– Independent Work Report Spring 2023 –

Reducing Object Hallucination in Visual Question Answering

Tanushree Banerjee
Advisor: Professor Olga Russakovsky

Abstract

Mitigating object bias is crucial to ensure VL models are trustworthy and reliable enough to be deployed in high stakes real world applications. This paper focuses on studying a specific type of object hallucinations triggered when VQA models are queried with an image and unrelated question. This paper proposes an evaluation procedure to evaluate the extent to which VQA models are effectively able to identify unrelated questions, as well as proposes and evaluates several possible approaches to identify unrelated questions. The best approach found achieves a 40% improvement over the random baseline. However, This approach does not perform significantly better than a naive linear classifier, indicating that investigating more sophisticated approaches may be necessary in order to effectively identify unrelated image-question pairs. Code is available at: https://github.com/tanushreebanerjee/iw-spring-23

The recent advancements in pre-trained large language models (LLMs) and vision-and-language pre-training (VLP) methods [10] have led to the development of AI systems capable of remarkable performance on vision-grounded text generation tasks, including image captioning and visual question answering (VQA) [5]. Such impressive performance presents an exciting opportunity for AI systems to make high-stakes processes much more accurate and efficient. For example, such systems could help assist physicians in making medical diagnoses [22], aid visually impaired users as a multi-modal assistant [29] and enable more intuitive interfaces to interact with technology such as robots, virtual assistants, et cetera [7].

Despite their impressive performance on existing benchmarks [4], these vision-language (VL) models are still prone to generating unfaithful or nonsensical output given the source input [17]. This major failure mode of VL models has been termed *hallucination* [17]. A major type of hallucination

in VL models is known as *object hallucination* or *object bias*, where the model generates text predicting non-existent or inaccurate objects from the input image [25]. Although the performance of state-of-the-art VL models has been pushed to the limit based on accuracy on standard benchmark datasets for tasks like image captioning and VQA, this has not translated into a decrease in object hallucination [6]. Moreover, object hallucination does not seem to diminish by scaling the dataset or the model size [14], making it a crucial failure mode to study in order to develop more reliable and robust VL models that can be trusted by humans enough to be deployed in high-stakes real-world applications.

Object hallucination severely limits the performance of VL models [10], and raises significant safety concerns in industry applications of such models [10]. For example, object hallucination in biomedical image captioning for diagnosis can significantly reduce test accuracy and have severe consequences for the patient [22]. Moreover, predicting non-existent or inaccurate objects is especially undesirable in applications involving aiding visually impaired users, where correctness is much more preferable than coverage [6]. In addition, since object hallucination may be indicative of over-fitting to the training data, mitigating object hallucination is crucial to improve the generalization capabilities of VL models, allowing them to adapt to unseen domains more easily [6]. Lastly, mistakes made by VL due to object bias are typically mistakes that a human would never make, since they aren't especially hard examples for humans to respond to. Thus, if a model is prone to object bias, the model is likely to make obvious mistakes for relatively easy examples. This would greatly reduce the trust human users have on the system, since mistakes made on easier examples are more likely to erode a user's trust on the system to a greater extent [9].

Despite the limitations and risks posed by object hallucination in downstream applications, this problem is relatively understudied in VL literature [17]. Therefore, it is crucial to solve the object hallucination problem to ensure such VL models are reliable and safe to deploy in high-stakes real-world downstream applications such as AI-assisted medical diagnoses, assisting the visually impaired, etc.

[4] has identified a few types of input questions that seem to be more likely to elicit object

hallucination VL models, as described in section 1.1. One such type of text input is questions unrelated to the input image. When prompted with a question unrelated to input image, a human would easily be able to recognise when a question is unrelated to an image, and appropriately respond, perhaps by abstaining to answer the question or otherwise indicating that the question asked by the user is unrelated to the image. Yet, existing pre-trained VQA models do not have this ability to abstain from answering a question, and thus attempt to answer the question despite it being unrelated to the question, thereby deviating from the desired human-like behavior on an example that a human would easily be able to respond to appropriately, thereby harming the user's trust on the system to a great extent [9]. Yet, such an ability to abstain from responding to unrelated questions has been largely understudied in multi-modal ML research [29], despite its importance in ensuring such models are reliable and trustworthy enough to be used in real world settings.

Thus, the goal of this paper is to systematically investigate the following two fundamental research questions about the object hallucination problem in VL models.

1. How can we give existing pre-trained VL models the ability to identify when a question is meaningless for a given image?
2. What would it take to effectively build a system that could consider such a scenario and thereby be more robust to object hallucination?

There has been some work done on investigating object hallucination on the image captioning task [6], but relatively little investigation has been done on the VQA task. Moreover, the VQA task in particular has been garnering an increasing interest from the ML research community [27]. The goal of VQA is to answer any question about any given image [15]. This task is especially challenging; not only does it require a high-level understanding of the visual scene and the question, but also adequate use of both modalities as well as grounding of the textual concepts in the image [7]. The successful resolution of VQA has vast real-world implications, such as assisting visually impaired users in comprehending their physical and online surroundings, enabling the use of natural language interfaces to explore large amounts of visual data, or even facilitating communication with robots through more efficient and intuitive interfaces [7]. Object hallucination in VQA

models can be especially harmful since users may base important decisions based on the output in downstream tasks, making it imperative that the output is trustworthy and reliable. Thus, for all the aforementioned reasons, this paper focuses its analysis on the VQA task.

In particular, after discussing previous relevant work done analysing and mitigating object hallucination in VL tasks, this paper first proposes a procedure to evaluate the prevalence of the type of object bias elicited when the model is probed with an image and an unrelated question in section 4.1. Next, section 2.4 explains different possible strategies to give a pre-trained VQA model the ability to identify and appropriately abstain to answer unrelated questions, thereby mitigating object hallucination. These strategies are then evaluated quantitatively using the proposed evaluation procedure in section 4.1. The predictions of the best performing model and a baseline linear classifier are compared qualitatively in section **??** to understand the extent to which these models can successfully identify unrelated questions, as well as the cases where they are unable to do so.

The best proposed approach (evaluated using the proposed evaluation procedure in section 4.1 achieves a 40% improvement in identifying unrelated questions compared to a random classifier. This approach identifies unrelated questions based on the difference in the softmax output of the pre-trained VQA model with a randomly generated image and the original input image, and involves training a Multi-Layer Perceptron (MLP) classifier on top of the pre-trained VQA model. This indicates that there is indeed some signal in the softmax scores of the model that can be exploited in order to identify unrelated questions. However, this approach is not significantly better at identifying unrelated questions than a baseline linear classifier trained on the maximum softmax output of the pre-trained VQA model. Thus, a more sophisticated approach involving a novel architecture and extracting more carefully constructed image features may be necessary to build models that are able to detect unrelated questions at least as well as humans. Work in this paper indicates that features based on object detection model output and image captioning model output may be a good starting point for designing such a more sophisticated approach.

1. Background and Related Work

1.1. Object hallucination in VL models

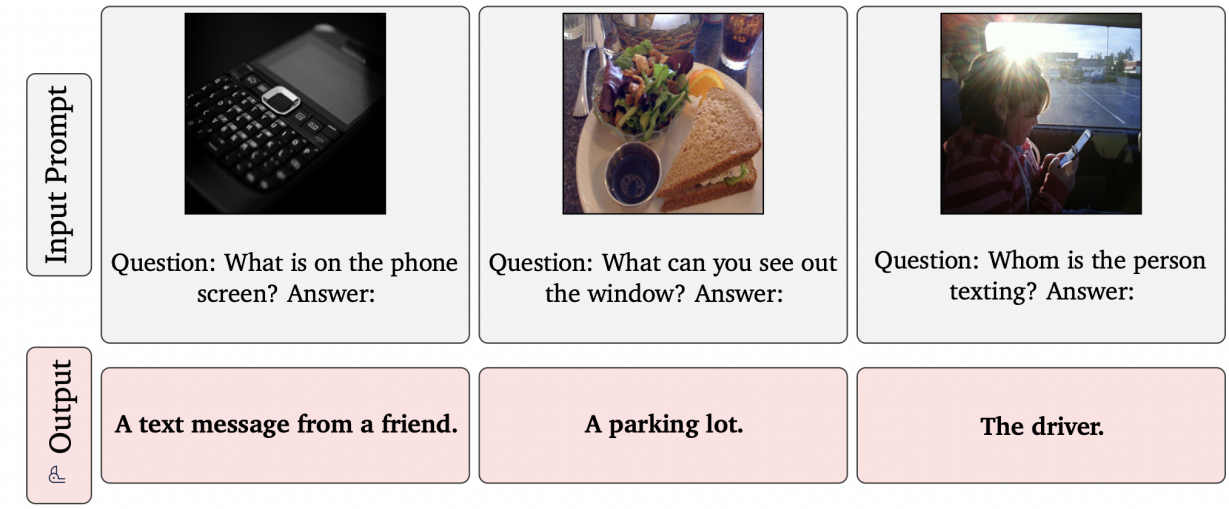

Figure 1: Object hallucinations in open ended VQA by the Flamingo model [4]

Object hallucination in VL models is defined as a prediction by the model containing non-existent or incorrect objects [4]. Three major types of object hallucination that often occur in open-ended VQA models as described in prior work [4] are explained below.

