Efficient Counterfactual Debiasing for Visual Question Answering

Camila Kolling*, Martin More†,§, Nathan Gavenski†, Eduardo Pooch†, Otávio Parraga†,§, Rodrigo C. Barros†,§
†MALTA Research Group, PUCRS, Brazil   § Teia Labs, Brazil   *Kunumi, Brazil
rodrigo.barros@pucrs.br

Abstract

Despite the success of neural architectures for Visual Question Answering (VQA), several recent studies have shown that VQA models are mostly driven by superficial correlations that are learned by exploiting undesired priors within training datasets. They often lack sufficient image grounding or tend to overly-rely on textual information, failing to capture knowledge from the images. This affects their generalization to test sets with slight changes in the distribution of facts. To address such an issue, some bias mitigation methods have relied on new training procedures that are capable of synthesizing counterfactual samples by masking critical objects within the images, and words within the questions, while also changing the corresponding ground truth. We propose a novel model-agnostic counterfactual training procedure, namely Efficient Counterfactual Debiasing (`ECD`), in which we introduce a new negative answer-assignment mechanism that exploits the probability distribution of the answers based on their frequencies, as well as an improved counterfactual sample synthesizer. Our experiments demonstrate that `ECD` is a simple, computationally-efficient counterfactual sample-synthesizer training procedure that establishes itself as the new state of the art for unbiased VQA.

1. Introduction

Over the past few years, the task of Visual Question Answering (VQA) [5] — answering a natural language question regarding the visual content of a given image — has attracted increasing attention of the scientific community. VQA provides interaction between image and text, being suitable for several real-world applications, such as chatbots for assisting visually-impaired people [12]. In those applications, we expect a model to answer truthfully based on the visual evidence contained in the image and the correct intention of the question. Unfortunately, this is not always the case even for state-of-the-art methods [2]. Instead of exploiting the image to find the correct answer, most models frequently rely on spurious correlations and follow the bias that naturally exists within the training data. Exploiting these "shortcuts" severely limits the generalization of VQA models in real-world scenarios, where the test distribution of facts (*e.g.*, colors, counts, objects, etc.) is often different from the training distribution [33].

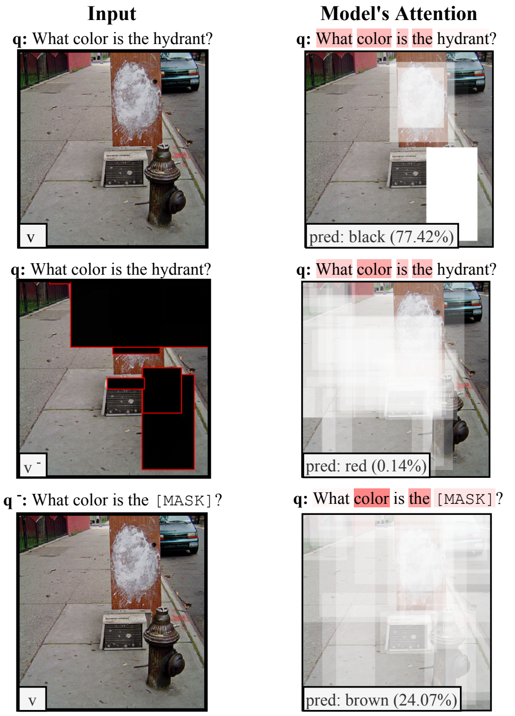

Input
Model's Attention
q: What color is the hydrant?
q: What color is the hydrant?
v
pred: black (77.42%)
q: What color is the hydrant?
q: What color is the hydrant?
v⁻
pred: red (0.14%)
q⁻: What color is the [MASK]?
q: What color is the [MASK]?
v
pred: brown (24.07%)

Figure 1: Example of our counterfactual sample synthesizer, which masks image regions and words according to their relevance to answer the question. (q, v) represents the original data, while (q, v^-) and (q^-, v) denote counterfactuals with masked image and question, respectively.

The *language bias* in VQA can be interpreted in two different ways [20]: (a) as the so-called *language prior* [3], where a strong correlation between questions and answers exists; and (b) the *visual priming bias* [15], in which the model assumes that the questioner is going to ask questions regarding objects present in the image. In both cases, VQA models may merely focus on the question rather than relying on the available visual content, as it should.

Recently, in order to evaluate the progress of VQA research towards mitigating biases, Agarwal *et al.* [2] proposed the VQA-CP (VQA under Changing Priors) benchmark. This benchmark has different question-answer distributions in the training and test splits, allowing to properly identify language biases that the models may have learned. The performance of many state-of-the-art VQA models [32, 13] drops significantly on VQA-CP compared to other datasets, confirming the existing biases.

The first prevailing solutions for avoiding biases in VQA were ensemble-based methods, *i.e.*, the introduction of an auxiliary question-only model to regularize the training of the targeted VQA model. However, as noted by Chen *et al.* [10], those approaches fail to embody two important characteristics: (a) visual explainability, *i.e.*, using the correct visual regions when making their decision; and (b) question sensitiveness, *i.e.*, perceiving linguistic variations within the questions. Thus, Chen *et al.* [10] proposed a different training procedure that synthesizes counterfactual image and question samples by masking critical information in the original data. Here "critical" means the most important cues to answer a specific question.

