Debiasing Multimodal Models via Causal Information Minimization

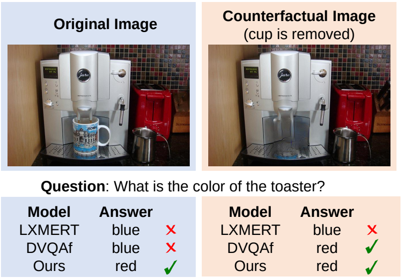
Vaidehi Patil **Adyasha Maharana** **Mohit Bansal**
UNC Chapel Hill
{vaidehi, adyasha, mbansal}@cs.unc.edu

Abstract

Most existing debiasing methods for multimodal models, including causal intervention and inference methods, utilize approximate heuristics to represent the biases, such as shallow features from early stages of training or unimodal features for multimodal tasks like VQA, etc., which may not be accurate. In this paper, we study bias arising from confounders in a causal graph for multimodal data, and examine a novel approach that leverages causally-motivated information minimization to learn the confounder representations. Robust predictive features contain diverse information that helps a model generalize to out-of-distribution data. Hence, minimizing the information content of features obtained from a pretrained biased model helps learn the simplest predictive features that capture the underlying data distribution. We treat these features as confounder representations and use them via methods motivated by causal theory to remove bias from models. We find that the learned confounder representations indeed capture dataset biases and the proposed debiasing methods improve out-of-distribution (OOD) performance on multiple multimodal datasets without sacrificing in-distribution performance.[1]

1 Introduction

The success of multimodal models in various tasks has been attributed to their ability to rely on spurious correlations present in the training data (Jabri et al., 2016; Agrawal et al., 2016; Zhang et al., 2016a; Goyal et al., 2017). An example of image bias in VQA is when the model tends to look at prominent objects in the image rather than focusing on the object about which the question is asked (Wen et al., 2021) (see example in Fig. 1). These models leverage such biases to perform well on in-distribution (ID) evaluation data (Agrawal et al., 2018). However, their poor performance on out-of-distribution data reveals that they merely rely on superficial features rather than capturing the true causal relationships between inputs and targets.



Question: What is the color of the toaster?

Model	Answer		Model	Answer	
LXMERT	blue	✗	LXMERT	blue	✗
DVQAf	blue	✗	DVQAf	red	✓
Ours	red	✓	Ours	red	✓

Figure 1: Multimodal models tend to rely on spurious correlations in the dataset to answer questions. Existing methods remove unimodal biases whereas our method removes biases from cross-modal interactions as well and is more invariant to irrelevant features (e.g., the coffee mug) in this example.

Existing methods attempt to diminish a model's reliance on these shortcuts by taking one or both of two primary strategies: (a) by balancing the sample groups with and without spurious correlation, e.g. via data augmentation (Gokhale et al., 2020) or sample synthesis (Chen et al., 2020, 2022; Kolling et al., 2022a), and (b) by explicitly

[1] Our code is available at: https://github.com/Vaidehi99/CausalInfoMin

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

eliminating the impact of spurious correlations during model training or inference (Huang et al., 2022; Lin et al., 2022; Pan et al., 2022). In the former approach, the identification of the unique set of spurious correlations in each sample becomes essential to curate augmented samples for achieving balance. Consequently, approaches that alleviate biases in features or predictions, independent of the availability of non-spurious data, are more desirable.

Recent research on debiasing models has emphasized causal theory's importance (Zhang et al., 2021; Liu et al., 2022; Bahadori and Heckerman, 2020) i.e., spurious correlations often stem from confounding variables inducing non-causal dependencies (Pearl et al., 2000). However, identifying and addressing biases affecting prediction accuracy remains challenging. Previous multimodal studies have used early-trained image features as contextual biases for multi-label image classification (Liu et al., 2022) or added unimodal training branches for Visual Question Answering (VQA) (Niu et al., 2021), missing biases arising from multimodal interactions. Hence, in this work, we represent the bias as confounder variables that have a direct causal effect on multimodal features and the corresponding predictions. Spurious correlations represent the simplest predictive features that explain biased datasets (Geirhos et al., 2020), thereby making them easily learnable by machine learning models under limited representation capacity (Yang et al., 2022). We capitalize on this notion to study a novel framework that combines information theory and causal graphs to learn confounder representations capable of capturing spurious features.

We explore two approaches to learn confounder representations from biased multimodal features: (a) *latent variable modeling* via a generative model and (b) *rate-distortion* minimization (Shannon, 1948). For (a), we employ an autoencoder to reconstruct basic predictive information from biased features and use these reconstructed features as substitutes for unobserved confounders (Huang et al., 2022). We then apply backdoor adjustment based on the average treatment effect (VanderWeele, 2015) using feature reweighting (Kirichenko et al., 2022) on these confounders, referred to as ATE-D. In (b), we minimize the bits required to encode biased feature representations while simultaneously minimizing cross-entropy loss for predicting ground truth targets from these features. This leads to the loss of diverse information, retaining simple, highly predictive features that arise from spurious correlations, used to compute the (unbiased) *total effect* (VanderWeele, 2015) of the input in TE-D.

