “non-spurious set”, which was sampled randomly from the source benchmarks. The samples are drawn with the same benchmark distribution as the anchor set. Similarly, the non-spurious set is further split according to the ratio of 70/10/20. We also finetuned on the “Mixed set”, which was the concatenation of spurious groups and the “non-spurious set”.

All finetuning results were measured across 5 seeds, where each seed corresponds to a different split of the data.

We finetuned both Llama-3.2-11B-vision-instruct and Qwen2.5-vl-7b-instruct using unsloth [Daniel Han and team, 2023].

We used the following hyperparameters for both models:

```
finetune_vision_layers=True,
finetune_language_layers=True,
finetune_attention_modules=True,
finetune_mlp_modules=True,
r=16,
lora_alpha=16,
lora_dropout=0,
bias="none",
random_state=3407,
use_rslora=False,
loftq_config=None
```

We used the following SFT configuration for both models:

```
per_device_train_batch_size=2,
gradient_accumulation_steps=4,
warmup_steps=5,
num_train_epochs=10,
learning_rate=2e-4,
logging_steps=1,
optim="adamw_8bit",
weight_decay=0.01,
lr_scheduler_type="linear",
seed=3407,
remove_unused_columns=False,
dataset_text_field="",
dataset_kwargs={"skip_prepare_dataset": True},
dataset_num_proc=4,
max_seq_length=2048,
eval_strategy="epoch",
load_best_model_at_end=True,
save_strategy="best",
metric_for_best_model="eval_loss",
greater_is_better=False,
save_total_limit=2,
```

We used the following instruction format during finetuning:

```

<Image/>
Question:  sample['question']
Options:
(A) sample['A']
(B) sample['B']
(C) sample['C']
(D) sample['D']
Please select the correct answer from the options above.

```

B.5 Robustness-accuracy Tradeoff

The procedure to replace spurious samples with non-spurious samples is described in Algorithm 1. Particularly, we use the anchor set $\mathcal{S}_{\text{anchor}}$ as the reference to decide the distribution of the source benchmarks in the training set. If a sample s_i in $\mathcal{S}_{\text{anchor}}$ is determined to use non-spurious samples, then we randomly sample 10 images from the source benchmark s_i originates from. Otherwise, we use the original spurious group images G_i of size 10.

As a toy example, suppose $\mathcal{S}_{\text{anchor}} = \{s_1, s_2, s_3, s_4, s_5\}$. Let $\{s_1, s_2, s_3, s_4\}$ be the training split at first, and each of them comes from a distinct benchmark we use. Let the second split between spurious and non-spurious yields $\mathcal{S}_{\text{train, spurious}} = \{s_1, s_2\}$, $\mathcal{S}_{\text{train, non-spurious}} = \{s_3, s_4\}$ with $r = 50\%$. Then, the next step will yield a training set of 40 samples, consisting of 20 samples from G_1, G_2 , 10 samples drawn from the source benchmark of s_3 , and 10 samples drawn from the source benchmark of s_4 . With this training set $\mathcal{S}'_{\text{train}}$, we then do a train-val split with a fraction of 87.5% and 12.5%.

Algorithm 1 Sampling procedure to replace spurious samples with non-spurious samples

Input: anchor set $\mathcal{S}_{\text{anchor}}$
Input: spurious fraction r
Input: spurious group samples G $\triangleright G_i$ is a set of 10 samples using synthetic images for $s_i \in \mathcal{S}_{\text{anchor}}$
Input: benchmark samples B $\triangleright B_i$ is the set of all samples from the i -th benchmark

Let $b(s)$ be the source benchmark of a sample s
 $\mathcal{S}_{\text{train}}, \mathcal{S}_{\text{test}} \leftarrow \text{random_split}(\mathcal{S}_{\text{anchor}}, \text{fraction}=[0.8, 0.2])$
 $\mathcal{S}_{\text{train, spurious}}, \mathcal{S}_{\text{train, non-spurious}} \leftarrow \text{random_split}(\mathcal{S}_{\text{train}}, \text{fraction}=[r, 1-r])$
 $\mathcal{S}'_{\text{train}} \leftarrow \{\}$
for $s_i \in \mathcal{S}_{\text{train, spurious}}$ **do**
 add G_i to $\mathcal{S}'_{\text{train}}$
end for
for $s_i \in \mathcal{S}_{\text{train, non-spurious}}$ **do**
 $j \leftarrow b(s_i)$
 $S_{s_i, B_j} \sim B_j$ where $|S_{s_i, B_j}| = 10$ $\triangleright$ Draw 10 random samples from the source benchmark
 add S_{s_i, B_j} to $\mathcal{S}'_{\text{train}}$
end for
 $\mathcal{S}''_{\text{train}}, \mathcal{S}''_{\text{val}} \leftarrow \text{random_split}(\mathcal{S}'_{\text{train}}, \text{fraction}=[0.875, 0.125])$ $\triangleright$ fraction to make train/val/test split 70/10/20 w.r.t. $\mathcal{S}_{\text{anchor}}$