Despite significantly reducing biases, their counterfactual bias mitigation approach significantly increases the computational cost during training, demanding multiple forward and backward passes through the network to compute elements required for optimization. Those extra passes translate to an increased number of operations, memory consumption, and total training time. Motivated by these limitations, though aware of the benefits of designing a counterfactual synthesizing approach, we propose a novel model-agnostic counterfactual training procedure, namely Efficient Counterfactual Debiasing (ECD). Specifically, our contributions are as follows:

- We propose a new counterfactual sample synthesizer, which masks the most relevant image region/question words (see Fig. 1). It forces the model to use the most informative data to properly answer the questions.

- We introduce a novel negative answer-assignment mechanism for providing the answers to the counterfactual samples synthesized by our method. It exploits the probability distribution of the answers based on their frequency in the original training set.

- During training, our approach significantly reduces the computational cost, memory consumption, and optimization time when compared to the state-of-the-art counterfactual debiasing procedure [10].

- We outperform all state-of-the-art methods in the main benchmark for evaluating unbiased VQA models, namely VQA-CP $v2$ [2].

2. Related Work

Due to the specificities of the collection process, real-world datasets usually contain some form of bias [29, 9]. As a result, machine learning models tend to reflect those biases as they are correlation machines capable of exploiting superficial correlations between the input data and the ground-truth annotation [1]. Some methods were developed to identify and mitigate specific types of biases; for instance, there are methods focused on visual recognition biases [30], while others focus on gender biases [7]. In multimodal tasks, *i.e.*, tasks that combine language and visual information, several studies evaluate unimodal baselines [27] or rely on external knowledge to address biases [23]. Despite being a multimodal task, several studies have shown the existence of a predominant language bias in VQA models [33, 22, 15, 3]. Three main solutions are currently used to reduce language biases in VQA:

1. Balancing Datasets: the straightforward solution, though a bit cumbersome and time-consuming, is to balance datasets [33]. A well-known approach was introduced by Goyal *et al.* [15], which collected real images and different types of questions to create the VQA 2.0 dataset. Although this strategy has reduced some forms of biases, models can still exploit language priors in the form of question-answer distributions. As shown in the VQA-CP benchmark [3], the performance of several models significantly drops when tested on these datasets with balanced distributions.

2. Building Models: another common solution is to design specific models for mitigating biases. So far, most "debiasing" models for VQA are ensemble-based methods [22, 9], which introduce an auxiliary question-only model to regularize the training of the VQA model. This approach usually requires training multiple sub-models separately.

3. Changing Training Procedure: one can also mitigate language priors in VQA by changing the training scheme. Gat *et al.* [14] proposed a novel regularization term. However, the most effective method so far in this category is CSS-VQA (Counterfactual Samples Synthesizing) [10]. It is a training procedure that masks critical objects in the original image or words in the question, forcing the model to focus on important objects and words for answering the question, since they are penalized for "guessing" (answering using other factors, such as question-answer distributions or irrelevant correlated features). The VQA model is trained with these counterfactual samples, as well as the original data, and does not require additional data annotation.

Our novel approach is part of the *changing training procedure* strategy, relying on counterfactuals synthesizing for addressing language biases. However, our method efficiently synthesizes counterfactuals, substantially reducing computational cost, memory consumption, and total training time while outperforming the state-of-the-art methods for unbiased VQA in terms of accuracy.

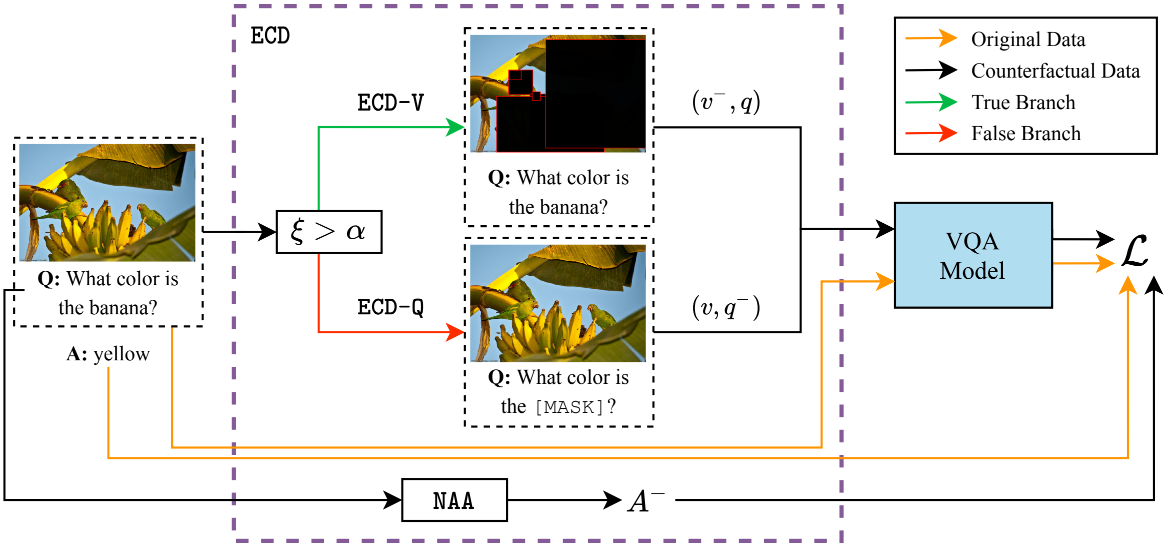


Figure 2: Architecture of ECD. Its core components are the counterfactual synthesizer, controlled by hyperparameter α, and the *Negative Answer Assignment* (NAA) module. ECD is agnostic to the VQA model architecture and loss function(s).