We evaluate the proposed methods on several multimodal tasks and along multiple dimensions i.e., in-distribution and out-of-distribution performance, efficiency and robustness. Results show that these methods not only outperform baseline models with lower training overhead, but also yield additional gains on top of unimodal debiasing methods. Our contributions are as follows:

1. We introduce two methods, TE-D and ATE-D, employing causally-motivated information loss to learn confounders from biased features and use confounders to debias models.
2. Our methods remove multimodal biases and yield up to 2.2% and 2.5% gains over LXMERT (Tan and Bansal, 2019), on VQA-CP and GQA-OOD (Kervadec et al., 2021) datasets respectively, and 0.7% gains on top of unimodal debiasing (Wen et al., 2021).
3. We propose a sufficiency score (λ) for quantifying the reliance of models on spurious features; results show that our methods improve robustness to spurious correlations.
4. We analyze the confounders learnt in ATE-D, TE-D and show that they encode biases.

2 Related Work

Data Augmentation. Balancing data (Zhang et al., 2016b) can involve training a generative model for sample synthesis (Agarwal et al., 2020; Sauer and Geiger, 2020), designing suitable data selection heuristics (Chen et al., 2020), or curating balanced/counterfactual samples (Goyal et al., 2017; Gokhale et al., 2020; Kolling et al., 2022c). Human explanations are used as additional training signals to promote reasoning (Ying et al., 2022; Wu and Mooney, 2019; Selvaraju et al., 2019). We debias models using existing biased data.

Inductive Bias in Model Architecture. Agrawal et al. (2018) explicitly design inductive biases to prevent the model from relying on training priors. Clark et al. (2019); Cadene et al. (2019); Ramakrishnan et al. (2018) rely on a separate QA branch to weaken the language prior in VQA models via adversarial or multi-task learning. Wen et al. (2021) use contrastive loss to remove unimodal biases for VQA. Peyrard et al. (2022) discover invariant correlations in data across different training distributions to enable generalization.

Causal Perspective. Lin et al. (2022) use causal intervention through backdoor adjustment (Glymour et al., 2016) to disentangle the biases for unsupervised salient object detection. Huang et al. (2022) use ATE to debias referring expression models. Niu et al. (2021) compute the Total Indirect Effect (TIE) of the multimodal branch to omit the influence of unimodal branches. Veitch et al. (2021) formalize counterfactual invariance and its relation to OOD performance. Liu et al. (2022) use features from early training as confounders and compute the Total Direct Effect (TDE) for multi-label image classification. We combine information theory and causal theory to learn confounders from biased representations and use them via ATE and TE causal mechanisms to debias a model.

3 Debiasing Methods: ATE-D and TE-D

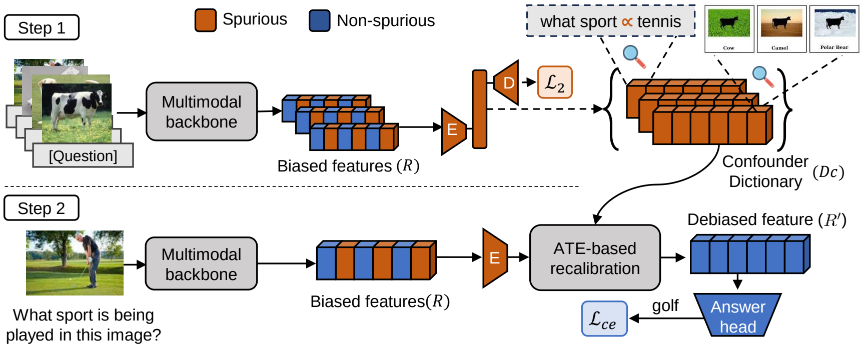


Figure 2: An illustration of our method ATE-D based on autoencoder-based confounder modeling and Average Treatment Effect causal mechanism (see Sec. 3.1). (Step 1) The confounders are modeled using autoencoder and (Step 2) biased features are debiased via recalibrated using confounders.

Neural nets are expected to trade-off between maximal compression of the learnt representations and maximal fitting to the labels (Information-Bottleneck) (Shwartz-Ziv and Tishby, 2022). Yang et al. (2022) show empirically that deep models preferentially encode dataset shortcuts under limited representation capacity. Hence, we propose information minimization by limiting representation capacity via low-dimensional vectors to learn bias/confounders. Similar approaches exist i.e. Kallus et al. (2018) recover latent confounders by performing low-rank matrix factorization on high-dimensional data, Sen et al. (2017) use low-dimensional variable to encode confounder. We explore two causal debiasing methods based on information minimization i.e., ATE-D and TE-D, as discussed next.

3.1 ATE-D: Deconfounding Using Average Treatment Effect

We follow a 2-step framework where we start with a pre-trained biased model, then (1) obtain the substitute confounders from the latent variables of autoencoder (Huang et al., 2022) and (2) use these confounders to debias the pretrained model using feature reweighing (Kirichenko et al., 2022).

Step 1: We collect the biased features $r \in R$ from a biased model for all samples in the training data and train an autoencoder composed of dense layers (F_{enc}, F_{dec}) to encode them into a lower dimension (see top, Fig. 2). The latent dimensions of the generative model capture the most common biases in the dataset and serve as a substitute for the confounders. We use a small-capacity network in order to capture the negative biases in the latent dimensions and avoid encoding positive features. F_{enc}, F_{dec} are trained using the reconstruction loss $L_{recon} = d(R, R)$, where $d(,)$ is the Euclidean distance function. We model the substitute confounders $\hat{c} \in \hat{C}$ for R ($\hat{}$ represents approximation) and cluster them to get a dictionary $D_{\hat{c}}$, which represents the main elements of $\hat{C}$ for efficient backdoor adjustment.