- **Type 1: Answers based on text input only [4].** In this type of object hallucination, the model produces answers that seem likely given the text input only, but are actually wrong given the image input [4]. An example of such an error is shown in Figure 1 on the left.
- **Type 2: Attempt to answer unrelated questions [4].** Similar to the type of object hallucination described above, this type of object hallucination is often elicited by adversarially prompting a VQA model with a question irrelevant to the input image [4], as shown in the example in the middle in Figure 1.
- **Type 3: Ungrounded guesses [4].** In this type of hallucination, the model gives an ungrounded guess when prompted with a question whose answer cannot be determined from the given input image, as shown in Figure 1 on the right.

Other work has attempted to classify hallucinations into two broad categories [17], described below.

- **Intrinsic hallucinations.** In this type of hallucination, the generated textual output contradicts the information provided in the input image [17].
- **Extrinsic hallucinations.** In this type of hallucination, the generated textual output can neither be supported nor contradicted by information in the input image, i.e. the generated textual output cannot be verified by the information in the input image [17].

Hallucinations of type 1 are a type of intrinsic hallucination, while hallucinations of type 2 and 3 are extrinsic hallucinations.

There has been more work done to mitigate intrinsic hallucinations [7], which are described in more detail in section 1.4 below. There has been relatively less work on addressing the extrinsic object hallucinations, which is described in section 1.5 below. Since extrinsic hallucinations and mitigation strategies for them are comparatively understudied in VL literature, and to narrow down the scope of the paper to make it appropriate for a single-term IW, this paper focuses on hallucinations of type 2 in VQA models as described above.

1.2. Attempts to quantify object hallucination in VL models

Prior works studying object bias in VL models have mainly focused on the image captioning task, so the only existing metrics for quantifying object bias are specific to image captioning, including the Caption Hallucination Assessment with Image Relevance (CHAIR) metric described below.

Caption Hallucination Assessment with Image Relevance (CHAIR) [6] This metric quantifies object hallucination on the image captioning task by calculating the proportion of words generated by the model that are actually in the image according to the ground truth sentences and object segmentations. There are two variants of this metric:

- **CHAIRi**, a per-instance metric capturing the proportion of hallucinated objects over all golden objects in all data samples, and can be seen as the probability of a generated object to be a hallucination.

- **CHAIRs**, a per-sentence metric capturing the proportion of generated sentences that contain at least one hallucinated object.

However, these metrics are only applicable to the outputs of an image-captioning model. Moreover, they require a pre-defined list of target object categories to recognize objects in the text, which is specific to the dataset and thus may not generalize to other datasets that contain images from a substantially different domain. Moreover, these metrics do not distinguish between different types of hallucination discussed in section 1.1 above.

Thus, in section 4.1, this paper proposes procedures to quantify how prone VQA models are to type 2 object hallucinations as defined in section 1.1. Focusing on evaluating a specific type of hallucination would help develop a better understanding of this type of hallucination, thereby allowing the development of targeted strategies to mitigate this specific type of hallucination in VQA models.

1.3. Prior attempts to mitigate object hallucination in VL

Prior work on the object hallucination problem in VL models has mainly focused on the image captioning task [6, 31, 10]. The corresponding research on mitigating object bias in image captioning models is detailed below.

Biten et al. [6] propose that the main cause of object bias is the co-occurrence of specific objects in input images with associated text in the training examples. They develop three data augmentation methods to level the co-occurrence statistics in the training dataset, which mitigates object hallucination without altering the model architecture. They only conduct their analysis on the image captioning task.

Xiao and Wang [31] suggest an uncertainty-aware beam search-based approach for decoding, demonstrating that reducing uncertainty mitigates hallucination in image captioning. In particular, a weighted penalty term added to the beam search objective balances the log probability and predicted uncertainty of the selected candidate words.

Dai et al. [10] investigate object bias in VL pre-training and develop a new pre-training objective

called object masked language modeling to address the hallucination problem. They only evaluate their proposed pre-training objective on the image captioning model as well.

Although the above approaches have achieved some success in mitigating object hallucination in image captioning, no prior work has attempted to study object hallucination in other VL tasks. Yet, even modern VL models with impressive performance on other VL tasks [17] have also been shown to be prone to object hallucination. Thus, this paper aims to fill this research gap by focusing on analyzing and object bias on the VQA task. In particular, this paper focuses on type 2 hallucinations as defined in section 1.1 in order to narrow the scope of the paper to be appropriate for a single-term independent work project.

1.4. Biases due to over-reliance on language priors

Cadene et al. [7] hypothesize that VQA models often exploit uni-modal biases to provide the correct answer without using the image information [7], causing a significant drop in performance when these models are evaluated on data outside their training set distribution. The output of the model on data outside the training set distribution often exhibits object hallucination of type 1 as described in section 1.1, and thus makes the models unsuitable for deployment in real-world downstream applications.

In order to mitigate biases arising from over-reliance on language priors, Cadene et al. [7] propose a new learning strategy called RUBi (Reducing Unimodal Biases). Their strategy implicitly forces the VQA model to use both input modalities instead of relying on the statistical co-occurrences between the textual question input and target model output in the training set. They add a question-only model branch on top of the base VQA model during the training phase, which influences the VQA model by dynamically adjusting the loss to compensate for the bias towards the textual modality. This strategy takes advantage of the fact that question-only models are by design biased towards the text modality to reduce the importance of the most biased training examples on the loss function. As a result, the gradients back-propagated through the model are reduced for the most biased examples and increased for the less biased ones. After training is complete, the question-only

branch is removed. Their proposed approach leads to a +5.94 percentage point improvement in accuracy over the state-of-the-art results on VQA-CP v2 [3], a dataset specifically designed to account for question biases [3].

Although this approach deals with hallucinations of type 1 as defined in section 1.1, this approach fails to address hallucinations of type 2 and type 3 as defined in section 1.1. Thus, this paper focuses on developing approaches for hallucinations of type 2, and leaves the analysis of hallucinations of type 3 for future work, since this is beyond the scope of the paper. However, the strategies proposed in this paper could easily be adapted to investigate type 3 errors, which are also extrinsic hallucinations like type 2 errors, as explained in section 1.1.

1.5. Attempts to give VQA models the ability to abstain

Although there has been no work done to mitigate object hallucinations of type 2 and type 3 as defined in section 1.1, Whitehead et al. [29] have done some work on introducing the ability to abstain in several prominent VQA models as well as on formalizing and exploring the notion of reliability of VQA models. They do so by re-framing the VQA task as a selective prediction problem [29] where the model must either predict an answer or abstain from answering. Formulating the VQA task in such a way requires 1) gauging the uncertainty in model predictions and 2) learning when to abstain from answering.

They measure the performance of VQA models under this framework by focusing on the following metrics:

1) coverage, i.e. what proportion of the questions are actually answered by the model and 2) risk, i.e. error rate on the questions that the model does to attempt to answer.

An ideal model would have high coverage along with low risk. However, in reality, there is a trade-off between the risk and coverage of a model. Thus, Whitehead et al. [29] propose a scalar metric called Effective Reliability, which accounts for abstention as well as a penalty for predicting an incorrect answer.

Under their proposed framework, Whitehead et al. [29] show that for many models, using the

maximum probability to determine when to abstain by thresholding the softmax scores leads the model to answer a very small fraction of the questions for a 1% risk of error, despite the model otherwise having high accuracy over standard benchmarks in the regular problem formulation without the option of abstention available. This inability to answer a majority of the questions at a low risk indicates that current VQA models would not be very useful in practical applications.

In order to address this problem, Whitehead et al. [29] explore two approaches to maximize coverage while minimizing risk: 1) calibration and 2) training a multi-modal selection function. While calibration often leads to a better risk-coverage trade-off compared with simply thresholding the softmax scores of the model, the multi-modal selection function based approach consistently improves the coverage of different VQA models across varying risks of error, and especially at low levels of risk. Their proposed Effective Reliability metric is also found to correlate with risk and coverage in a meaningful way [29] .

1.6. Detecting intrinsic difficulty of questions.

There is some previous literature on VQA which involves categorizing and detecting questions that are intrinsically hard to answer or are not answerable, regardless of the model's ability. For instance, the VizWiz VQA dataset [16] contains labels for questions that are unanswerable [16], as well as reasons for annotation entropy, such as low image quality or question ambiguity. Work in Davis, 2020 [11] defines a similar categorization for unanswerable questions in VQA. Teney et al., 2016 [28] show that precision/recall computed based on model confidence scores can be reflective of the ambiguities in the ground truth answers.