3. Efficient Counterfactual Debiasing

The VQA task can be formulated as a multi-class classification problem with *softscores* as targets. Given a dataset $\mathcal{D}$ consisting of N triplets $(i_j, q_j, a_j)_{j\in[1,N]}$ with $i_j \in \mathcal{I}$ an image, $q_j \in \mathcal{Q}$ a question in natural language, and $a_j \in \mathcal{A}$ an answer, one must optimize parameters θ of function $f_\theta : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ to produce accurate predictions. The typical learning strategy of VQA consists of minimizing the standard binary cross-entropy loss $\mathcal{L}$ over dataset $\mathcal{D}$ to provide the correct final answer to the question.

This section introduces our novel debiasing method for VQA, namely Efficient Counterfactual Debiasing (ECD). We first present the visual features used in our experiments in Sec. 3.1. Note, however, that our method is model agnostic, so one could use any other set of features (*e.g.*, [18, 25]). Then, we introduce the training procedure for debiasing the models in Sec. 3.2, detailing the counterfactual sample generation as well as the creation of the labels for the counterfactuals via a Negative Answer Assignment (NAA) module.

3.1. Visual Features

Several previous VQA studies [13, 22] make use of a trainable top-down attention mechanism over convolutional features to recognize relevant image regions. Anderson *et al.* [4] introduced complementary bottom-up attention that first detects common objects and attributes so that the top-down attention can directly model the contribution of those higher-level concepts. This so-called UpDn approach is often used in recent work [6, 18, 8, 9, 10] and significantly improves the performance of VQA models.

Specifically, for a single given image i_j with n_v detected objects, there is an image encoder $e_v : \mathcal{I} \rightarrow \mathcal{V} \in \mathbb{R}^{n_v \times d_v}$ to output a set of object features $\mathcal{V} = \{v_1, ..., v_{n_v}\}$, where v_j is a feature vector of dimension d_v for the j^{th} detected object. For each question q_j with n_t words, UpDn uses a question encoder $e_t : \mathcal{Q} \rightarrow \mathcal{T} \in \mathbb{R}^{n_t \times d_t}$ to output a set of word features $\mathcal{T} = \{w_1, ..., w_{n_t}\}$, where w_j is a feature vector of dimension d_t for the j^{th} word. Finally, UpDn has a multimodal fusion mechanism $m : \mathcal{V} \times \mathcal{T} \rightarrow \mathbb{R}^{d_m}$, and a classifier $c : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{|\mathcal{A}|}$. These functions are usually combined as $f(i_j, q_j) = c(m(e_v(i_j), e_t(q_j)))$. For the sake of clarity, we omit subscript j in the following sections.

3.2. Training Procedure

The overall structure of ECD is illustrated in Fig. 2. Our method consists of three main steps: (a) selecting relevant image regions or question words using either ECD-V (Sec. 3.2.1) or ECD-Q (Sec. 3.2.2); (b) synthesizing counterfactual samples via the Negative Answer Assignment procedure (Sec. 3.2.3); and (c) training the VQA model with both original and counterfactual samples.

Algorithm 1 describes the counterfactual synthesizing procedure in detail. ECD receives as input a set of visual features $\mathcal{V}$, questions $\mathcal{Q}$ and its corresponding answers $\mathcal{A}$, trade-off weight α, and top-N answers β. Hyperparameter α controls whether we synthesize counterfactual samples masking visual or textual information. Hyperparameter β defines the top-N most frequent answers that are removed from the ground truth and will not be assigned to the counterfactual samples. We sample ξ from a uniform distribution and test whether $\xi > \alpha$; if so, we execute module

Algorithm 1: Efficient Counterfactual Debiasing

```
Function ECD ($\mathcal{V}, \mathcal{Q}, \mathcal{A}, \alpha, \beta$):
    $\xi \sim U[0,1]$
    if $\xi > \alpha$ then                                  ▷ ECD-V
        $\mathcal{V}^- \leftarrow$ OBJ_SEL($\mathcal{V}$)
        $\mathcal{V} \leftarrow \{\mathcal{V} \cup \mathcal{V}^-\}$
    else                                                    ▷ ECD-Q
        $\mathcal{Q}^- \leftarrow$ WORD_SEL($\mathcal{Q}$)
        $\mathcal{Q} \leftarrow \{\mathcal{Q} \cup \mathcal{Q}^-\}$
    end
    $\mathcal{A}^- \leftarrow$ NAA($\mathcal{Q}, \beta$)
    $\mathcal{A} \leftarrow \{\mathcal{A} \cup \mathcal{A}^-\}$
    return $\mathcal{V}, \mathcal{Q}, \mathcal{A}$
end
```

ECD-V, which applies a visual mask to the critical visual features v^- of each instance; otherwise, we execute module ECD-Q, which applies a textual mask to the critical words in the question q^-. Then, we concatenate the counterfactual instance with its original version, and we do this for the entire batch of instances. Since we lack ground-truth annotations of answers a^- for these counterfactual samples, we generate them using NAA. The original and counterfactual samples are returned by ECD and forwarded to training the VQA model at hand. We compute the loss function and update the model's parameters, (with any desired optimizer) and ECD is called once again for the next data batch.