Step 2: Kirichenko et al. (2022) show that non-spurious features can be emphasized in biased features by reweighing them using a balanced dataset. However, creating balanced data is non-trivial for complex tasks like VQA. To overcome this challenge, we instead create an instantiation of backdoor adjustment that reweighs features based on their similarity with the substitute confounders (see bottom, Fig. 2). We hypothesize that this leads to lower weights for the simple spurious features and higher weights for more complex predictive features, alleviating the over-reliance on spurious features for prediction. For a sequence of biased features $r = [r_1, r_2, ..., r_k]$, we recalibrate each r_i according to their similarity with the confounders in D_c. The weight w_i corresponding to r_i is $1 - \frac{1}{len(D_{\hat{g}})} \sum_{g_j \in D_{\hat{g}}} s(F_{enc}(r_i), g_j)$, where $s(.)$ is the cosine-similarity function (see ATE-based recalibration in Fig. 2 and see Appendix for explanation of recalibration as an instantiation of back-door adjustment). The resulting debiased features $R' = [r'_1, r'_2, ..., r'_k]$, where $r'_i = w_i * r_i$, are then used for prediction as shown in Fig. 2.

3.2 TE-D: Debiasing Using Rate-Distortion & Total Effect

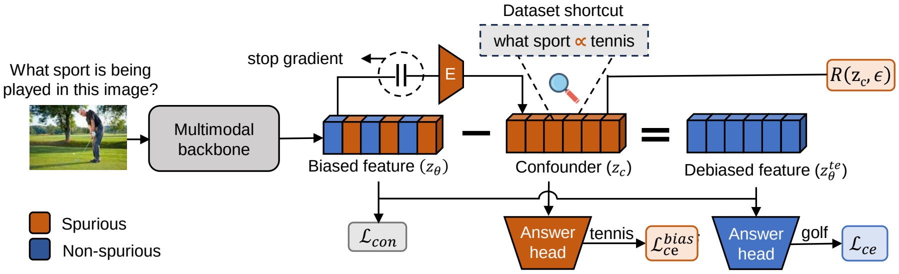


Figure 3: An illustration of our method TE-D based on rate distortion & Total Effect (see Sec. 3.2). The biased features are used to learn confounders via rate-distortion minimization and cross-entropy loss (L_{ce}). The confounders are subtracted from the biased features to get debiased feature.

The rate-distortion function $R(Z, \epsilon)$ measures the minimum number of bits per vector required to encode the sequence $Z = \{z_1, z_2, ...z_n\} \in \mathcal{R}^{n \times d}$ such that the decoded vectors $\{\hat{z}\}_{i=1}^{n}$ can be recovered up to a precision ϵ^2 i.e., $R(Z, \epsilon) = \frac{1}{2}\log_2 \det(I + \frac{d}{n\epsilon^2} ZZ^T)$ where $\frac{1}{n}ZZ^T$ is the estimate of covariance matrix for the Gaussian distribution (Chowdhury and Chaturvedi, 2022) and assuming that the vectors are i.i.d. samples from $\mathcal{N}(0, 1)$. Rate-distortion values are higher for distribution with high variance (diverse features). Hence, we minimize the rate-distortion to learn confounder representations in TE-D (see Fig. 3). Given a *biased model* with parameters θ, we first obtain the biased feature z_θ. Then, we encode the z_θ into a lower dimension to promote information loss, along with a classification head ($\mathcal{L}_{ce}^{conf}$) to encourage retaining predictiveness of the information present in the encodings, which we treat as the confounder representation z_c. Finally, we enforce rate-distortion minimization ($R(z_c, \epsilon)$) on z_c for promoting the loss of complex feature information. We enforce a stop gradient (see in Fig. 3) prior to the encoder in order to prevent the training signals for learning confounder representations from seeping into the parameters of the biased model.

In order to isolate the causal effect of M, we need to cut off the link $C \rightarrow M$ (see Fig. 7(c)). This can be achieved by computing the total effect i.e., $A_{m,c} - A_{m*,c}$, where m and $m*$ represent the treatment and no-treatment conditions respectively, while c represents the confounder resulting from $M = m$. We implement this at the feature level by representing $A_{m,c}$ with the biased features z_θ and $A_{m*,c}$ with the confounder features z_c. Next, we take the difference of those features to secure z_θ^{te} which represents the direct effect of M. i.e. $z_\theta^{te} = z_\theta - z_c$. We further aid the debiasing process by enforcing a contrastive loss between the three sets of features $z_\theta, z_c, z_\theta^{te}$ as:

$$\mathcal{L}_{con} = \log \frac{\mathbf{e}^{s(z_\theta^{te}, z_\theta)}}{\mathbf{e}^{s(z_\theta^{te}, z_\theta)} + \mathbf{e}^{s(z_\theta^{te}, z_c)}} \tag{1}$$

where $s(.)$ is the cosine similarity function. The contrastive loss penalizes the model when the confounder is correlated with the biased feature z_θ and hence, promotes debiasing of the multimodal backbone itself. In summary, we jointly optimize the model for learning confounder representations via $\mathcal{L}_{ce}^{conf}, R(Z_c, \epsilon)$ and debiasing with the help of the learned confounders via $\mathcal{L}_{con}, \mathcal{L}_{ce}$ i.e., $\theta_{deconf} = \text{argmin}_\theta \mathcal{L}_{con} + \mathcal{L}_{ce} + \mathcal{L}_{ce}^{conf} + \alpha R(Z_c, \epsilon)$, where α is the weight for rate-distortion loss.