I hypothesize that such questions that are intrinsically hard to answer or are unanswerable are more prone to eliciting hallucinations from VQA models. This paper also hypothesizes that for questions that are unrelated or unanswerable given the input image, the model's softmax output would reflect the uncertainty of the model in answering the question, and thus the signal from these softmax outputs could be exploited to detect unrelated questions. This hypothesis will drive the procedure for investigating hallucinations of type 2 as defined in section 1.1. This procedure will be

explained in detail in section 4.1.

2. Approach

2.1. Hypotheses

A key hypothesis for the essential issue that causes type 2 and type 3 error, and extrinsic hallucinations more generally, drives the exploration in this paper – that VQA models are not explicitly trained to identify unrelated questions and never penalized for attempting to answer unrelated or unanswerable questions in the training stage. Thus, they cannot discern unanswerable or unrelated questions and attempt to answer them anyway, even if it makes the prediction with relatively low confidence. Thus, the key novel idea of this paper is to leverage the impressive ability of existing pretrained VQA models, while adding modifications that would allow these models to recognise when an image is irrelevant to the input question, as well as strategies inspired by previous work that could guide the model to better identify irrelevant questions while still maintaining reasonable performance on questions that are related to the image.

2.2. Dataset

This paper uses 3000 images from the original validation split of the VQAv2 dataset [15]. This subset of 3000 images is further split into train, validation and test splits of a 1000 images each for all the experiments in this paper. Any reference to the train, validation or test splits in subsequent sections of this paper refer to the splits thus created, and not to the original train, validation and test splits of the VQAv2 dataset [15]. The VQAv2 dataset [15] is one of the standard benchmark datasets for the VQA task, and hence is chosen for the experiments in this paper.

The VQAv2 dataset [15] contains open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer [15]. The full dataset consists of 265,016 images from the COCO [21] and abstract scenes [33] datasets. Each image is annotated with at least 3 questions, with an average of 5.4 questions per image [15]. There are 10

ground truth answers per question, along with 3 plausible (but likely incorrect) answers per question [15].

2.3. Generating unrelated questions

In order to generate unrelated questions for each image in our train, validation and test sets (derived from the original VQAv2 [15] dataset as described in section 2.2 above), this paper randomly permutes all the questions associated with each image in the train, validation and test set. The image and question thus obtained after randomly permuting all questions are assumed to be 'unrelated'. For the first 500 examples in the train, validation and test set, the query question is the original question, while the query image is the original image. For the last 500 images in the train, validation and test set, the query image is the original image, and the query question is the randomly permuted question, which is assumed to be unrelated to the image.

In order to validate how reasonable the assumption that the image-question pair thus obtained are indeed unrelated to each other, 30 randomly sampled images are manually inspected, and the proportion of image-question pairs that are indeed unrelated to each other is calculated. The image-question pairs in the random sample are included in **??**.

Based on manual inspection, 83.33% of the questions in the random sample of size 30 were genuinely unrelated to each other. Thus, this paper considers it reasonable to assume that image-question pairs obtained after randomly permuting all questions in the train, validation and test splits are indeed unrelated to each other. However, this means that the evaluation procedure described in section 4 would underestimate a model's ability to be able to detect and appropriately respond to unrelated questions, thereby overestimating a model's tendency to hallucinate when probed with questions unrelated to the input image.

2.4. Approaches for detecting unrelated questions

Proposed approaches that give off-the-shelf pre-trained VQA models the ability to recognise unrelated questions and abstain from answering them are described below.

2.5. Random classifier

This baseline is evaluated in order to determine whether the other proposed approaches in this paper are indeed picking up on any signal that would allow them to meaningfully distinguish image-question pairs that are unrelated from image question pairs that are related. The baseline classifier is constructed as follows.

1. A random score between 0 and 1 inclusive is generated for each image-question pair
2. For each image-question pair in the validation set, a prediction is made and recorded as follows for varying threshold values.
   (a) If the random score is above the threshold, the image-question pair is predicted as related.
   (b) Otherwise, the image-question pair is predicted as unrelated.
3. The evaluation procedure described in section 4 is conducted using the predictions for the validation set at the different threshold values.
4. For each image-question pair in the test set, a prediction is made and recorded using the same procedure as for the pairs in the validation set as described above.
5. The evaluation procedure described in section 4 is conducted using the predictions for the test set at the different threshold values.

2.5.1. Threshold on maximum softmax score This maximum softmax score is hypothesized to capture the confidence of the model that the most probable answer according to the output of the model is indeed correct. This paper hypothesizes that for unrelated image-question pairs, the model would be likely to be less confident that its prediction is correct than for related image-question pairs, and thus have a lower maximum softmax output than for related image-question pairs. Based on this hypothesis, another strategy to exploit the softmax output of the pre-trained VQA model to detect unrelated image-question pairs is described below.

1. For each image-question pair in the validation set, the maximum softmax output of the pretrained VQA model is recorded.
2. For varying threshold values, the predictions on the validation set are obtained and recorded as follows.

   (a) If the maximum softmax output for the image-question pair is above the threshold, the image-question pair is predicted as related.
   (b) Otherwise, the image-question pair is predicted as unrelated.
3. The evaluation procedure described in section 4 is conducted using the predictions for the validation set at the different threshold values.
4. For each image-question pair in the test set, a prediction is made and recorded using the same procedure as for the pairs in the validation set as described above.
5. The evaluation procedure described in section 4 is conducted using the predictions for the test set at the different threshold values.

2.5.2. Linear classifier trained on maximum softmax score An alternative strategy to use the maximum softmax score of the pre-trained VQA model in order to detect unrelated questions is to add a linear classifier on top of a pre-trained VL model. This classifier is trained on the maximum softmax output of the pretrained VQA model as input, and the gold labels (1 for unrelated image-question pairs, 0 for unrelated image-question pairs). The procedure is described in more detail below.

1. For each image-question pair in the validation set, the maximum softmax output of the pretrained VQA model is recorded.
2. A logistic regression model is trained on the maximum softmax output of the pretrained model as found in the previous step for each image-question pair in the validation set as the input, and the target labels 1 for unrelated pairs, 0 for related pairs. The logistic regression model used in this paper is described in section 3
3. The trained logistic regression model is used to get predicted probabilities for the label 0 and 1 for all the image-question pairs in the test set.
4. For varying thresholds, the predictions on the test set are obtained ad recorded as follows.
   (a) If the predicted probability of the label 1 (corresponding to an unrelated image-question pair) for the image-question pair is above the threshold, the image-question pair is predicted as unrelated.

(b) Otherwise, the image-question pair is predicted as related.

5. The evaluation procedure described in section 4 is conducted using the predictions for the test set at the different threshold values.

2.5.3. Intermediate image captioning Changpinyo et al. [8] propose a method that automatically derives VQA examples at volume, by leveraging the abundance of existing image-caption annotations combined with neural models for textual question generation. This indicates that perhaps image captions may contain information about relevant questions that could possibly be asked about the image and thus help identify unrelated questions. The approach is described in detail below. This method is based on the hypothesis that if there exists an object in the question that is not in the image caption generated from the input image, then the image-question pair is likely to be unrelated.

1. For each image-question pair in the validation set, generate image captions for each image in the pair using a pre-trained image captioning model. The pre-trained image-captioning model is described in section 3.
2. Find all the objects in the image by using a pre-trained object detection model. The pre-trained object detection model used in this paper is described in section 3.
3. Get a prediction for each image-question pair in the validation set as follows.
   (a) For each word in the question, do the following to obtain the list of objects in the question.
   (b) Find all the words in the question that are nouns in the question using the following procedure.
      i. Use WordNet [13] to find all the synsets of the word. A synset is a set of synonyms that share a common meaning [13]. Each synset contains one or more lemmas, which represent a specific sense of a specific word [13].
      ii. if at least one of the lemmas in one of the synsets of the word has the part-of-speech (POS) tag indicating that the lemma is a noun, consider the word an noun. Otherwise, do not consider the word a noun
      iii. For each noun in the question, find the similarity of the noun in the question with each of the object categories in the list of object categories covered by the pre-trained object

detection model. If the similarity score is above a certain threshold, consider the noun an object. Otherwise, the noun is not an object.

A. In this paper, the pre-trained object detection model used, as described in section 3, is pre-trained on the COCO dataset [21]. Thus, the list of 80 object categories labelled in the COCO dataset [21] is used.

B. The appropriate threshold for the similarity score used in this step is determined by finding a value that works well and makes sense on the validation set empirically. A value of 0.6 is found to work best empirically.

C. The similarity of the noun with each object in the list is computed using the linear similarity metric. This metric gives a score representing how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node) and that of the two input Synsets. The relationship is given by the equation $\frac{2IC(lcs)}{IC(s1)+IC(s2)}$. IC gives the amount of information conveyed by a particular unit of language in a particular context

iv. if the noun is considered an object, consider all the objects from the list for which the similarity score is above the chosen threshold as objects in the question.

v. If there is at least one object in the question that is not in the image caption, consider the image-question pair as unrelated. Otherwise, consider the image-question pair as related.