3.2.1 **ECD-V**

The ECD-V module synthesizes a counterfactual image by masking critical objects in the original image. For such, we first need to select critical objects present in the visual cue to answer the question. The extraction of objects highly related with a question and answer (QA) pair is performed as follows. First, we assign the part-of-speech tags to each word in the QA using the *spaCy part-of-speech* tagger [17] and extract the nouns of the QA. Then, we measure the cosine similarity between the GloVe embeddings [21] of the object categories (extracted from the detected objects) and the extracted nouns. We select the h objects with the highest similarity scores as $\mathcal{V}^+$, following the work of Wu *et al.* [31], which made use of the critical regions in their loss function to prevent the network from focusing on them when the model prediction is wrong. Here, however, we use $\mathcal{V}^+$ as a means of building the counterfactual visual input $\mathcal{V}^-$ as the complement of set $\mathcal{V}^+$ [10]. We show an example of elements in $\mathcal{V}$ and $\mathcal{V}^-$ in Fig. 1.

3.2.2 **ECD-Q**

The ECD-Q module synthesizes a counterfactual question by masking critical words present in the original question. It employs a *word selection* function, which first extracts question-type words (*e.g.*, "What color" in Fig. 1) for each question q and assign them as non-critical. Then, it selects all the remaining words that are not classified as stop words (*e.g.*, "is", "the") as critical. The counterfactual question q^- is built by replacing all the critical words in q with a special token [MASK]. We assign this final masked question to the question encoder e_t in order to extract t^-. We show an example of q^- and q in Fig. 1, in which q is "What color is the banana?" and q^- is "What color is the [MASK]?".

3.2.3 Negative Answer Assignment

To assign ground-truth answers for counterfactual pairs, we design the NAA mechanism (Fig. 3). We extract the negative answers by exploring *softscores* [10, 28, 26], which represent the reliability of ground-truth answers. An answer is considered reliable when at least 3 out of 10 human respondents have provided the given answer [5]. Specifically, we extract *softscores* by analyzing the number of occurrences of each answer in the 10 ground-truth answers annotated per question type. We normalize the number of occurrences of an answer for a specific question by the total number of occurrences of that question type in the training set. We use this per-question-type probability distribution as the intrinsic bias from the original source to improve fairness and model generalization, *i.e.*, the most biased answers are the ones that occur more frequently. This process is simple and can be performed once with negligible computational cost before the training procedure begins. We select the top-N answers (hyperparameter β) with the highest predicted probabilities as $\mathcal{A}^+$. We then define the ground-truth answers as $\mathcal{GT}$ and the negative answers $\mathcal{A}^-$ as all answers of $\mathcal{GT}$ but those in $\mathcal{A}^+$, *i.e.*, $\mathcal{A}^- := \{a | a \in \mathcal{GT} \wedge a \notin \mathcal{A}^+\}$.

The model does not have enough information to predict the correct answer for the counterfactual sample since the most relevant image regions or question words are masked. For instance, if the image contains object and the question is "What color is the object", the model could asso-

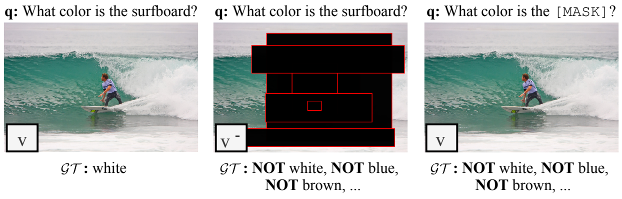

q: What color is the surfboard?
q: What color is the surfboard?
q: What color is the [MASK]?
V
V⁻
V
$\mathcal{GT}$: white
$\mathcal{GT}$: NOT white, NOT blue, NOT brown, ...
$\mathcal{GT}$: NOT white, NOT blue, NOT brown, ...

Figure 3: Example of the Negative Answer Assignment procedure in action, which generates the correct labels for the counterfactual samples by removing from the $\mathcal{GT}$ the top-N answers with the highest frequencies in the dataset, *i.e.*, the most biased answers for that particular question type.

Model	Venue	VQA-CP $v2$ test (%)				VQA 2.0 val (%)			
		Overall	Yes/No	Number	Other	Overall	Yes/No	Number	Other
HAN [19]	ECCV 18'	28.65	52.25	13.79	20.33	-	-	-	-
GVQA [3]	CVPR 18'	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
UpDn [4]	CVPR 18'	39.74	42.27	11.93	46.05	63.48	81.18	42.14	55.66
+ AReg [22]	NeurIPS 18'	41.17	65.49	15.48	35.48	62.75	79.84	42.35	55.16
+ MuRel [8]	CVPR 19'	39.54	42.85	13.17	45.04	-	-	-	-
+ GLR [16]	ACL 19'	42.33	59.74	14.78	40.76	51.92	-	-	-
+ RUBi [9]	NeurIPS 19'	47.11	68.65	20.28	43.18	63.10	-	-	-
+ SCR [31]	NeurIPS 19'	48.47	70.41	10.42	47.29	62.30	77.40	40.90	56.50
+ LMH [11]	EMNLP 19'	52.45	69.81	44.46	45.54	61.64	77.85	40.03	55.04
+ LMH + CE [14]	NeurIPS 20'	54.55	74.03	49.16	45.82	-	-	-	-
+ LMH + CSS [10]	CVPR 20'	58.95	**84.37**	49.42	48.21	59.91	73.25	39.77	55.11
+ LMH + ECD (ours)	Under Review	**59.92**	83.23	**52.59**	**49.71**	57.38	69.06	35.74	54.25

