Mitigating Lies in Vision-Language Models

Junbo Li, Xianhang Li, Cihang Xie
University of California, Santa Cruz

Abstract

In this work, we bring new insights into the honesty of vision-language models, particularly in visual question answering (VQA). After a throughout revisit of the existing 'lie' behavior in pure language models, our work makes an unprecedented extension of 'lies' to vision-language models. The results indicate that the lie prefixes have a more obvious misleading effect on vision-language models than on language models. We also propose a novel visual prefix and prove that the consistent vision-language prefix is more threatening to vision-language models. To defend the models from the stated 'lies', we put forward an unsupervised framework based on Gaussian mixture modeling and obtain improvement with 3% against the language prefix and 12% against the vision-language prefix.

1 Introduction

Vision-language models are gaining popularity in various scenarios, where their robustness and safety are worth noticing. Researchers have investigated different aspects of the model robustness in vision-language tasks (including visual question answering (VQA) [1, 9, 8, 7] and visual grounding [3, 2]). *E.g.*, for the VQA task, the robustness includes two aspects: the model resilience to distribution shift in both training set and answer type, and the model brittleness to rephrased languages and manipulated images.

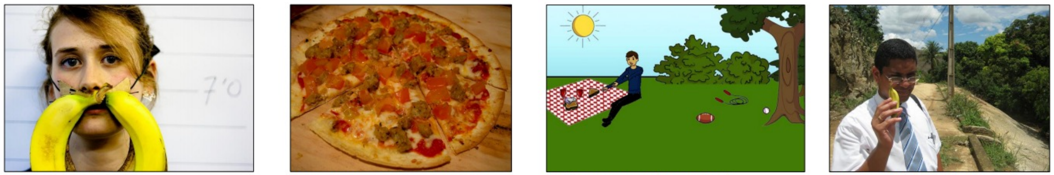
The robustness study of multi-modality models can derive from that of single-modality models. Recently, [5] explores the honesty of language models. The authors find that some prefixes can mislead the language model to output false text, specifically in the case of binary-type questioning and answering (QA). They call this phenomenon 'lie' behavior, which is different from previous works aiming to prevent models from accidentally outputting false results. Viewing binary-type QA as a classification task, they propose a completely unsupervised learning framework, which only performs clustering on the test set. Their unsupervised clustering not only proves but also mitigates the model's 'lie' behavior, and it works much better than a direct generation.

Such 'lie' behavior may also be a severe drawback in current vision-language models, but the related research remains untouched yet. In this work, we investigate the 'lie' behaviors in vision-language models and propose practical solutions. Our results show that vision-language models are much easier to be fooled by the language lie prefix. Besides the single-modality language prefix, we find that the consistent language prefix and vision manipulation are more threatening to vision-language models. To solve these problems, we propose completely unsupervised methods based on Gaussian mixture models. Under the consistent language prefix and vision manipulation settings, our solution successfully mitigates the 'lie' behavior with accuracy improvements of 3% and 12%.

2 Preparations and basic formulations

In this section, we first recap the study in [5], which explores lie behaviors in a pure language setting. Then we present the basic settings in our work.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Figure 1: Examples in VQA-v2 datasets.

2.1 Lie behavior in language models

[5] focuses on binary-type question-answering (QA) problems. They found that some wrong prefix, *e.g.*, ‘is the United States inside of England? yes’, can make language models have wrong output. They call this prefix a ‘lie’. They then construct a contrastive pair for each question, and show that **unsupervised** clustering for pair encoding can mitigate the lie behavior.

2.2 Basic formulations for vision-language setting

Task selectioon [5] discusses QA, specifically with binary-type answers, *e.g.*, true or false, yes or no. As an extension to the vision-language setting, we consider visual-question answering (VQA) [4] problem.

Dataset selection The commonly used dataset in VQA is VQA-v2 [6], which includes multiple different categories of question-answer types, including yes or no, numbers, *etc*. Here we clarify two notations $\mathbb{A}_0$ and $\mathbb{A}_1$ about the dataset:

$$\mathbb{A}_0 = \{(x_i, q_i)\} : \textit{the original dataset VQA-v2, where } (x_i, q_i) \textit{ is an image-question pair.}$$

$$\mathbb{A}_1 \subset \mathbb{A}_0 : \textit{the sub-dataset of yes-or-no type.}$$

Besides multiple types of answers in $\mathbb{A}_0$, we also note here that each image in $\mathbb{A}_0$ has multiple different corresponding questions and respective answers, as shown in Figure 1 [4]. This will be illustrated later. For an image x, denote $Q(x, \mathbb{A})$ to be a list of all questions on x in $\mathbb{A} \subset \mathbb{A}_0$, and $A(x, \mathbb{A})$ to be a list of all corresponding answers.