(c) Use the validation set to determine all the hyperperameters of the system, such as the threshold for object similarity, etc.

4. using the best hyperperameters found over the validation set, get the predictions on the image-question pairs in the test set using the same procedure followed for the validation set.

2.5.4. Intermediate object detection . This paper hypothesizes that if there is an object in the question that is not in the image, then the image-question pair is likely to be unrelated. Thus, an approach based on this hypothesis is described below.

1. For each image-question pair in the validation set, get predictions for whether each image-question pair is related or unrelated using the following procedure.
   (a) Get a list of objects in the image using a pretrained image object detection model. This paper uses a pretrained model described in section 3.
      i. Find all the words in the question that are nouns in the question using the following procedure.
      ii. Use WordNet [13] to find all the synsets of the word. A synset is a set of synonyms that share a common meaning [13]. Each synset contains one or more lemmas, which represent a specific sense of a specific word [13].
      iii. if at least one of the lemmas in one of the synsets of the word has the part-of-speech (POS) tag indicating that the lemma is a noun, consider the word an noun. Otherwise, do not consider the word a noun
      iv. For each noun in the question, find the similarity of the noun in the question with each of the object categories in the list of object categories covered by the pre-trained object detection model. If the similarity score is above a certain threshold, consider the noun an object. Otherwise, the noun is not an object.
         A. In this paper, the pre-trained object detection model used, as described in section 3, is pre-trained on the COCO dataset [21]. Thus, the list of 80 object categories labelled in the COCO dataset [21] is used.
         B. The appropriate threshold for the similarity score used in this step is determined by finding a value that works well and makes sense on the validation set empirically. A value of 0.6 is found to work best empirically.
         C. The similarity of the noun with each object in the list is computed using the linear similarity metric. This metric gives a score representing how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer (most specific ancestor node) and that of the two input Synsets. The relationship is given by the equation $\frac{2IC(lcs)}{IC(s1)+IC(s2)}$. IC gives the amount of information conveyed

by a particular unit of language in a particular context

v. if the noun is considered an object, consider all the objects from the list for which the similarity score is above the chosen threshold as objects in the question.

vi. If there is at least one object in the question that is not in the image caption, consider the image-question pair as unrelated. Otherwise, consider the image-question pair as related.

(b) Use the validation set to determine all the hyperperameters of the system, such as the threshold for object similarity, etc.

(c) using the best hyperperameters found over the validation set, get the predictions on the image-question pairs in the test set using the same procedure followed for the validation set.

2.5.5. MLP classifier trained on all softmax output scores Another possible approach to detect unrelated questions is to train an MLP classifier on all the softmax outputs of the pretrained VQA model. This method is based on the hypothesis that perhaps the pattern and distribution of the softmax scores across several logits might give a better indication about whether the input image-question pair are indeed related or not. For example, more evenly distributed softmax values may indicate that the input image-question pair are unrelated, since this would indicate that the model is more confused about its prediction. Training on all the softmax outputs rather than just the maximum softmax output would give the model more useful information about the distribution of confidence over different logits rather than simply the confidence over the most likely answer according to the model. The procedure is described in detail below.

1. Run a pre-trained VQA model on the image-question pairs in the validation set, and record all the softmax outputs of the model over all the logits in the final output layer of the pretrained model. The pretrained model used for VQA is described in section 3
2. Train an MLP classifier to classify each image-question pair as related or unrelated based on the softmax outputs from the previous step as input, and the target labels 0 and 1 for related and unrelated questions respectively. The hyperperameters of the MLP classifier, including number of hidden layers and regularization weight α are determined using grid search. The best

hyperperameters found are described in section 5.

3. Use the trained MLP classifier to obtain the predicted probability that the image-question pair is unrelated for each image-question pair in the test set
4. get the predictions for varying thresholds for the predicted probability, and conduct the evaluation procedure described in section 4.1 to evaluate the performance of the model.

2.5.6. Text debias This method is based on the hypothesis that when presented with a question and a random image (i.e. an image with each pixel's RGB value is drawn from a uniform random distribution), a pretrained VQA model would give an output that is determined solely from the question inputted as text. Thus, for unrelated images, the model would also tend to give a response that is more likely to be determined from the text, since past work has shown that pretrained VQA models tend to make predictions biased towards the text input [7]. Thus, if the difference in the softmax output of the model given the random image and the question and the target image and question is small, the target image is more likely to be unrelated to the question, since the output is closer to the output the model would give if the model was presented with only the text question and a random image. Here, it is assumed that a random image would provide no signal to the VQA model. This procedure is described in more detail below.

1. For each image-question pair in the validation set, all the softmax output values of the pretrained VQA model is recorded. The softmax output of the question with a random image (where each pixel value is drawn from a random uniform distribution) is also recorded.
2. The difference in the softmax output of the target image and question as input and the softmax output of the random image and question is computed. This difference vector is sorted in descending order and denoted as d_i for each example i in the validation set.
3. Several models are trained using the distinct procedures described below.
   (a) MLP classifier on softmax difference: Trained on each d_i as input, and target label (related or unrelated) as the output. Best hyperperameters of the MLP model, including number of hidden layers and regularization weight are determined using grid search.

(b) Linear classifier on softmax difference: Trained on each d_i as input, and target label (related or unrelated) as the output.

(c) Linear classifier on the maximum softmax difference: Trained on the max value in each d_i as input and the target label (related or unrelated) as the output.

4. For each image-question pair in the test set, all the softmax output values of the pretrained VQA model is recorded. The softmax output of the question with a random image (where each pixel value is drawn from a random uniform distribution) is also recorded.
5. The difference in the softmax output of the target image and question as input and the softmax output of the random image and question is computed. This difference vector is sorted in descending order and denoted as f_i for each example i in the test set.
6. For each model trained above, probabilities for whether each image-question pair i in the test set is unrelated is obtained by running inference on each trained model and corresponding f_i as input for each pair i in the test set. The output probabilities for the appropriate label are then obtained.
7. For each trained model, the predictions are derived for varying thresholds as follows. For each threshold value, if the predicted probability that the image-question pair is unrelated is greater than the threshold value, then the pair is labelled as unrelated. Otherwise, the pair is predicted as unrelated.
8. the evaluation procedure described in section 4 is conducted to evaluate the performance of this approach.

2.5.7. Image Debias In order to compare the degree to which pre-trained VQA models are biased towards the image rather than the text when presented with an unrelated image question pair, this approach modifies the method described in section 2.5.6 to check whether the behavior of a pretrained VQA model when presented with a random image and question or meaningless text and original image would be similar to the behavior of the model when presented with a question and image unrelated to each other. Thus, if the difference in the softmax output of the model given the target image and a meaningless text input and the target image and target question is small, the

target image is more likely to be unrelated to the target question, since the output is closer to the output the model would give if the model was presented with only the text question and a random image, based on the hypothesis that the behavior of a pretrained VQA model when presented with a random image and question or meaningless text and original image would be similar to the behavior of the model when presented with a question and image unrelated to each other. Here, it is assumed that meaningless text would provide no signal to the VQA model, so the output would be solely determined by the image input. This procedure is described in more detail below.

1. For each image-question pair in the validation set, all the softmax output values of the pretrained VQA model is recorded. The softmax output of the question with the target image and meaningless text input "N/A" is also recorded.
2. the difference in the softmax output of the target image and question as input and the softmax output of the target image and meaningless text input is computed. This difference vector is sorted in descending order and denoted as d_i for each example i in the validation set.
3. Several models are trained using the distinct procedures described below.
   (a) MLP classifier on softmax difference: Trained on each d_i as input, and target label (related or unrelated) as the output. Best hyperperameters of the MLP model, including number of hidden layers and regularization weight are determined using grid search.
   (b) Linear classifier on softmax difference: Trained on each d_i as input, and target label (related or unrelated) as the output.
   (c) Linear classifier on the maximum softmax difference: Trained on the max value in each d_i as input and the target label (related or unrelated) as the output.
4. For each image-question pair in the test set, all the softmax output values of the pretrained VQA model is recorded. The softmax output of the meaningless text input "N/A" with the target image is also recorded.
5. The difference in the softmax output of the target image and question as input and the softmax output of the targte image and meaningless text input "N/A" is computed. This difference vector is sorted in descending order and denoted as f_i for each example i in the test set.

6. For each model trained above, probabilities for whether each image-question pair i in the test set is unrelated is obtained by running inference on each trained model and corresponding f_i as input for each pair i in the test set. The output probabilities for the appropriate label are then obtained.
7. For each trained model, the predictions are derived for varying thresholds as follows. For each threshold value, if the predicted probability that the image-question pair is unrelated is greater than the threshold value, then the pair is labelled as unrelated. Otherwise, the pair is predicted as unrelated.
8. The evaluation procedure described in section 4 is conducted to evaluate the performance of this approach.