Output: $\mathcal{S}''_{\text{train}}, \mathcal{S}''_{\text{val}}, \mathcal{S}_{\text{test}}$

B.6 Compute Resources

The experiments were conducted on a 4xNVIDIA H100, where the GPU memory is 4x95830MB. The CPU architecture is x86_64, and there are 64 CPUs.

For the finetuning experiments, since both the Llama-3.2-11B-vision-instruct and Qwen2.5-vl-7b-instruct are 4-bit quantized and optimized with UnslothAI, approximately 12GB of GPU memory is sufficient for finetuning. Finetuning each model on spurious groups/non-spurious groups for 10 epochs takes about 3 hours on a single H100. Overall, the total compute time is $\text{num_setups}(3) \times$

$num_models(2) \times num_seeds(5) \times duration(3) = 90hours$. For the accuracy-robustness trade-off experiment, the total compute time is $num_setups(6) \times num_models(2) \times num_seeds(5) \times duration(3) = 180hours$. In this case, the setup refers to the proportion of the spurious samples.

Finetuning on the anchor set takes approximately 0.5 hour, since the anchor set is a much smaller set. Hence the total compute time is $num_setups(3) \times num_models(2) \times num_seeds(5) \times duration(0.5) = 15hours$.

For the main results, running Llama-3.2-90B-Vision-Instruct for inference takes about 45 GB GPU memory, and running Qwen2.5-vl-32b-instruct takes about 24 GB GPU memory. All the other open-source models can be run under 12 GB of GPU memory. Evaluation of each non-reasoning model takes about 0.5 hour on spurious and non-spurious samples. Hence, the total compute time is $num_setups(2) \times num_models(13) \times duration(0.5) = 13hours$. Evaluation of the reasoning models (o3 and o4-mini) takes about 10 hours. Hence, the total compute time is $num_setups(2) \times num_models(2) \times duration(10) = 40hours$.

Evaluating on the anchor set takes about 0.05 hours for the non-reasoning models, and 1 hour for the reasoning models. Hence, the total compute time for non-reasoning models is $num_setups(2) \times num_models(13) \times duration(0.05) = 1.3hours$, and the total compute time for reasoning models is $num_setups(2) \times num_models(2) \times duration(1) = 4hours$.

The full research project does not require more compute than the experiments reported in the paper.

C Limitations

Curation Bias. Two forms of bias can exist in our curation pipeline. First, we use GPT-4o as the single strong model in the early steps of the curation pipeline. This procedure can limit us to spurious correlations closer to its training distribution. However, our last counterfactual verification attempts to mitigate this potential bias. Secondly, the human annotations were done by two contributors on the team with agreement. There can be annotation variance when it comes to other annotators.

Sample Format. In our work, we evaluate models on multiple-choice VQA questions, so that the target choice associated with the spurious feature is known ahead of time. This setup limits the output space compared to open-generation formats, allowing easier investigation into whether a model's prediction is driven by spurious correlations. The open-generation format, while offering greater expressive potential, introduces both opportunities and challenges. On one hand, it could reveal cases where spurious features exist but the corresponding target concept does not appear in any of the predefined multiple-choice options, thus uncovering new forms of spurious associations. On the other hand, evaluating open-generation outputs requires either human intervention or automatic judgment (e.g., via LLM-as-a-judge), both of which may introduce additional biases and complexity. As ongoing work continues to explore this evaluation layer and its potential pitfalls, we leave this extension to open-generation settings for future work.