Model selection We choose OFA [10] as the backbone, and use the checkpoint obtained after finetuning on $\mathbb{A}_0$ [1].

3 VQA with language lie prefix

3.1 Lie prefix with arbitrary answer type

Lie prefix in [5] is a group of question-answer pairs that have no semantic relations to the target question-answer pair. In the vision-language setting, we argue that the language prefix should be consistent with the visual input. Hence, we need the dataset to contain multiple questions for one image, as shown in Figure 1. For each image and question pair $(x, q) \in \mathbb{A}_1$, we randomly select a question q' from $Q(x, \mathbb{A}_0)$ and the corresponding answer a' from $A(x, \mathbb{A}_0)$ as the prefix. So a' could be any type of answer. After adding a language prefix, the language part changes from q to $q' + a' + q$. We find that the lie prefix can severely mislead the output answer. Accuracy drops from 90.6% to 45.01%, and around 37.6% answers are neither ‘yes’ nor ‘no’. The result shows that, compared with language models, VQA models could be far more easily misled by lie prefixes.

3.2 Lie prefix with yes-or-no answer type

3.2.1 Generate answers directly using OFA model

To avoid the issue mentioned above, we restrict the lie prefix to only the yes-or-no type. This requires an additional restriction on the dataset. We choose $\mathbb{A}_2 \subset \mathbb{A}_1$ such that for any $(x, q) \in \mathbb{A}_2$,

[1] https://github.com/ofa-sys/ofa

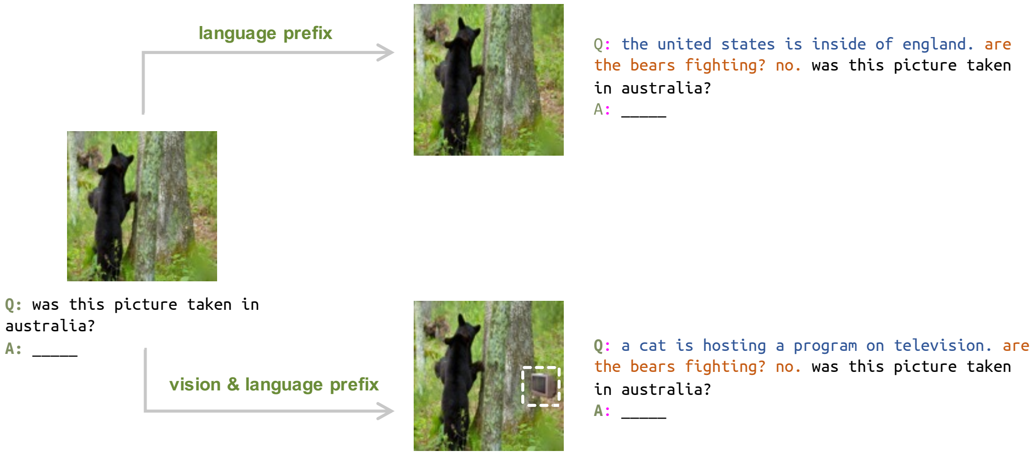

language prefix
Q: the united states is inside of england. are the bears fighting? no. was this picture taken in australia?
A: _____
Q: was this picture taken in australia?
A: _____
vision & language prefix
Q: a cat is hosting a program on television. are the bears fighting? no. was this picture taken in australia?
A: _____

Figure 2: The top is an image with a language prefix. The blue text is inherited from [5]. The yellow text is another question and its corresponding false answer from $\mathbb{A}_2$ for the same image. The bottom is an image with vision & language prefix, where a television patch is added. The blue text is a sentence related to but not consistent with the added object. The yellow text is the same as the language-prefix case.

$|Q(x, \mathbb{A}_2)| > 1$. This means for each image and question pair in $\mathbb{A}_2$, we can always find another different yes-or-no question belonging to the same image in $\mathbb{A}_2$.

Now for each $(x, q) \in \mathbb{A}_2$, we select a pair (x, q') with the corresponding answer a' from $\mathbb{A}_2$ instead of $\mathbb{A}_0$. We also inherit the prefix used in [5] and add it to the beginning. We show an example with all these prefixes in Figure 2. Experiments show that with the yes-or-no prefix, almost all answers belong to ['yes', 'no']. Now accuracy drops from 90.6% to 64.3%. This indicates that the lie behavior can cause around 25% performance drop.