4 Measuring Necessity, Sufficiency of Spurious Features in Multimodal Tasks

Necessity. *Type 1 Features* (Joshi et al., 2022) are neither necessary nor sufficient for predicting the label e.g., 'person' (visual feature) when the VQA question is "How many trees are in the picture?"

(see left, Fig. 4). We assess model robustness by comparing performance with and without these features. Unbiased models remain unaffected, but biased ones rely on such features due to misleading correlations. Type 1 features predominantly arise from the image in multimodal tasks, as depicted in Fig. 4. Thus, we evaluate the necessity of these features using counterfactual images (Agarwal et al., 2020) (see Sec.5).

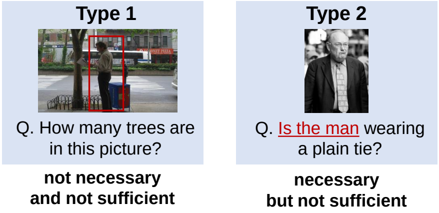

Type 1
Q. How many trees are in this picture?
not necessary and not sufficient
Type 2
Q. Is the man wearing a plain tie?
necessary but not sufficient

Figure 4: Types of spurious features (red) in VQA based on necessity and sufficiency.

Sufficiency. *Type 2 Features* are necessary but not sufficient for predictions, like "Is the man" (see right, Fig. 4). To assess model robustness to such features, we propose a new metric - sufficiency score of a feature. Ying et al. (2022) define models' certainty of predictions as the KL-divergence between the predicted output distribution and a uniform distribution across all samples in a group. We define the sufficiency score (λ) as the percentage of the model's certainty that can be attributed to the spurious component of the input in making a prediction. For a data sample (x, y), where the input x consists of the spurious feature x^s and the remaining context x^c, i.e., $x = [x^s; x^c]$, we compute the sufficiency λ as $\lambda = \frac{\sum_{i=1}^{G} \mathrm{KL}(f(y_i|x_i^s)||\mathbf{U})}{\sum_{i=1}^{G} \mathrm{KL}(f(y_i|x_i)||\mathbf{U})}$. Here, $\mathbf{U}(.)$ represents the uniform distribution, $f(.)$ denotes the trained model, and G is a group of samples. In the case of the multimodal Visual Question Answering (VQA) task, where $x_i = (q_i, v_i)$, we evaluate sufficiency of Type 2 features that arise in the textual modality q_i. To compute $f(y_i|q_i^s, v_i)$, we mask q_i^c in the query input sent to $f(.)$.

5 Experiment Setup

5.1 Datasets

- VQA-CP (Agrawal et al., 2018): It is a re-organization of the VQAv2 (Antol et al., 2015) such that the distribution of question type-answer correlation is different between the train and test splits. This evaluation helps demonstrate the method's ability to debias in a setting where language bias is dominant.

- VQA-CP + IV-VQA: We evaluate it on a new version of the VQA-CP test set where we replace the image in each sample with their invariant counterparts from the IV-VQA dataset from (Agarwal et al., 2020). IV-VQA dataset has images replaced with their edited version obtained after removing irrelevant objects in a way that the predicted answer does not change. This adds another layer of hardness to the benchmark along the image dimension. This evaluation helps demonstrate the method's ability to debias in a setting where both language and vision biases are dominant.

- GQA(Hudson and Manning, 2019), GQA-OOD(Kervadec et al., 2021):
GQA evaluation helps measure visual reasoning as well as compositional question-answering abilities. GQA-OOD is a re-organization of the GQA dataset that introduces distribution shift in validation and test sets based on question type similar to VQA-CP.

- NLVR2 (Suhr et al., 2019): It helps the generalization to multimodal tasks other than question answering. It helps evaluate reasoning abilities about sets of objects, comparisons, and spatial relations.

All our experiments are run with a single seed value.

Baselines. We use D-VQA$_f$ (feature perspective only) (Wen et al., 2021) based on LXMERT as the baseline for experiments with VQA-CP and train from scratch due to the aforementioned reasons. We also present results from D-VQA (both feature & sample perspective) for comparison, however, note that methods using data balancing are not comparable to causal debiasing methods (see Sec. 1).