Table 1: Accuracies on VQA-CP $v2$ test set and VQA 2.0 validation set, when trained on their respective training sets. We report the results for several state-of-the-art VQA models and debiasing approaches. Best results are shown in bold.

ciate irrelevant parts of the visual/textual information with a specific answer based on spurious dataset correlations. Hence, we are penalizing the model if its prediction is one of the top-N most biased answers for that particular question type. Additionally, this mechanism penalizes the model when it outputs an answer with high confidence and the necessary information is not present, acting as a regularizer. NAA generates probability distributions over possible answers that are more stable when the model does not know what to answer or lacks critical information to answer the question. By analyzing Fig. 1, when the VQA model has all the necessary information, it correctly predicts "black" with a confidence of 77%. Conversely, when it lacks enough visual or textual information, it predicts "red" and "brown" with a confidence of 0.14% and 24.07%, respectively.

4. Experimental Analysis

We evaluate ECD for debiasing VQA models mainly on the VQA-CP $v2$ test set [3], but we also present experimental results on the VQA 2.0 validation set [15] and VQA-CP $v1$ test set for completeness. We follow the standard VQA evaluation procedure [5], reporting model accuracy in four categories: *yes/no*, *number*, *other*, and *overall* questions. For a fair comparison, we perform the same data preprocessing steps presented in UpDn [4], using the available implementation[1]. For the experiments in which we measure computational efficiency, we use an NVIDIA Tesla M40 GPU and the PyTorch profiler.

We incorporate the loss function of LMH [11], shown as LMH + ECD, and compare it with previous approaches on VQA-CP $v2$ and VQA 2.0. Tab. 1 displays the results, grouped according to the VQA model that is used. ECD achieves state-of-the-art performance in the VQA-CP $v2$ dataset, reaching 59.92% in overall accuracy, ≈ 1% better in absolute terms than the previous state-of-the-art, CSS [10]. The largest gains in ECD are regarding the question type *number* (3.17% in absolute terms), followed by the category *other* (1.5%). ECD only underperforms in the *yes/no* question type, which is the type of question with the fewer number of possible answers. With only two possible answers, there is a larger chance of statistical fluctuations due to random guessing. For questions with more complex answers such as *number* (counts) [31] and *other*, ECD easily outperforms the previous state of the art, which is reflected in the *overall* accuracy.

Model	Overall	Yes/No	Number	Other
UpDn [4]	39.74	42.27	11.93	46.05
+ RUBi [9]	44.81	69.65	14.91	31.13
+ LMH [11]	55.27	76.47	26.66	45.68
+ LMH + CSS [10]	60.95	**85.60**	40.57	44.62
+ LMH + ECD (ours)	**61.78**	84.40	**45.16**	**47.15**

Table 2: Results for the VQA-CP $v1$ test set.

The data distributions for VQA 2.0 splits are similar and hence bias-prone, while the distributions for VQA-CP $v2$ are purposefully not [3]. Thus, when analyzing the difference between results for VQA-CP $v2$ and VQA 2.0 for all methods in Tab. 1, we can see that all VQA approaches that do not prioritize debiasing display large positive margins in terms of *overall* accuracy, clearly indicating that they are exploiting the biases in the data distributions. For that reason, several studies [9, 10] argue that the discrepancy between *overall* results should be small, which is the case of ECD whose margin in absolute values is 2.54%, confirming that our debiasing strategy does not significantly harm performance in balanced data distributions.

We also compare our approach with previous methods on VQA-CP $v1$ (Tab. 2). LMH + ECD achieves state-of-the-art performance, reaching 61.78% in overall accu-

[1] Code available at https://github.com/hengyuan-hu/bottom-up-attention-vqa.