3.2.2 Unsupervised learning on vision-language encoding

[5] mitigates this lie behavior by unsupervised clustering. For a given binary-type question q with lie prefix $q' + a'$, they construct a contrastive pair by answering 'yes' and 'no' respectively: $q' + a' + q +$ *'yes'* and $q' + a' + q +$ *'no'*. For an encoder h, the final representation is:

$$q' + a' + q \rightarrow H(q' + a' + q) = (h(q' + a' + q + \textit{'yes'}) - h(q' + a' + q + \textit{'no'}))\,[-1].$$

Here $[-1]$ means to take the encoding of the last token. $H(\cdot)$ is used to train an unsupervised binary classifier, because they argue that for questions with opposite answers, $H(\cdot)$ would lie on different spaces. They propose two clustering methods based on principal component projection and variance minimization.

Our OFA model has an encoder-decoder structure, where the image tokens and text tokens are concatenated in one sequence during encoding. For an image-question pair (x, q), the encoding is $h(x, q) \in \mathbb{R}^{N \times d}$, where $N \approx 100$ is the number of tokens, $d = 1024$ is the embedding dimension. Researchers in [5] simply choose the last token, however, we find this does not work for vision-language models. We use average pooling to obtain more information. So the representation is:

$$(x, q'+a'+q) \rightarrow H(x, q'+a'+q) = mean(h(x, q'+a'+q+\textit{'yes'}) - h(x, q'+a'+q+\textit{'no'})) \in \mathbb{R}^d.$$

Different from [5], we use mixture Gaussian models for unsupervised clustering on $H(\cdot)$. Results show that the accuracy improves from 64.3% to 67.3%.

4 VQA with vision-language lie prefix

In this section, we explore a novel visual prefix. It is straightforward to add texts in the front of text questions as prefix guidance. However, there lacks a counterpart prefix of images. We propose to add some meaningful semantic objects to the original image. Moreover, we add a sentence related to the objects at the beginning of the text. Figure 2 presents an example of our prefix design.

We also make inferences on $\mathbb{A}_2$ with both direct generation and unsupervised clustering. Since the vision & language prefix shares the same data structure as before, the pipeline is totally the same. First, the accuracy of generation using OFA model drops from 90.6% to 52.4% with the vision & language prefix, which is nearly random guessing. However, unsupervised clustering with a Gaussian mixture model still remains a good performance as before, which is 64.7%. This illustrates that unsupervised clustering helps mitigate the lie prefix on both images and texts.

We list some objects and sentences in Table 1.

objects	sentences
television	a cat is hosting a program on television.
truck	a truck is flying to the moon.
computer	a monkey is typing in front of a computer.
motorbike	a dog is driving a blue motorbike on water.

Table 1: Objects and corresponding sentences.

5 Experiments summary and conclusions

The main results are summarized in Table 2. We recap some main conclusions here.

dataset	vision prefix	language prefix 1	language prefix 2	direct generation	out of [‘yes’, ‘no’]	unsupervised clustering
$\mathbb{A}_1$	none	none	none	90.6%	1.2%	
$\mathbb{A}_1$	none	none	$Q(x, \mathbb{A}_0)$	45.0%	37.6%	
$\mathbb{A}_2$	none	inherit from [5]	$Q(x, \mathbb{A}_2)$	64.3%	0.0%	67.3% (↑**3.0%**)
$\mathbb{A}_2$	objects	objects related	$Q(x, \mathbb{A}_2)$	52.4%	0.0%	64.7% (↑**12.3%**)

Table 2: The main results in our work. Language prefixes 1 and 2 are added to the front of the text. Direct generation and out of [‘yes’, ‘no’] refers to the accuracy and percentage when using the original OFA model. Unsupervised clustering refers to the accuracy of Gaussian mixture models.

Vision-language models are more easily misled by the language prefix. Although [5] also focuses on binary-type questions, their language prefix is of arbitrary types other than [‘yes’, ‘no’]. Different from our results, theirs do not show such a strong misleading effect.

Consistent vision-language prefix is more dangerous than language prefix for vision-language models. Simple language prefix still keeps the accuracy at a reasonable level. However, the consistent vision-language prefix degraded the model to random guessing.

Unsupervised clustering helps mitigate lie behaviors. The results show that with good representation extraction, unsupervised clustering shows a better performance than a direct generation. When tested with a consistent vision-language prefix, the direct generation performs like a random guess, while the unsupervised clustering still maintains stable performance as before.

Acknowledgments and Disclosure of Funding

This work is supported by a gift from Open Philanthropy.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

[2] Arjun Akula, Varun Jampani, Soravit Changpinyo, and Song-Chun Zhu. Robust visual reasoning via language guided neural module networks. *Advances in Neural Information Processing Systems*, 34:11041–11053, 2021.

[3] Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*, 2020.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. 2022.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[7] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *arXiv preprint arXiv:1906.08430*, 2019.