	VQA-CP				IVQA-CP				Additional
	Overall	Yes/No	Num	other	Overall	Yes/No	Num	other	#MFLOPS
LXMERT Tan and Bansal (2019)	41.2	44.1	13.9	47.2	35.0	43.3	12.7	36.8	-
+ IRM Peyrard et al. (2022)	42.7	44.1	15.2	49.5	36.5	43.2	12.8	39.3	-
+ ATE-D (ours)	42.2	43.6	14.6	49.0	35.8	42.9	13.2	38.2	**0.7**
+ TE-D (ours)	43.4	48.3	14.4	48.8	36.7	46.5	12.8	38.1	8.8
+ CD-VQA Kolling et al. (2022b)	42.1	42.7	14.8	49.3	36.3	44.7	12.9	38.7	-
+ GenB Cho et al. (2023)	**52.8**	**67.3**	**29.8**	49.7	**41.3**	**50.7**	**16.7**	**39.4**	50.2
D-VQA$_f$ Wen et al. (2021)	43.9	47.5	15.7	**49.8**	37.3	45.8	13.9	39.2	18.9
D-VQA$_f$ + ATE-D	43.9	47.2	**15.9**	49.9	37.4	45.7	13.9	39.3	19.6
D-VQA$_f$ + TE-D	**44.6**	**47.8**	15.7	**50.8**	**37.8**	**46.2**	13.9	**40.1**	27.7
D-VQA	52.4	65.5	29.7	51.8	44.6	62.9	26.4	39.9	25.0

Table 1: Accuracy results on the VQA-CP (Agarwal et al., 2020) and IVQA-CP test sets. Higher is better. 'Additional MFLOPs' represents extra MFLOPS introduced by a method over LXMERT.

6 Results & Discussion

6.1 Does causal debiasing help improve out-of-distribution generalization?

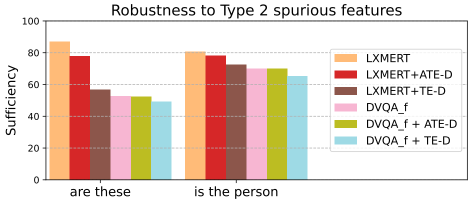


Figure 5: Using our sufficiency metric (λ, lower is better), we show that our debiased models rely less on Type 2 spurious features than baseline models.

We evaluate the effect of causal debiasing on improving generalization by evaluating our methods on three multimodal datasets. First, we observe that our methods, ATE-D and TE-D, demonstrate 1% and 2.2% gains over LXMERT on the VQA-CP test set (see Tab. 1). TE-D improves the accuracy of Yes/No category by 4.2% which has higher bias presence as seen in Fig. 5 and outperforms D-VQA$_f$, a state-of-art unimodal debiasing method for VQA (feature perspective only), by 0.8% (p=0.04) [2] in the Yes/No category, while the latter achieves better overall accuracy on VQA-CP. However, our methods can be used to debias features in any backbone and task, in contrast to D-VQA$_f$ that has been designed for VQA. Moreover, D-VQA$_f$ trains a debiased model from scratch while TE-D debiases a biased model with a few epochs of fine-tuning (see efficiency in Sec. 6.4). We see 1.8% and 2.3% gains in GQA-OOD accuracy with ATE-D and TE-D over the LXMERT baseline (see Tab. 2). The GQA-OOD dataset is further divided into OOD-Head and OOD-Tail splits which represent the samples containing answers from the head and tail of the answer distributions respectively; our methods achieve improvements in both groups. These gains are obtained along with gains in in-distribution (ID) accuracy on GQA (see Appendix). Additionally, we see 0.4%, 0.5% gains with ATE-D, TE-D respectively on NLVR2, an ID evaluation setting for visual entailment task (see Appendix). This shows that our methods do not hurt in-distribution performance and are task-agnostic. See Appendix for discussion on the improved efficiency of our methods.

6.2 What kind of biases are captured by confounder representations?

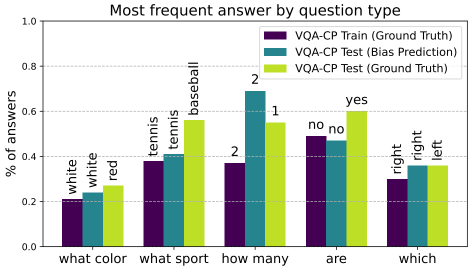


Figure 6: Most frequent answer by question type in VQA-CP train, test, and bias predictions from TE-D.

In **ATE-D**, we find that up-weighting confounder-like features (instead of down-weighting, see Sec. 3.1) significantly impairs OOD accuracy, confirming that confounder representations encode spurious correlations. We then train a non-linear probe on these confounder representations for VQA, achieving a 25% accuracy. The probe's predicted answer distribution exhibits lower entropy than unbiased features, indicating higher bias in the encoded semantic concepts within the confounders.

The bias representations in **TE-D** capture the most prominent input-output biases in the VQA-CP train set, accounting for answers in 0.34% of the answer vocabulary but covering approximately 67% of the

[2] Statistical significance is computed with 100K samples using bootstrap (Noreen, 1989; Tibshirani and Efron, 1993). All other gains are statistically significant.

	GQA	**GQA OOD**		
	ID	Tail	Head	All
LXMERT (Tan and Bansal, 2019)	59.8	49.8	57.7	54.6
+VILLA (Gan et al., 2020)	-	49.9	57.2	54.5
+MANGO (Li et al., 2020)	-	-	-	54.9
+X-CGM (Jiang et al., 2021)	-	49.9	57.5	55.6
+ATE-D (ours)	**60.0**	50.8	59.9	56.4
+TE-D (ours)	59.9	**51.4**	**60.1**	**56.8**

Table 2: Accuracy results on GQA ID, OOD datasets for various debiasing methods. Higher is better.

train questions. The classifier head connected to these bias representations achieves 28% accuracy on the VQA-CP test set, while the overall causal model accuracy is 44%. The most frequent answers predicted by this classifier head on the VQA-CP test set align with those in the VQA-CP train set, showing that the captured confounders effectively represent dataset biases (see Fig. 6).