2.5.8. ask-LLM It has been shown in prior work that large language models exhibit surprising zero-shot reasoning capabilities [23]. Thus, based on this insight, this approach relies on the hypothesis that a large language model could perhaps reason about whether a question is related to an image given a list of objects in the image as well as the queried question. This procedure is detailed below. This paper uses the BART model [19] pretrained on the MultiNLI (MNLI) dataset [30]. This model is described more in section 3

1. For each image-question pair in the validation set and test set, the following steps are conducted.
   (a) A caption for the image is found by using a pretrained image captioning model and running inference on the model with the image as the input. The image captioning model is described in section 3.
   (b) The following template is used to construct the input given to a large pretrained language model:

   "Image caption: image caption for the queried image

   Question: queried question

   Is the question related to the image?"
   (c) the prediction probabilities of the LLM are recorded.
2. Over a varying threshold, the predictions are computed as follows. For each threshold, the

prediction of each image question pair is determined as follows. If the probability that the model predicts that the image-question pair is unrelated is above the threshold, the pair is predicted as unrelated. Otherwise, the pair is predicted as unrelated.

3. The method is evaluated over the test set by following the evaluation procedure described in section 4.

3. Implementation

3.1. Off-the-shelf models

All off-the-shelf pretrained models used in this paper are used using the HuggingFace Models API [**?**].

3.1.1. Base VQA model The Vision-and-Language Transformer (ViLT) model [18] is used in this paper as the baseline VQA model. This model is finetuned on the VQAv2 dataset [15]. The pre-trained model checkpoint "dandelin/vilt-b32-finetuned-vqa" is loaded using the HuggingFace Models API [1].

3.1.2. Object detection model The You Only Look at One Sequence (YOLOS) model [12] is used in this paper as an off-the-shelf model for object detection. It was pre-trained on ImageNet-1k [26] and fine-tuned on COCO 2017 object detection [21], a dataset consisting of 118k/5k annotated images for training/validation respectively. The pre-trained model checkpoint "hustvl/yolos-tiny" is loaded using the HuggingFace Models API [2].

YOLOS is a Vision Transformer (ViT). Despite its simplicity, a base-sized YOLOS model is able to achieve similar performance on the COCO validation 2017 dataset [21] as more complex frameworks such as Faster R-CNN [24].

3.1.3. Image Captioning model A pre-trained BLIP model [20] is used as an off-the-shelf image captioning model, pre-trained on the COCO dataset [21].

3.2. Large language model (LLM)

For the ask-LLM approach described in section 2, this paper uses the BART-large model [19] trained on the Multi-NLI dataset [30]. MNLI, or the Multi-Genre Natural Language Inference is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information [30]. In particular, a method proposed by Yin et al. [32] is used to perform zero-shot classification by using pre-trained NLI (Natural Language Inference) models as a ready-made zero-shot sequence classifiers [32].

Their proposed method works by posing the sequence to be classified as the NLI premise and to construct a hypothesis from each candidate label [32]. For example, if one wants to evaluate whether a sequence belongs to the class "politics", one could construct a hypothesis of "This text is about politics..." [32]. The probabilities for entailment and contradiction are then converted to label probabilities [32].

This method is found to be surprisingly effective in many cases by Yin et al. [32], particularly when used with larger pre-trained models like BART [19]. Thus, this method proposed by Yin et al. [32] is used in this paper by calling the "zero-shot-classification" pipeline from HuggingFace and loading the "facebook/bart-large-mnli" checkpoint [**?**].

4. Evaluation

4.1. Evaluation metric

The CHAIR metric [6] to measure hallucination in image captioning did not generalize to other VL tasks. Moreover, it only worked for the image captioning task, and for datasets where the list of objects in the dataset was available. However, in real life, this approach would not generalize and would not scale to truly open-domain datasets with many different object classes.

Thus, this section proposes an evaluation procedure on the test set that can help analyze how well a model is able to correctly abstain from answering unrelated questions as well while also reasonably attempting to answer related questions. To develop the evaluation procedure, this section

further refines the definitions of risk and coverage in Whitehead et al. [29] to focus specifically on how well a model can detect questions unrelated to the image in order to more explicitly address the goals of this paper. The definitions of risk and coverage as used in subsequent sections of this paper are described below.

Risk: This is defined as the ratio between the number of unrelated questions the model attempts to answer (u_a) and the total number of unrelated questions (u_t) in the dataset being evaluated. This ratio is denoted as R in the rest of the paper

$$R = \frac{u_a}{u_t}$$

Considering the detection of unrelated questions as a binary classification task, with the positive class corresponding to unrelated questions, R can be rewritten as one minus the recall score of the positive class, as shown below. Note that $u_a = fn$ and $u_t = tp + fn$ where tp and fn refer to the number of true positives and false negatives respectively.

$$\begin{aligned} R &= \frac{fn}{tp+fn} \\ &= 1 - \frac{tp}{tp+fn} \\ &= 1 - recall_{positive} \end{aligned}$$

Coverage: This is defined as the ratio between the number of related questions the model attempts to answer (r_a) and the total number of related questions in the dataset being evaluated (r_t). This ratio is denoted as C in the rest of the paper.

$$C = \frac{r_a}{r_t}$$

Considering the detection of unrelated questions as a binary classification task, with the positive class corresponding to unrelated questions, coverage can be rewritten as the recall score of the

negative class (i.e. related questions), as shown below. Note that $r_a = tn$ and $r_t = tn + fp$ where tn and fp refer to the number of true negatives and false positives respectively.

$$C = \frac{tn}{tn + fp}$$

$$= recall_{negative}$$

An ideal model that is able to appropriately abstain from answering questions unrelated to the given image would exhibit higher coverage at lower levels of risk. However, typically, there is a trade-off between the coverage ratio offered by the model and the level of risk associated with the model, i.e. a model would only be able to achieve a lower risk ratio, (or equivalently, a higher $1 - R$ value) at the cost of a lower coverage ratio. The C and $1 - R$ values represent the per-class accuracies of the model.

Thus, this paper considers the area under the curve (AUC) of the graph of C versus $1 - R$ values as a metric to evaluate a model's ability to detect unrelated questions and respond to them appropriately. Looking at the AUC value allows us to compare each approach's risk-coverage trade-off without choosing any particular threshold or risk tolerance.

This metric is chosen in addition to looking at the AUC of the precision-recall graph since the C versus $1 - R$ graph offers a more intuitive visualization of the coverage-risk trade-off, effectively capturing how well a model trades off coverage in favor of lower risk. Looking at the coverage-risk trade-off of any approach makes more sense since the highest coverage model at a particular tolerable level of risk for any particular application can be easily compared across different models, and the best approach with the highest coverage for the tolerable risk can be chosen accordingly. For instance, high risk tasks would have lower tolerable values of R, where incorrect predictions would have severe consequences (such as VQA systems for blind users, medical diagnoses, etc.).

Proposed evaluation procedure:

- Step 1: For half the examples in the dataset, randomly permute all questions in the dataset and assign them to each image in that portion of the dataset. From the human study conducted, it is

reasonable to assume that the image-question pairs thus obtained are unrelated to each other. The other half of the dataset is kept as is, so the questions are assumed to be related to the images in this half of the dataset being evaluated.

- Step 2: For several risk thresholds, record the output of the VQA model on each image-question pair obtained after Step 1. For half the dataset, the image question pairs are assumed to be unrelated, and for the other half, they are assumed to be related.
- Step 3: Compute the coverage for each risk threshold used in step 2.
- Step 4: Plot the C vs $1-R$ curve, and compute the AUC. The higher this AUC is for a given model, the better the model is at identifying and appropriately responding to unrelated questions, indicating that the model is less prone to object hallucination when prompted with an unrelated question.