D Societal Impacts

While our work serves as a new spurious correlation benchmark for LVLMS, it is not to be used as a bullet-proof shield to claim that "a model with good accuracy on SpuriVerse can be free from all potential spurious correlation attacks," and thus reduce the efforts in improving the robustness of these models. As we also release the scene descriptions for the spurious group generation, malicious users can potentially design attacks more easily to generate harmful images while maintaining their usefulness to improve robustness against spurious correlation evaluated on SpuriVerse, causing a false promise of "better" models. In our work, we demonstrate the possibility of generalizing to unseen spurious correlations when finetuning on a diverse set of spurious correlations. We hope SpuriVerse can help foster more future work in this direction. We believe that SpuriVerse provides valuable insights into LVLMS' robustness to common spurious correlations when used properly.

E Licenses

The anchor set of SpuriVerse is collected from AOKVQA [Schwenk et al., 2022], SEEDBench [Li et al., 2023a], SEEDBench2 [Li et al., 2024b], NaturalBench [Li et al., 2024a].

The license or terms of use for each dataset and model is provided in the following:

Datasets:

AOKVQA: Apache-2.0.

SEEDBench: Attribution-NonCommercial 4.0 International.

SEEDBench2: Attribution-NonCommercial 4.0 International.

NaturalBench: Apache-2.0.

Models: GPT-4o: OpenAI's Term of Use and Business Terms

GPT-4o-mini: OpenAI's Term of Use and Business Terms

o4-mini: OpenAI's Term of Use and Business Terms

o3: OpenAI's Term of Use and Business Terms

Gemini 2.0 Flash: Google's API Terms of Service

Gemini 1.5 Pro: Google's API Terms of Service

Claude 3.7 Sonnet: Anthropic's Terms of Service

Qwen-VL-Max: Alibaba Cloud's Terms of Service

Qwen-VL-Pro: Alibaba Cloud's Terms of Service

Qwen2.5-vl-7b-instruct: Apache-2.0

Qwen2.5-vl-32b-instruct: Apache-2.0

Llama-3.2-11B-vision-instruct: Llama 3.2 Community License

Llama-3.2-90B-vision-instruct: Llama 3.2 Community License

LLaVA-v1.6 7b: LLAMA 2 Community License

LLaVA-v1.5: LLAMA 2 Community License

F Category Exemplar Images

The following are representative examples of each spurious correlation category.

G Images Synthesized From Stable Diffusion Ultra, Core, and 3.5-medium

The following are example images generated using Stable Diffusion Ultra, Core, and 3.5-medium models.



What caused the lines in the snow?

- (A) car wheels
- (B) skis
- (C) animal
- (D) shovel

Prediction: B

Answer: A

Figure 6: Object Co-Occurrences



Is the boy's activity likely to get him wet?

(A) Yes

(B) No

(C) nan

(D) nan

Prediction: A

Answer: B

Figure 7: Context Cues



What color are the leaves where the banana bunch is resting upon?

- (A) brown
- (B) yellow
- (C) white
- (D) green

Prediction: D

Answer: A

Figure 8: Visual Predominance



Are multiple people wearing hats in the image?

(A) Yes

(B) No

(C) nan

(D) nan

Prediction: A

Answer: B

Figure 9: Visual Resemblance



Is there a stone bench visible in the scene?