6.3 Does causal debiasing improve robustness to spurious features?

Type 1 Spurious Features. Our IVQA-CP test set (Sec. 5) shares question annotations with VQA-CP but has images edited to remove irrelevant objects (Agarwal et al., 2020). The LXMERT model finetuned on VQA-CP shows a significant drop i.e., from 41.2% to 35.0%, on IVQA-CP (Tab. 1), indicating the evaluation's challenging nature. Our methods, ATE-D and TE-D, achieve 0.8% and 1.7% improvements respectively over LXMERT on the IVQA-CP test set, enhancing robustness to Type 1 features. D-VQA$_f$ performs explicit visual debiasing and hence, exhibits the highest robustness to Type 1 features in IVQA-CP.

Type 2 Spurious Features. A prominent source of Type 2 spurious features in VQA is the first few words of a question, as seen in Fig. 4. We select two question types i.e. questions starting with "Are these" and "Is this person", which are strongly biased in the training set of VQA-CP, and compute the sufficiency of the phrases for model predictions by masking the remaining question (see Sec. 4). As shown in Fig. 5, we find that causal debiasing methods lower the sufficiency score of the spurious feature for both of these question types, suggesting that they indeed alleviate the reliance of these models on spurious features for making predictions. TE-D and D-VQA$_f$ achieve similar sufficiency scores, suggesting that they are equally effective at improving robustness by giving more importance to the context. TE-D achieves lower λ than ATE-D which aligns with its larger acc. gains (see Tab. 1).

6.4 Is cross-modal debiasing more effective and efficient than unimodal debiasing?

D-VQA$_f$ outperforms cross-modal debiasing in Tab. 1, but when D-VQA$_f$ is treated as the biased model in TE-D, additional improvements of 0.7% (p=0.03) are achieved, indicating that cross-modal interactions contribute to bias not addressed by unimodal debiasing. Cross-modal feature-based confounders effectively mitigate biases involving multiple modalities. Our causal debiasing methods demonstrate higher efficiency compared to D-VQA, with ATE-D adding 0.7 MFLOPS and TE-D adding 3% additional parameters and 8.8 MFLOPS to LXMERT. In contrast, D-VQA adds 5% additional parameters and 18.9 MFLOPS during training, requiring more time as it is trained from scratch. Efficiency results for GQA and NLVR are same as those reported for VQA.

7 Conclusion

We propose ATE-D and TE-D to mitigate biases in models by imposing causally-driven information loss on biased features to learn confounders. Results across multimodal tasks and datasets show that the learned confounders capture biases successfully, and our methods effectively eliminate biases.

Acknowledgments

We thank Peter Hase, Zhuofan Ying, Jaemin Cho, and Nitish Joshi for their useful insights about this work, and the reviewers of this paper for their helpful feedback. This work was supported by ARO W911NF2110220, DARPA MCS N66001-19-2-4031, ONR N00014-23-1-2356, DARPA ECOLE Program No. HR00112390060. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. 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.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Mohammad Taha Bahadori and David Heckerman. 2020. Debiasing concept-based explanations with causal analysis. In *International Conference on Learning Representations*.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in Neural Information Processing Systems*, 32:841–852.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *CVPR*.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 95–112. Springer.

Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, and In So Kweon. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. 2022. Learning fair representations via rate-distortion maximization. *arXiv preprint arXiv:2202.00035*.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, Los Alamitos, CA, USA. IEEE Computer Society.

Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. 2022. Deconfounded visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 998–1006.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision – ECCV 2016*, pages 727–739, Cham. Springer International Publishing.

Jingjing Jiang, Ziyi Liu, Yifan Liu, Zhixiong Nan, and Nanning Zheng. 2021. X-ggm: Graph generative modeling for out-of-distribution generalization in visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 199–208.

Nitish Joshi, Xiang Pan, and He He. 2022. Are all spurious features in natural language alike? an analysis through a causal lens. *arXiv preprint arXiv:2210.14011*.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. 2018. Causal inference with noisy and missing covariates via matrix factorization. *Advances in neural information processing systems*, 31.

Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, and Rodrigo C. Barros. 2022a. Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3001–3010.

Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, and Rodrigo C Barros. 2022b. Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3001–3010.

Camila Kolling, Martin More, Nathan Gavenski, Eduardo Pooch, Otávio Parraga, and Rodrigo C. Barros. 2022c. Efficient counterfactual debiasing for visual question answering. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2572–2581.

Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *CoRR*, abs/2012.08673.

Xiangru Lin, Ziyi Wu, Guanqi Chen, Guanbin Li, and Yizhou Yu. 2022. A causal debiasing framework for unsupervised salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1610–1619.

Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. 2022. Contextual debiasing for visual recognition with causal mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12755–12765.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. 2022. Causal inference with knowledge distilling and curriculum learning for unbiased vqa. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–23.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

Maxime Peyrard, Sarvjeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, Saurabh Tiwary, and Robert West. 2022. Invariant language modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5728–5743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31.

Axel Sauer and Andreas Geiger. 2020. Counterfactual generative networks. In *International Conference on Learning Representations*.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2591–2600.

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. 2017. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ravid Shwartz-Ziv and Naftali Tishby. 2022. Opening the black box of deep neural networks via information. *Information Flow in Deep Neural Networks*, page 24.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

Tyler VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208.

Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. 2021. Debiased visual question answering from feature and sample perspectives. In *Advances in Neural Information Processing Systems*.

Jialin Wu and Raymond Mooney. 2019. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32.

Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. 2022. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems*, 35:20351–20365.

Zhuofan Ying, Peter Hase, and Mohit Bansal. 2022. Visfis: Visual feature importance supervision with right-for-the-right-reason objectives. In *Advances in Neural Information Processing Systems*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016a. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016b. Yin and yang: Balancing and answering binary visual questions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5014–5022. IEEE Computer Society.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4803–4813, Online. Association for Computational Linguistics.

A Limitations

While we evaluate robustness to spurious features, we do so on specific question types for Type 2 features and specific Type 1 features (irrelevant objects in the image). Getting an all-inclusive robustness metric for evaluating debiasing methods would be insightful. Approaches that debias using data augmentation or sample balancing, although cumbersome, are more effective than feature-based debiasing approaches, including those proposed in our paper. More analysis is required to understand how the merits of sample-perspective and feature-perspective methods can be merged efficiently.

B Broader Impact

In this work, the biases that we try to mitigate stem from the spurious correlations present in the dataset that lead to a drop in performance in OOD settings. This helps models learn causal associations between inputs and targets and thus brings them closer to real-world deployment as it helps mitigate unethical use of these models . However, vision-language models may encode other societal stereotypes and biases present in the data they are trained on and also introduce new ones. VL models explored in this paper are not immune to these issues. We are hopeful that our focus on modeling biases and alleviating them is a step towards more inclusive models.

C Causal Theory Preliminaries

In this section, we discuss our proposed causal graph for multimodal tasks and the two causal mechanisms relevant to our debiasing methods.

Causal Graph. Causal graphs are directed acyclic graphs $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where the edges $\mathcal{E}$ are used to represent causal relationships between random variables $\mathcal{V}$. An example is shown in Fig. 7(a), where $\mathbf{M}$ has a *direct effect* on $\mathbf{A}$.When the variable $\mathbf{Q}$ has an *indirect effect* on $\mathbf{A}$ through a variable $\mathbf{M}$ i.e. $\mathbf{Q} \rightarrow \mathbf{M} \rightarrow \mathbf{A}$, the variable $\mathbf{M}$ is said to be a *mediator* in the causal graph. If a variable $\mathbf{C}$ has a direct causal effect on both $\mathbf{M}$ and $\mathbf{A}$, it is said to be a *confounder*.

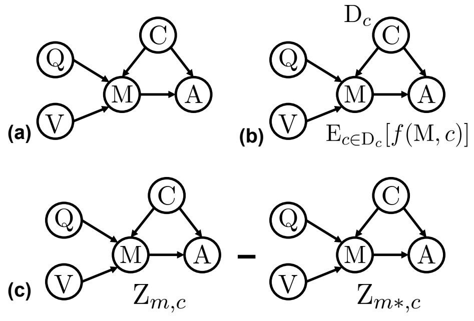


Figure 7: Demonstration of (a) our proposed causal graph for multimodal tasks, (b) Average Treatment Effect (ATE), and (c) Total Effect (TE) on (a). Values in grey indicate the 'no-treatment' condition.

Causal Perspective for Multimodal Tasks. Models developed for multimodal tasks are designed to use the combined data stream of vision (V) and language (Q) for solving the task. However, the unimodal data variables may act as confounders and give rise to spurious features in the model e.g. via

$Q \rightarrow M, Q \rightarrow A$. Existing approaches that leverage causal theory for debiasing multimodal models aim to eliminate the direct unimodal effects. However, consider the VQA example in Fig. 1. A potential spurious correlation that may lead to incorrect predictions from models on similar examples is that in most training instances where the question asks the color of an object, the object is present in the center of the image. Spurious correlations arising from such multimodal interactions are ignored in existing causal graphs for multimodal tasks. Hence, we propose to model the spurious correlation as a confounder $\mathbf{C}$ that affects the mediator $\mathbf{M}$ and the answer $\mathbf{A}$ (see Fig. 7(a)). This allows us to model the biases encoded in the multimodal features as confounder $\mathbf{C}$ and eliminate the bias using causal intervention.

In order to debias VQA models, we adopt two causal mechanisms i.e., the Average Treatment Effect (ATE) and Total Effect (TE), which essentially refer to the same effect but differ in how they deal with the confounder (VanderWeele, 2015; Tang et al., 2020). In ATE, C is treated as a distribution, and c is sampled without assuming a causal association with the treatment $M = m$. In TE, c is causally associated with the treatment $M = m$ in each sample. We explore both mechanisms in our experiments and discuss their theories below.

Average Treatment Effect. The aim of causal inference is to estimate the independent effect of an intervention on a treatment variable M on an outcome of interest A i.e. to estimate the conditional probability distribution $P(A|do(M))$. However, standard models are optimized to infer the observational conditional probability $P(A|M)$ and in the presence of confounders i.e. variables $c \in C$ that affect both A and M

$$P(A|M) \neq P(A|do(M)) \tag{2}$$

where the *do*-operation implies the causal effect of $M \rightarrow A$. $P(A|do(M))$ can be estimated using backdoor adjustment by controlling for all values of the confounders $c \in C$, i.e.,

$$P(A|do(M)) = E_{c \sim C}[P(A|M, c)] \tag{3}$$

This translates to an empirical sum over all possible values of the confounder in practice, also known as average treatment effect (ATE) (see Fig. 7(b)). When the confounders are known and observed, the confounder values are selected using suitable rules and heuristics (Pearl et al., 2000).