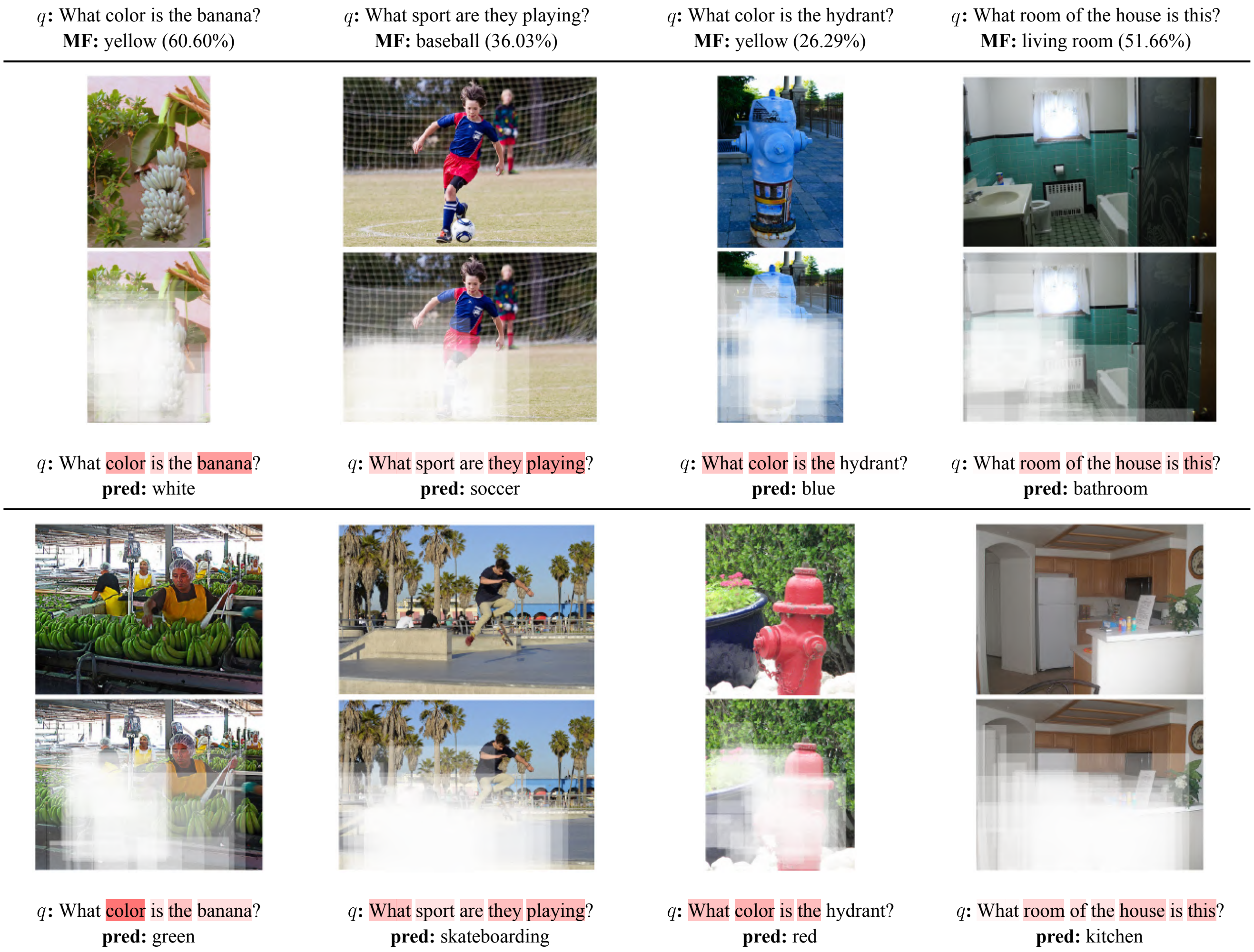

q: What color is the banana?
MF: yellow (60.60%)
q: What sport are they playing?
MF: baseball (36.03%)
q: What color is the hydrant?
MF: yellow (26.29%)
q: What room of the house is this?
MF: living room (51.66%)
q: What color is the banana?
pred: white
q: What sport are they playing?
pred: soccer
q: What color is the hydrant?
pred: blue
q: What room of the house is this?
pred: bathroom
q: What color is the banana?
pred: green
q: What sport are they playing?
pred: skateboarding
q: What color is the hydrant?
pred: red
q: What room of the house is this?
pred: kitchen

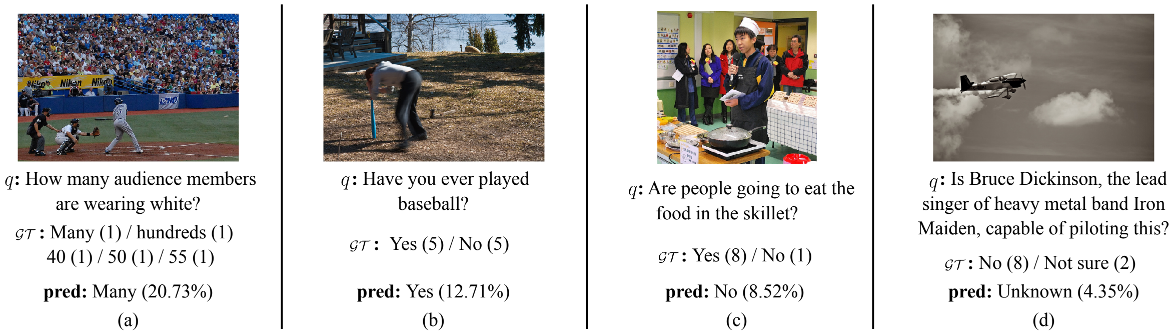
Figure 4: Examples of qualitative results using ECD. "**MF**" indicates the most frequent answer in training for the question being asked. We show the results for two images using the same question. The original image is followed by the model's visual/textual attention and its prediction.

racy, an improvement of 0.85% in absolute terms over LMH + CSS [10]. Once again ECD shows a large improvement for *number* (counts) and *other* questions. Therefore, in both datasets that were built to assess the level of bias in VQA models (VQA-CP $v1$ and VQA-CP $v2$), ECD presents state-of-the-art performance by significantly outperforming all main methods in the literature. We later show that ECD is substantially more computationally efficient in terms of memory consumption, number of operations, and total training time when compared to CSS [10] in Sec. 5.3.