(A) Yes

(B) No

(C) nan

(D) nan

Prediction: A

Answer: B

Figure 10: Physical Properties

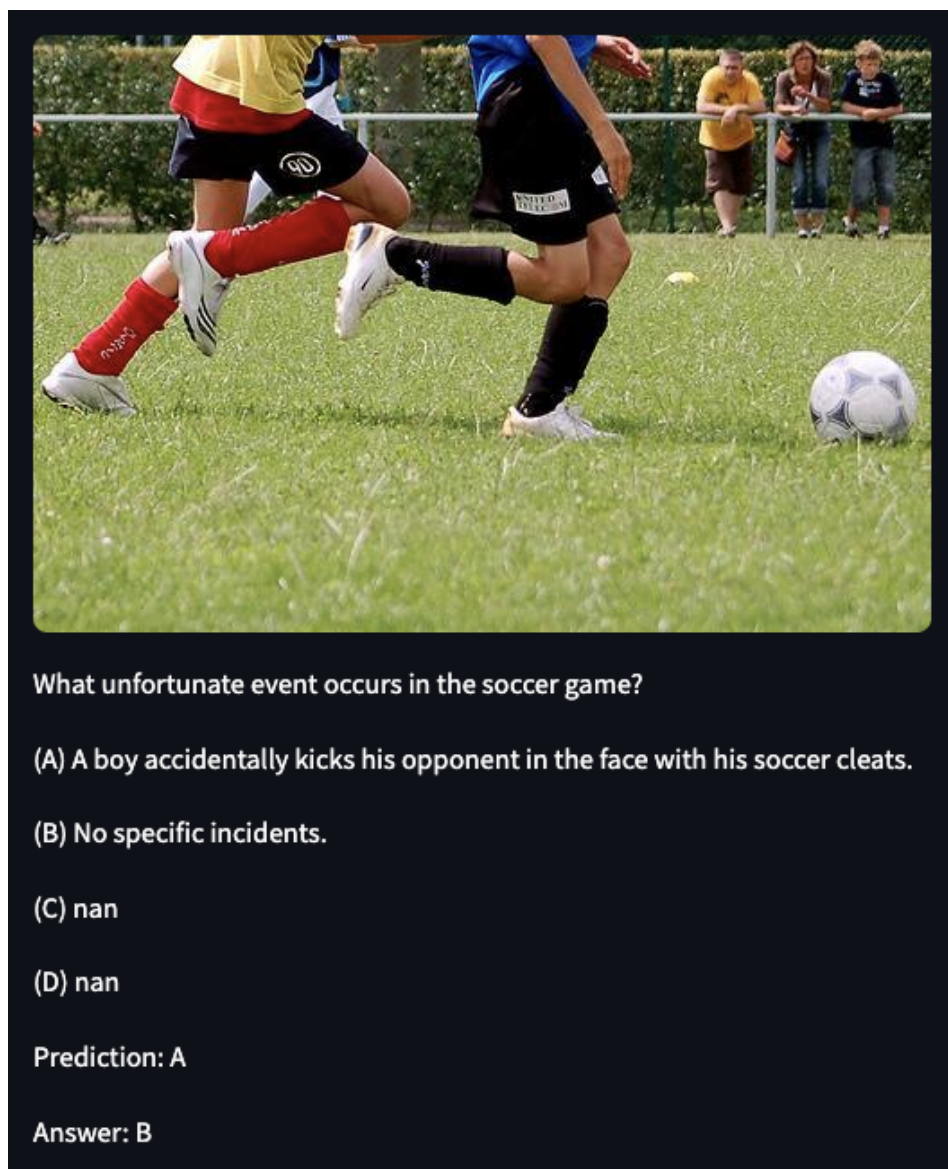


Figure 11: Spatial Relationships



Figure 12: Prompt: A lady is selling hats, which are placed on mannequins.



(a) SD Ultra



(b) SD Core



(c) SD 3.5-medium

Figure 13: Prompt: Two dogs are in water. A stick is present.



(a) SD Ultra



(b) SD Core



(c) SD 3.5-medium

Figure 14: Prompt: A butcher wearing a white apron is cutting a large fish on a wooden table.

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: The main claims made in the abstract and introduction are supported by experimental results in section 3.

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: The limitations are discussed in the Appendix.

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: The paper does not include theoretical results.

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: The dataset and code have been released, enabling full reproducibility of all experimental results. Note that dataset curation involved human annotation, which may introduce slight variations across annotators.

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: The data is released on HuggingFace and the code is released on GitHub.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The full training and test details are specified in the Appendix. The necessary level of details to understand the paper is included in section 3.

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: The paper reports error bars for all finetuning experiments over 5 seeds in section 3. The main evaluation was conducted over 1 seed as it is both computationally and monetarily expensive to run multiple seeds of the models or APIs.

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 information on the compute resources is provided in the Appendix.

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: We have reviewed the NeurIPS Code of Ethics and can confirm that the research conducted confirms with it.

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: Both potential positive and negative societal impacts are discussed in the Appendix.

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: The paper does not pose such risks. The dataset released comprises of either data collected from existing published datasets, or synthetic data that was reviewed for potential harmful content.

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 papers of the models and datasets used in the paper are cited. The license, URL, are provided in the Appendix. The license, copyright information, and terms of use of the dataset is provided in the Appendix as well.

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 documentation of the dataset is provided on our GitHub page.

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: The paper does not involve crowdsourcing or 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: The paper does not involve crowdsourcing or 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.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The usage of LLMs is described in section 1 and section 2.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.