5. Results

5.1. Summary of quantitative results

Approach	$1-R$ vs C AUC	Precision-recall AUC
Random baseline	0.50	0.51
Softmax Thresholding	0.69	0.65
Linear Classifier	0.69	0.64
Multi-Layer Perceptron Classifier (MLP)	0.58	0.64
Text bias (MLP on softmax difference)	**0.70**	**0.67**
Text bias (linear classifier softmax difference)	0.63	0.58
Text bias (max softmax linear classifier)	0.67	0.64
Image Debias (Max softmax difference threshold)	0.67	0.63
Image bias (Max softmax MLP)	0.67	0.63
Image bias (Softmax MLP)	0.71	0.65
ask-LLM	0.58	0.58

Table 1: Summary of results using the evaluation metric described in section 4 for all the proposed approaches described in section 2

Table 1 gives a summary of the performance of each successful proposed approach for identifying unrelated questions described in section 2 using the evaluation procedure described in section 4. The best performing approach is the text debias approach using a MLP classifier trained on the

	precision	recall	f1-score	support
0	76	87	81	15
1	85	73	79	15
accuracy	81	80	80	30
macro avg	81	80	80	30

Table 2: Human is given the image captions and the question (a)

all the softmax differences as described in section 2. In particular, this approach attains a 40% improvement over the random baseline, indicating that this approach is somewhat successful in detecting unrelated questions and is able to successfully extract and exploit at least some information that would allow us to detect unrelated image question pairs, thereby achieving our aim. However, this approach is not significantly better than the naive linear classifier trained on the max softmax outputs, thereby indicating that even the best approach of this paper fails to extract significantly more meaningful extra information than the naive linear classifier. This may indicate the need for more sophisticated approaches involving a novel architecture or training method with perhaps better designed input image features to help detect unrelated questions. The following discussion attempts to analyse the outputs of selected models in more depth both quantitatively and qualitatively in order to understand the nature of the output and where these approaches succeed as well as fail in order to discover potential future avenues for further investigation.

The intermediate image captioning and intermediate object detection approaches were found to not be successful, i.e. they gave results worse than the baseline random classifier, indicating that they did not learn any useful information that encoded whether an image-question pair is indeed unrelated. Thus, they are not included in the table above.

The plots for the precision-recall curve and the $1-R$ vs C curve are given in Appendix **??**. A more detailed quantitative and qualitative analysis of a selection of both successful and failed approaches is provided in the discussion in the rest of this section, along with hypotheses for why these approaches failed and succeeded.

	precision	recall	f1-score	support
0	79	73	76	15
1	75	80	77	15
accuracy	77	77	77	30
macro avg	77	77	77	30

Table 3: Human is given the list of objects in the image and the question (b)

5.2. Human baseline studies

To motivate the intermediate image captioning and intermediate object detection approaches, this paper conducts a human baseline study over 30 random examples from the validation set to check if a human is able to identify when a question is unrelated to an image given just (a) the image captions and the question and (b) the list of objects in the image and the question. The results are given in table 2 and table 3. Note that for all subsequent analysis sections, the positive class, represented by label 1 refers to unrelated questions. It is found that a human can perform significantly better than the random classifier, indicating that it is reasonable to expect that both approach (a) and (b) provide enough information to be able to classify the image-question pairs as unrelated. The human performance is slightly better in approach (a) with image captions than approach (b) with the list of objects in the image, suggesting that the approach with image captions may be more promising for detecting unrelated questions, and we would expect a better performance with the image captioning based approach.

However, when the proposed approaches are indeed evaluated using the proposed evaluation procedure in section 4, they do not succeed, i.e. they achieve worse performance than the random classifier. In particular, for approach (a), the AUC of the $1 - R$ vs C curve is 0.56, while the same value for approach (b) is 0.42. Thus, corroborating with the inference from this paper's human study above, the image captioning based approach performs better than the object detection based approach. However, both approaches are still significantly worse than the naive linear classifier trained on max softmax output, indicating that they are not much more successful.

This worse performance may be due to the fact that not all objects in the image caption may indeed be found in the image, and there may be inaccuracies in matching the nouns in the caption

as found using this paper's proposed approach to objects in the COCO dataset list of 80 objects [21]. Moreover, not all objects in the question may necessarily appear in the image even for related questions.

Yet, the fact that the image captioning model performs better than the random classifier may indicate that the model does indeed learn some useful information from the image caption about whether the image-question pair is indeed unrelated. Image features based on output from the image captioning model may be promising to study and would be a promising direction for future work.

5.3. Analysis of naive approach of training a linear classifier on the maximum softmax output

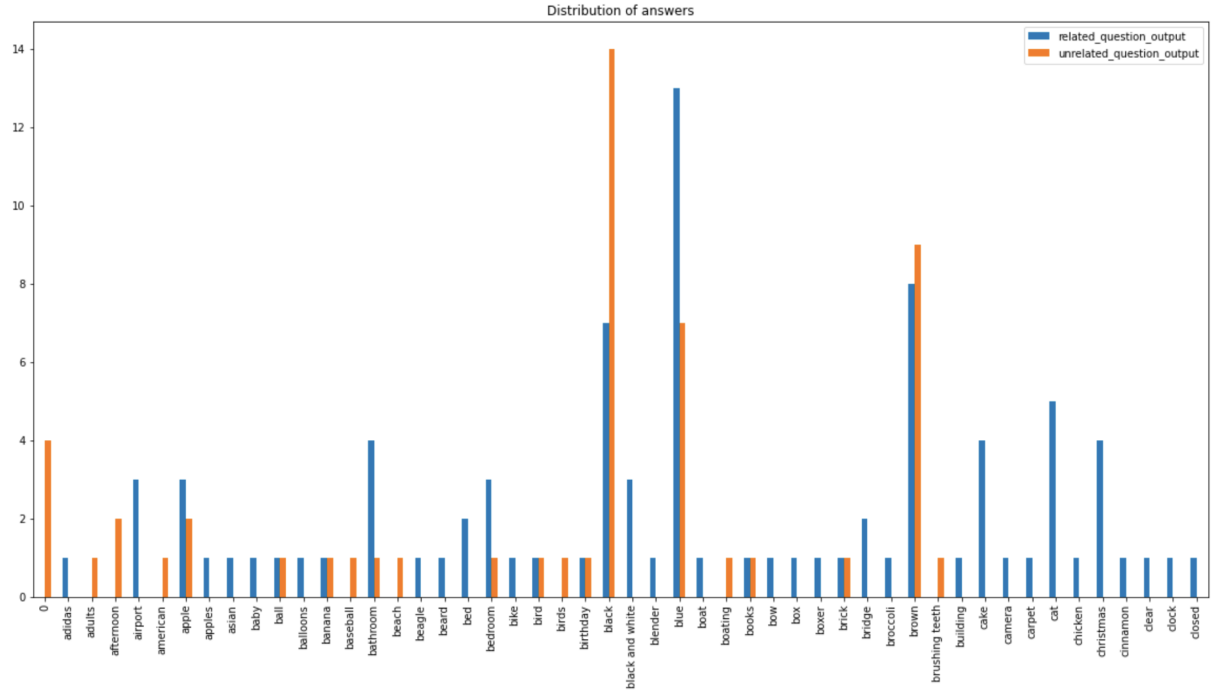

Figure 2: Distribution of answers for related and unrelated image-question pairs in the test set for the linear classifier trained on maximum softmax values

The linear classifier trained on the maximum softmax output achieves a performance significantly better than the random baseline, indicating that the model is indeed learning some useful information that meaningfully correlates with unrelated image-question pairs.

The figure 2 shows the distribution of output for both related and unrelated questions on the test set. In general, it seems that for a large proportion of the unrelated pairs, the most common outputs