4.1. Qualitative Results

By inspecting Fig. 1, we can further understand how training a VQA model using ECD affects both its attention and outputs. When a model has access to both visual and textual information (v, q), its confidence is considerably high and the attention maps indicate that relevant textual and visual information were considered. However, when we mask the visual features from the detected objects (v^-, q) — the hydrant in Fig. 1 — we expect an unbiased model not to be very confident in its predictions, since it would be making essentially an educated guess. This is precisely what occurs in practice, and we show in Fig. 1 that the model's attention disperses and its confidence substantially decreases. Similarly, when we mask words within the question (v, q^-), the VQA model also displays a disperse attention map and its confidence is once again low, due to the lack of critical information for answering the question.

Fig. 4 presents some qualitative results for ECD. Even though the most frequent answer in the training distribution



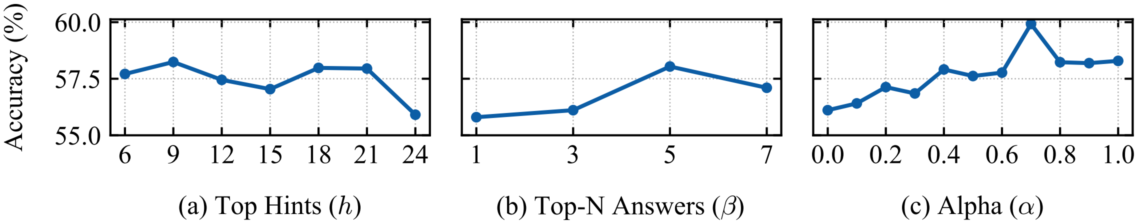
q: How many audience members are wearing white?
$\mathcal{GT}$: Many (1) / hundreds (1) 40 (1) / 50 (1) / 55 (1)
pred: Many (20.73%)
(a)

q: Have you ever played baseball?
$\mathcal{GT}$: Yes (5) / No (5)
pred: Yes (12.71%)
(b)

q: Are people going to eat the food in the skillet?
$\mathcal{GT}$: Yes (8) / No (1)
pred: No (8.52%)
(c)

q: Is Bruce Dickinson, the lead singer of heavy metal band Iron Maiden, capable of piloting this?
$\mathcal{GT}$: No (8) / Not sure (2)
pred: Unknown (4.35%)
(d)

Figure 5: Examples of failure cases of ECD for *number* and *yes/no* questions in the VQA-CP $v2$ dataset. $\mathcal{GT}$ is the ground truth answers, followed by their frequencies. The ECD's prediction *pred* is followed by the model's confidence.



(a) Top Hints (h) (b) Top-N Answers (β) (c) Alpha (α)

Figure 6: The influence of the three hyperparameters in ECD on the overall score in VQA-CP $v2$ validation set: (a) the number of top hints h (masked image regions); (b) β, the number of removed top-N most frequent answers for the counterfactuals, and (c) α, which is the visual/textual mask selection probability.

for a specific question is different from the ground-truth answer in the test set, ECD is capable of mitigating the language prior bias and correctly answering the question. Such a behavior indicates that VQA models trained with ECD use both question and image cues to answer questions. Additionally, the attention maps show that the model uses critical visual and textual information to answer the question. In Fig. 5, we show some typical failure cases of our approach. In all cases, the confidence of the model is relatively low. Note that these questions: (a) contain disagreements between human annotators; (b) they are not directly related to the image; and (c) they are subjective or would require additional external information for answering them.

5. Ablation Studies

We perform several ablation studies in order to analyze the influence of different components and hyperparameters within ECD. We build ECD on top of the ensemble-based method LMH [11] as its backbone VQA model.

5.1. Hyperparameters

In this section, we analyze the influence of the three hyperparameters within ECD: h, which is the number of masked regions; α, which provides the trade-off between masking the images and the questions; and β, which is the top-N most frequent answers that ECD removes from the $\mathcal{GT}$ of the counterfactuals. In each experiment, we freeze the other hyperparameters in $h = 9$, $\alpha = 0.5$, and $\beta = 5$.

The influence of masking the different number of critical objects (h) is shown in Fig. 6a. The method seems to be robust across different values of h, since by varying h ECD generates quite similar results ($\approx 58\%$). When $h = 9$, ECD achieves the best results. Regarding β, we can see in Fig. 6b that the best achieved result is when $\beta = 5$, though once again the method seems to be somewhat robust across different values of the β most frequent answers.

The influence of different values of α is shown in Fig. 6c. When $\alpha = 0$, we only mask the visual features (ECD-V). In contrast, when $\alpha = 1$, we only mask the textual features (ECD-Q). The method seems to yield best performance when $\alpha = 0.7$, *i.e.*, when we mask the questions $\approx 70\%$ of the time (and the images $\approx 30\%$). We notice that the overall accuracy gradually increases as α increases. This indicates that ECD-Q has a greater impact than ECD-V in ECD. Since our goal is to mitigate language-priors, we penalize the model when it predicts an answer without considering the visual cue, *i.e.*, when it explores spurious correlations between question types and answers. When we mask

Model	Method	Overall	Yes/No	Number	Other
UpDn [4]	Original	39.74	42.27	11.93	46.05
	Original*	39.59	42.36	12.47	45.58
	+ ECD-Q	40.69	41.71	13.41	47.63
	+ ECD-V	40.53	**45.35**	12.68	45.65
	+ ECD	**41.78**	42.74	**14.89**	**48.66**
RUBi [9]†	Original	44.23	–	–	–
	Original*	45.13	45.79	19.36	51.86
	+ ECD-Q	46.47	47.02	21.17	53.11
	+ ECD-V	45.81	46.41	20.28	52.50
	+ ECD	**46.69**	**47.24**	**21.46**	**53.32**
LMH [11]†	Original	52.05	–	–	–
	Original*	53.69	75.41	36.06	47.15
	+ ECD-Q	58.29	82.27	50.64	47.82
	+ ECD-V	56.11	77.71	46.99	47.29
	+ ECD	**59.92**	**83.23**	**52.59**	**49.71**