[8] Gouthaman Kv and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. In *European Conference on Computer Vision*, pages 18–34. Springer, 2020.

[9] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019.

[10] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

A X-Risk Analysis

'Lie' behavior refers to the phenomenon that models give wrong results when adding some wrong guidance while giving correct results with no such guidance. This paper makes the first exploration of 'lie' behavior for vision-language models. Our formulation and experiments show that vision-language models are much easier to be misled by lie prefix than pure language models. Besides pure language manipulation, we also show that consistent vision and language manipulation are a more threatening risk for vision-language models.

We then propose a completely unsupervised method to solve this potential problem. We build contrast pairs of features of the same vision-language input and conduct unsupervised learning. Results indicate that our method (based on Gaussian mixture modeling) outperforms direct generation, and it retains higher accuracy even with stronger lie prefixes on both vision and language.

A.1 Long-Term Impact on Advanced AI Systems

In this section, please analyze how this work shapes the process that will lead to advanced AI systems and how it steers the process in a safer direction.

1. **Overview.** How is this work intended to reduce existential risks from advanced AI systems?
   Answer: Our work can reduce existential risks from advanced AI systems. To illustrate this, we first make a comparison of our work and previous work. Our exploration derives from a previous work that first propose the concept of lie behaviors and explore this in pure language models. They propose unsupervised methods that mitigate the problem. Our formulation shows that the related formulation is very different from pure language models. For example, 1. the lie prefix should be related to visual input, 2. the prefix on visual input also needs to be considered.

   Although there are many differences in formulations, our work shows that purely unsupervised methods can also work in vision-language models. Advanced AI systems would contain multi-modality modules including both vision and language. Therefore, a thorough formulation and analysis of risks in vision-language systems are crucial. Moreover, our work shows the generalizability of our proposed unsupervised learning framework. Combined with the previous work, our work verifies that the unsupervised learning framework can learn the intrinsic information of high-dimensional cross-modality input.

2. **Direct Effects.** If this work directly reduces existential risks, what are the main hazards, vulnerabilities, or failure modes that it directly affects?
   Answer: We study the vulnerability of current well-performed vision-language models in VQA tasks. A simple language prefix as in previous works [5] can significantly reduce performance (e.g. 90.6% v.s. 64.3 %), which shows the current system can take a high risk of being fooled. In our work, we further demonstrate a stronger 'attack' on this system by manipulating the image and language simultaneously. We add a small picture with a corresponding description onto the original image, which not only fools the language model but also takes the vision encoder into consideration. Compared with the previous language-prefix-only method, our method further reduces the performance by **12 % (**64.3% v.s. 52.4%). It shows that our current VQA models can be easily attacked by adding meaningful object and object-related prefixes to both vision and language encoders. To address this issue, we propose to use an unsupervised clustering method to implicitly de-noise the feature that contains the 'lie' prefix. We find a Gaussian mixture model can bring us **3%** and **12%** improvement in both two settings. It inspires future studies to develop advanced methods which de-noise the features with lie prefixes and reduces risks.

3. **What's at Stake?** What is a future scenario in which this research direction could prevent the sudden, large-scale loss of life? If not applicable, what is a future scenario in which this research direction be highly beneficial?
   Answer: In our work, we show that taking the vision encoder into consideration for robustness evaluation is necessary. A meaningful image with a concrete description can mislead the vision-language model than only adding a language prefix. We believe a future study can be a more in-depth robustness evaluation of both vision and language models instead of only considering one of them. Second, we find a simple Gaussian mixture model can significantly reduce the effect of lie behavior. We believe our work contributes to the

future development of advanced clustering or de-noising algorithms to enhance robustness. Moreover, one specific scenario or method can be: we can mix the different images and prefixes together when training the vision language model to see if we can improve the robustness, similar to the mixup methods in image processing.

4. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? No
5. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? No
6. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? No
7. **Competitive Pressures.** Does work towards this approach strongly trade-off against raw intelligence, other general capabilities, or the economic utility? No

A.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

1. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: First of all, our goal is not to improve the performance of the model, so we did not retrain or restructure the model to achieve better performance. Our goal is to analyze the robustness of the existing model, so we select the trained model for evaluation. Under this scenario, we design a more effective method to fool our model by manipulating the image and language input simultaneously. Second, our proposed method to improve the robustness also does not modify the model but performs unsupervised clustering at the output level of the model, and we find that this method is very effective in improving the robustness of the vision and language models.
2. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
   Answer: In our findings, a classical unsupervised clustering method works better than well-designed clustering methods in [5]. One way to further improve robustness can be mixing the different images and prefixes when training vision-language models.
3. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? Yes
4. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the Internet, or similar capabilities? Yes
5. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? Yes
6. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? Yes