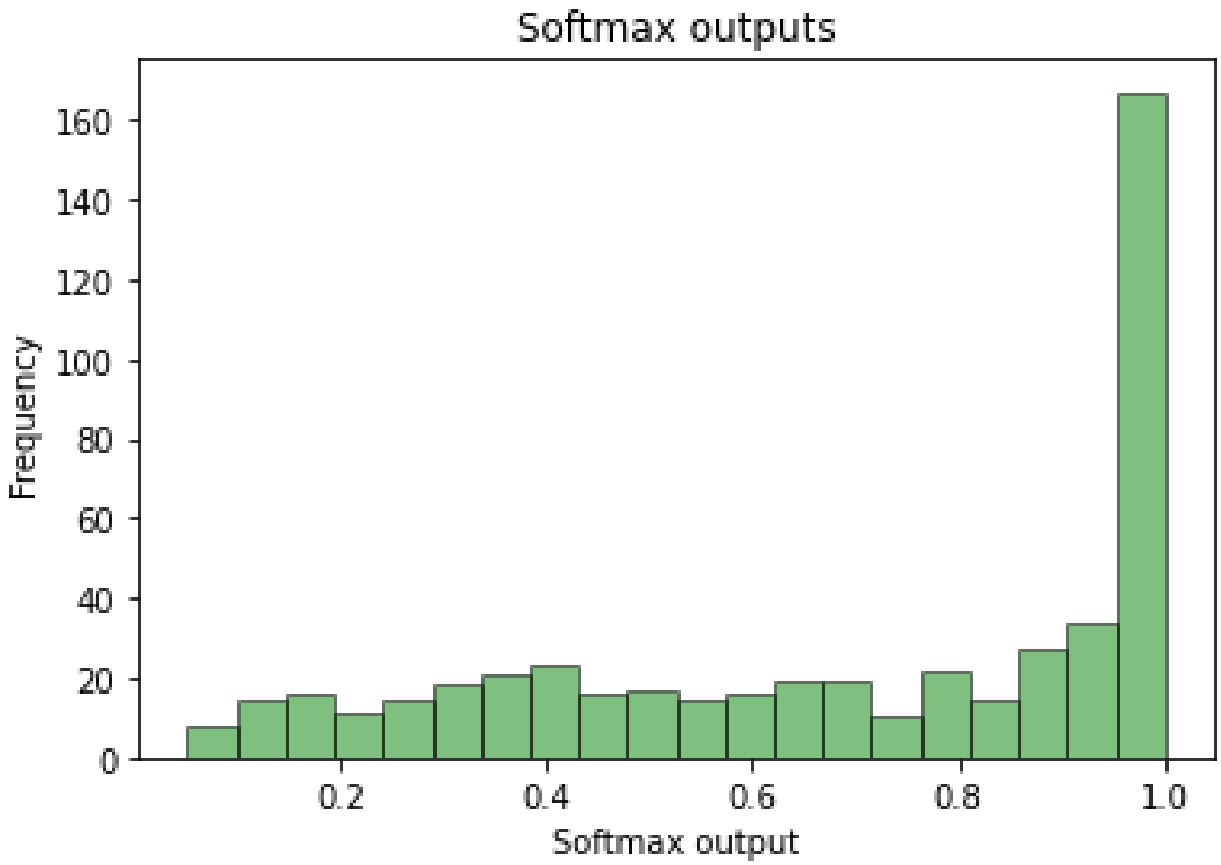

Figure 3: Max softmax values distribution for related image-question pairs

indeed correspond with the outputs for the related pairs. This indicates that perhaps the VQA model is biased towards certain outputs or tends to revert to outputting the most common output answer of "black and white", "black", "brown", and other colors when presented with an unrelated image. Such predictions are likely to correspond to object hallucination - it seems that these colours are probably very common in the dataset so the model might learn to just detect these colours in the image and output the result. Alternatively, the model may be looking at objects in the question, and simply outputting the color of those objects when confused about how to answer the question if it is unrelated to the image.

5.4. Analysis of softmax outputs of pretrained VQA model

The figures 3 and 4 show the distribution of softmax outputs for related and unrelated image-question pairs. The softmax scores for related questions seem to be significantly more concentrated towards

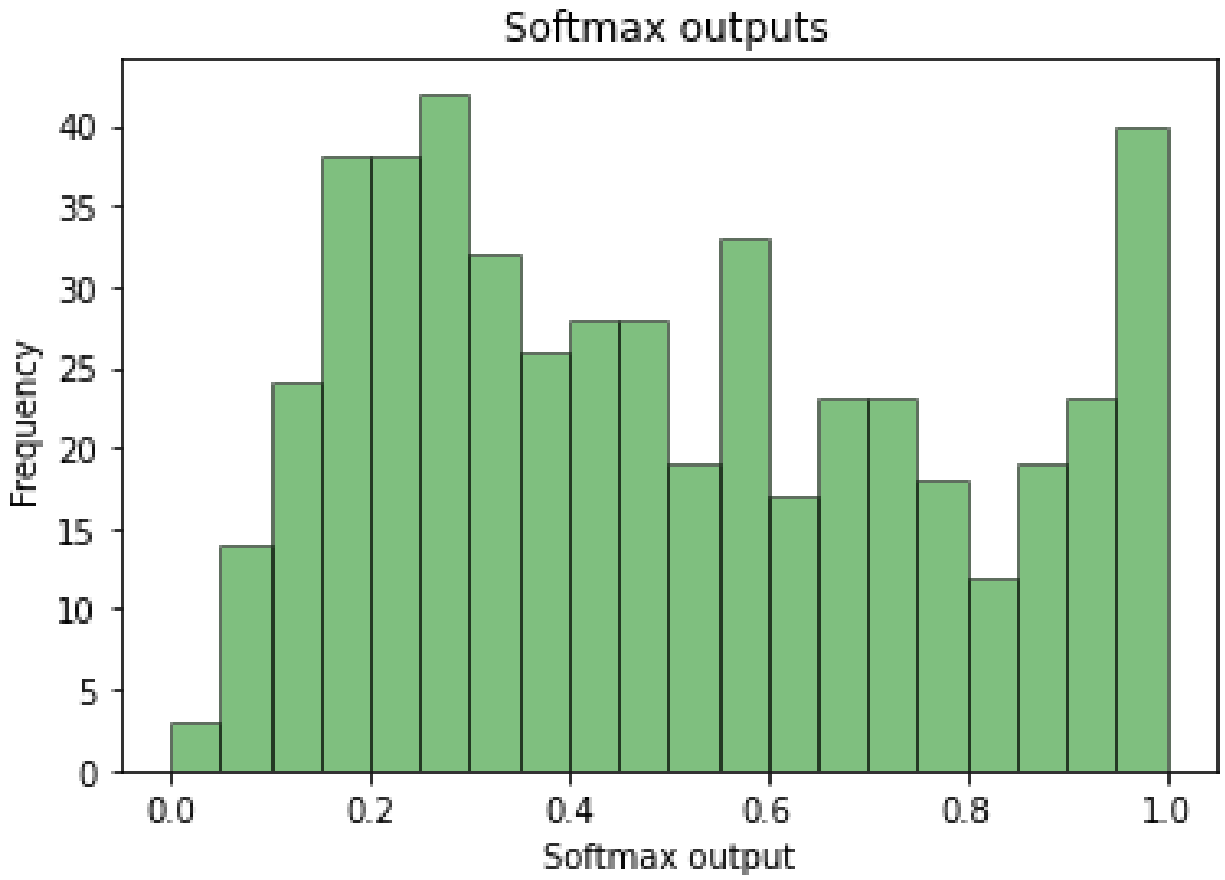

Figure 4: Max softmax values distribution for unrelated image-question pairs

higher values close to 1, while the softmax scores for unrelated pairs seem to be a lot more spread out. This indicates that there may be at least some signal that would allow us to detect unrelated questions from the softmax output of a pretrained VQA model. However, a significant proportion - around 8% - of unrelated pairs also have max softmax scores close to 1, indicating that the model is surprisingly very confident on its prediction for unrelated pairs, which is counter intuitive and unexpected. However, this may be due to the nature of how we obtain our unrelated pairs - not all pairs are guaranteed to be related based on the human study described in a previous section, which showed that only around 83% of the randomly permuted questions were indeed unrelated to the image (based on 30 random samples).

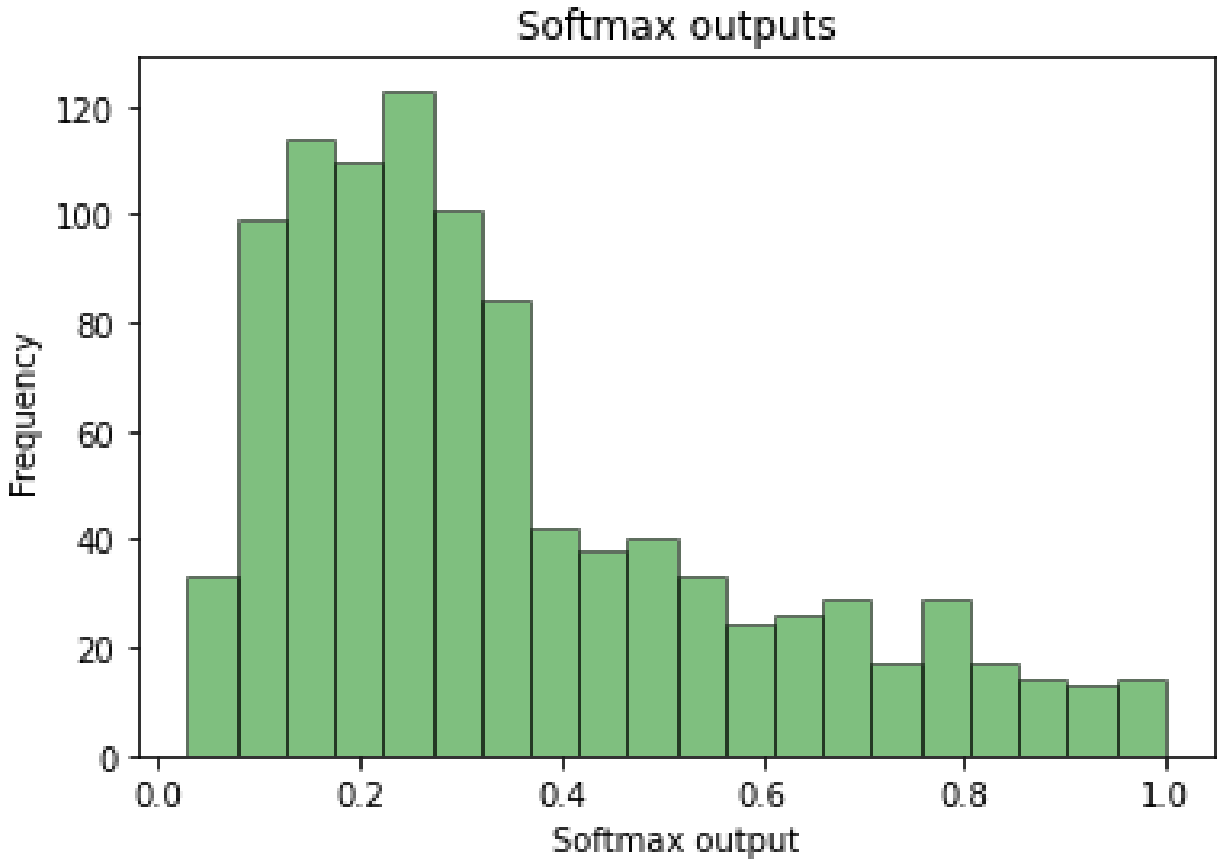

Figure 5: Max softmax values distribution for questions inputted with a random image