Table 3: Accuracies on VQA-CP $v2$ test set for several VQA architectures. ECD denotes the model with both ECD-V and ECD-Q. Original* represents the results based on our execution of the original implementations; † represents the ensemble-based methods with UpDn architecture.

the visual object, we are only forcing the model to look at the critical regions to answer the question, whereas when we mask the textual information, we are forcing the model to actually use the visual cue to answer the question while penalizing the model when it predicts "biased" answers.

5.2. Architecture Agnosticism

Given that ECD is model-agnostic, it can be effortlessly incorporated into different VQA models without changing the underlying architecture. In Tab. 3, we experimentally demonstrate the generality and effectiveness of our learning scheme by showing the results of incoporating our approach into three different architectures, including Updn [4], RUBi [9] and LMH [11].

Note that applying our method on these architectures leads to important gains over the baselines trained with their original learning strategy. We report a gain of 6.23% in LMH. Furthermore, we can see that when we apply our full method (with both ECD-V and ECD-Q), models often achieve the best performance.

5.3. Computational Efficiency

In order to understand how ECD performs when compared to more costly methods such as CSS [10], we measured the training time using the hardware mentioned in Sec. 4. CSS [10] was so far the state-of-the-art method for counterfactual synthesizing. It performs several forward and backward passes and rely on Grad-CAM [24] analysis to build the counterfactual samples and ground-truth labels. In contrast, we do not rely on any gradient-based methods to generate counterfactual samples.

We compute the number of multiply-accumulate operations (MAC) and memory usage of both CSS and ECD during training. To measure only the memory usage of the forward and backward passes without taking into account model and data sizes, we use the PyTorch profiler to list all operations performed when forwarding a single sample.

We remove all operations that are not model-specific, such as sending data to the GPU or any other framework-specific command. In terms of memory usage, we observe a reduction of $\approx 50\%$ from CSS ($\approx$ 5Gb) to ECD ($\approx$ 2.5Gb), as displayed in Tab. 4. This reduction is due to the amount of forward and backward passes necessary for training CSS (3 forward and 2 backward passes), while our approach needs a single forward and backward pass. Furthermore, ECD is more efficient in terms of MAC: while CSS requires 0.42 GMAC, ECD uses only 0.14 GMAC, reducing $\approx 67\%$.

We train each method for 30 epochs using the same batch size and hardware, and we observe that ECD is $\approx 41\%$ faster than CSS. Finally, it is important to point out that ECD achieves state-of-the-art results despite being a much more computationally-efficient approach to synthesize counterfactuals towards unbiased VQA.

Computational Efficiency	CSS	ECD	Reduction
Number of Operations (GMAC)	0.42	0.14	66.67%
Memory Consumption (GB)	5.02	2.52	49.80%
Total Training Time (h)	7.04	4.14	41.19%

Table 4: Contrasting ECD and CSS regarding computational efficiency. Training time is accumulated over iterations.

6. Conclusions

In this paper we have introduced ECD, a novel model-agnostic counterfactual sample synthesizing procedure for unbiased VQA. ECD design is simple, efficient, and yet effective. It synthesizes counterfactual training samples by masking critical objects from the images or words from the questions. We have designed a new Negative Answer Assignment mechanism to generate ground-truth labels for these counterfactual samples, acting both as a regularizer and a penalty factor to guide the model towards unbiased predictions. We have shown the effectiveness of ECD through an extensive experimental analysis, and we have executed several ablation studies to show the influence of each component and hyperparameter choice. We have demonstrated that ECD substantially outperforms the previous state-of-the-art methods on VQA-CP $v2$, a dataset specifically designed to account for language biases.

As future work, we intend to extend ECD to other multimodal tasks that suffer from language biases (*e.g.*, image captioning and retrieval). Additionally, we want to investigate whether changing the visual encoder of the framework significantly affects in model debiasing.

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Bias-resilient neural network. *CoRR*, abs/1910.03676, 2019.

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[3] 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.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[5] 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.

[6] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109, 2019.

[7] Shikha Bordia and Samuel R Bowman. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*, 2019.

[8] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019.

[9] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *Advances in neural information processing systems*, pages 841–852, 2019.

[10] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog, 2017.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[14] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020.

[15] 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.

[16] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *ACL workshop*, 2019.

[17] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1), 2017.

[18] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[19] Mateusz Malinowski, Carl Doersch, Adam Santoro, and Peter Battaglia. Learning visual question answering by bootstrapping hard attention. In *ECCV*, 2018.

[20] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018.

[23] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[25] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, 2019.

[26] Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*, 2018.

[27] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. *arXiv preprint arXiv:1811.00613*, 2018.

[28] Weidong Tian, Rencai Zhou, and Zhongqiu Zhao. Cascading top-down attention for visual question answering. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[29] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[30] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8919–8928, 2020.

[31] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *Advances in Neural Information Processing Systems*, pages 8604–8614, 2019.

[32] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[33] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.