5.5. Analysis of output of text-only from text debias approach

It is found that the VQA model gives the exact same text output for 70.2% of the test set with the original image and question as with the random image and question. This indicates that for a majoring of the image-question pairs, the model gives the response solely based on the text rather than learning to look at the image in order to answer the question. This may be plausible and a reasonable response in many cases - the question narrows down the possible set of valid responses. However, this may also indicate to the model that the text is often enough to perform the VQA task, causing the model to be biased towards the text and thereby contributing to object hallucination when presented with an unrelated image. Thus, training based interventions that force the model to account for textual biases may be crucial for mitigating object hallucination, especially those of type 2.

Figure 5 shows the distribution of max softmax values for questions inputted with a random image. The distribution has a mode closer to 0.2, as opposed to being concentrated at a high value close to 1 for related image question pairs as seen in figure 3. Thus, the text debias approach proposed seems to be a promising approach for detecting unrelated questions. However, many unrelated image question pairs, as seen in figure 4 have high softmax values, which would not be recognised my this approach, explaining the types of cases where this approach might fail.

5.6. Comparing the linear classifier trained on max softmax to the best text debias approach with an MLP trained on differences between the text-only model softmax output and the softmax output with the image and text

Both the naive linear classifier trained on the max softmax output (referred to as model c) and the text debias approach achieve similar performance based on the evaluation procedure described in section 4 (referred to as model d). In order to qualitatively assess the cases where one model performs better than the other and the failure modes of both models, this section looks at the top 5 most confident mistakes made by both models (with confidence determined gy the max softmax score value) as well as the top 5 most confident mistakes made by model c that were not made by model d and vice versa. These are images and their associated questions are included in Appendix A.

From the top five most confident mistakes made by model d, it seems that the model makes mistakes on relatively harder examples. It falsely predicts the image pairs as unrelated in images with a lot of clutter, such as in 7. Moreover, it seems to miss detecting unrelated pairs where a human could be interpreted to be as an animal, such as in 6. The linear model, in contrast, is able to discern between a human an animal, as in 6. However, it is arguable that a human is indeed a type of animal, making the question in 6 indeed related to the image. The linear model seems to correctly predict related pairs as related when the objects referred to in the question are smaller, or if there is a lot of clutter in the image, as in 7. The text debias model also seems to often miss references from the background, such as references to the floor in the input question. However, the

linear model gets some of the more easier examples wrong, such as 11, where there clearly isn't any person in the image, while the question is very explicitly referring to person, providing a clear signal that the image-question pair is unrelated. The linear model also falsely predicts image-question pairs as unrelated for questions that are harder to answer due to poor lighting in the image, such as 12, which would thus naturally have lower max softmax values and thus be detected as unrelated despite actually being related.

6. Conclusion and future work

This paper first proposes a procedure to evaluate the prevalence of the type of object bias elicited when the model is probed with an image and an unrelated question, along with possible approaches to detect unrelated image-question pairs, thereby mitigating a major type of object hallucination. Using this proposed evaluation procedure several proposed approaches of detecting unrelated questions are evaluated, along with a thorough analysis of the outputs of these approaches as well as hypotheses for why they fail on certain types of examples, as well as types of examples where they succeed. The analysis in this paper suggests that softmax outputs of a pretrained VQA model provide some signal about whether a question is unrelated to an image. However, the best approach in this paper - the text debias model - does not perform significantly better than the naive linear classifier trained on the maximum softmax outputs of the pretrained VQA model. This may indicate that this paper has reached the limit on the information that can be extracted from the signal provided by the softmax output about whether an image-question pair is unrelated, and more sophisticated approaches involving novel training methods, loss functions as well as better image features and novel architectures may be necessary to identify unrelated image-question pairs. Investigating such approaches seem to be a viable area of future work and are crucial to explore in order to investigate how we can mitigate object bias to make VL models more reliable and trustworthy.

References

[1] "dandelin/vilt-b32-finetuned-vqa · Hugging Face — huggingface.co," https://huggingface.co/dandelin/vilt-b32-finetuned-vqa, [Accessed 01-May-2023].

[2] "Hustvl/yolos-tiny." [Online]. Available: https://huggingface.co/hustvl/yolos-tiny

[3] A. Agrawal *et al.*, "Don't just assume; look and answer: Overcoming priors for visual question answering," 2017. Available: https://arxiv.org/abs/1712.00377

[4] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," 2022. Available: https://arxiv.org/abs/2204.14198

[5] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," 2017. Available: https://arxiv.org/abs/1707.07998

[6] A. F. Biten, L. G. i Bigorda, and D. Karatzas, "Let there be a clock on the beach: Reducing object hallucination in image captioning," *CoRR*, vol. abs/2110.01705, 2021. Available: https://arxiv.org/abs/2110.01705

[7] R. Cadene *et al.*, "Rubi: Reducing unimodal biases in visual question answering," 2019. Available: https://arxiv.org/abs/1906.10169

[8] S. Changpinyo *et al.*, "All you may need for VQA are image captions," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1947–1963. Available: https://aclanthology.org/2022.naacl-main.142

[9] H. Choung, P. David, and A. Ross, "Trust in ai and its role in the acceptance of ai technologies," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–13, 2022. Available: https://doi.org/10.1080/10447318.2022.2050543

[10] W. Dai *et al.*, "Plausible may not be faithful: Probing object hallucination in vision-language pre-training," 2022. Available: https://arxiv.org/abs/2210.07688

[11] E. Davis, "Unanswerable questions about images and texts," *Frontiers in Artificial Intelligence*, vol. 3, jul 2020. Available: https://doi.org/10.3389%2Ffrai.2020.00051

[12] Y. Fang *et al.*, "You only look at one sequence: Rethinking transformer in vision through object detection," 2021.

[13] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. Available: https://mitpress.mit.edu/9780262561167/

[14] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, nov 2020. Available: https://doi.org/10.1038%2Fs42256-020-00257-z

[15] Y. Goyal *et al.*, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," 2016. Available: https://arxiv.org/abs/1612.00837

[16] D. Gurari *et al.*, "Vizwiz grand challenge: Answering visual questions from blind people," 2018. Available: https://arxiv.org/abs/1802.08218

[17] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, nov 2022. Available: https://doi.org/10.1145%2F3571730

[18] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," 2021. Available: https://arxiv.org/abs/2102.03334

[19] M. Lewis *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019.

[20] J. Li *et al.*, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.

[21] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," 2015.

[22] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 26–36. Available: https://aclanthology.org/W19-1803

[23] A. Radford *et al.*, "Language models are unsupervised multitask learners," 2019.

[24] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2016.

[25] A. Rohrbach *et al.*, "Object hallucination in image captioning," 2018. Available: https://arxiv.org/abs/1809.02156

[26] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[27] M. Stefanini *et al.*, "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain," in *International Conference on Image Analysis and Processing*, 2019.

[28] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," *CoRR*, vol. abs/1609.05600, 2016. Available: http://arxiv.org/abs/1609.05600

[29] S. Whitehead *et al.*, "Reliable visual question answering: Abstain rather than answer incorrectly," 2022. Available: https://arxiv.org/abs/2204.13631

[30] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1112–1122. Available: http://aclweb.org/anthology/N18-1101

[31] Y. Xiao and W. Y. Wang, "On hallucination and predictive uncertainty in conditional language generation," 2021. Available: https://arxiv.org/abs/2103.15025

[32] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," 2019.

[33] C. L. Zitnick, R. Vedantam, and D. Parikh, "Adopting abstract images for semantic scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 627–638, 2016.

A. Qualitative comparison of most confident mistakes of text debias approach and linear model

Figure 6: Man playing tennis.

Question: What kind of fuel does this train use?
texttual debias model pred: False
Probas: 0.3971008344892008

Figure 7: A ship docked in a city.

Figure 8: Boy with a skateboard.

Figure 9: People eating doughnuts.

Figure 10: Dogs sleeping on a couch in a living room.

Question: If she lets go, will what she is holding fly away?
linear model pred: 0
Probas: 0.3869563562077811

Figure 11: A bathroom with a bathtub.

Figure 12: A girl eating food on a plate.