Total Effect. We need to isolate the causal effect of $M = m$ on A, free from the influence of the confounders C. According to causal theory, the total effect (TE) of treatment $M = m$ on A can be computed as,

$$TE = A_{m,C_m} - A_{m*,C_m} \tag{4}$$

where $M = m*$ represents the "no treatment" condition and C_m represents the confounder under the treatment condition i.e $M = m$. By retaining the confounder in both sides of the difference, we eliminate the direct effect of C_m on M (see Fig. 7(c)).

C.1 ATE-D

Step-2 of ATE-D:

Inspired by feature reweighing (Kirichenko et al., 2022), we instantiate backdoor adjustment by recalibrating r_i based on confounder similarity i.e., $E_{\hat{c} \in D_{\hat{c}}}[f(R, \hat{c})]$ (see Fig. 7(b)) as,

$$P(A|do(Q), do(V)) = P(A|do(M)) \tag{5}$$

$$E_C[P(A|M, C)] = E_{\hat{c} \in D_{\hat{c}}}[P(A|M, \hat{c})] \tag{6}$$

$$\approx P(A|E_{\hat{c} \in D_{\hat{c}}}[f(M, \hat{c})]) \tag{7}$$

See appendix of Huang et al. (2022) for complete proof. In our analysis, we instantiate $f(.)$ as the cosine similarity function in $s(.)$, as discussed in Sec 3.1.

D Analysis

While OOD generalization accuracies are indicative of the model learning causal relationships between the inputs and labels, another way to probe causal learning is to investigate if the models are robust to spurious features present in the dataset. In order to evaluate this, in this section, we discuss

Hyperparameter	**LXMERT**	**ATE**	**TE**
Learning Rate	5e-5	5e-5	5e-5
Epochs	20	5	5
Max Gradient Norm	1.0	1.0	1.0
Weight Decay	0.0	0.01	0.01
Batch Size	32	32	32
Max Length	128	128	128
Warmup Ratio	0.1	0.1	0.1
LR Decay	Linear	Linear	Linear
Optimizer	AdamW	AdamW	AdamW
Bias dimension factor	-	-	4
Confounder dictionary size	-	10	-

Table 3: Training hyperparameters for different models trained on the VQA-CP dataset.

an analysis framework for probing the behavior of models toward spurious features and propose a new metric for evaluation. Joshi et al. (2022) define the probability of necessity (PN) of a feature X_i for predicting the label Y as the probability that the ground truth label Y changes when the feature X_i is changed. Similarly, they define the probability of sufficiency (PS) of a feature X_i for predicting the label Y as the probability that setting $X_i = x_i$ in a sample where $X_i \neq x_i$ is absent changes its ground truth label Y. Based on this framework, spurious features are categorized into (a) *low PN, low PS features*: These features are irrelevant to the ground truth label e.g., person in the image when the VQA question is "How many trees are in the picture?" (see Fig. 4) (b) *High PN, low PS features*: These features are necessary but not sufficient to make predictions i.e. the model should rely on other features in their presence. For instance, when a model always answers "yes" to all questions starting with "Is the man.." irrespective of the image, the model is biased towards the feature "Is the man.." (see Fig. 4). Henceforth, we refer to the low PS, low PS, and high PN, low PS features as *Type 1* and *Type 2* features respectively. We use this framework to analyze the various debiasing methods in our experiments.

Sufficiency. In order to evaluate the robustness to *sufficiency* of type 2 features, we propose a novel metric for quantifying the sufficiency of a feature towards a prediction. We define the certainty of predictions as the KL divergence between the predicted output distribution and uniform distribution across all samples in the group (Ying et al., 2022). We define the sufficiency score (λ) as the certainty of a model's prediction when only the non-spurious features are the input to the model. Further, in order to make this metric comparable across models, we normalize this with the certainty of the model's predictions when the complete sample i.e., spurious as well as non-spurious features, is the input to the model. This results in a metric that represents the percentage of certainty of the model that can be attributed to the non-spurious component of the input. For a data sample (x, y), let the input x be comprised of the spurious feature x^s and the remaining context x^c i.e. $x = [x^s; x^c]$. The sufficiency λ is computed as follows:

$$\lambda = \frac{\sum_{i=1}^{G} \mathrm{KL}(f(y_i|x_i^s)||\mathbf{U})}{\sum_{i=1}^{G} \mathrm{KL}(f(y_i|x_i)||\mathbf{U})} \tag{8}$$

where $\mathbf{U}(.)$ represents the uniform distribution, $f(.)$ is the trained model, and G is a group of samples. A good debiasing technique should increase the sufficiency of non-spurious features. For the multimodal VQA task where $x_i = (q_i, v_i)$, we focus on the type 2 features emerging in the text modality q_i. To compute $f(y_i|q_i^c, v_i)$, we mask q_i^s in the query before sending it as input to $f(.)$.

E Results

E.1 Analysis of confounder features

We compare the most frequent answer in the VQA-CP training and test sets with those from the predictions of the bias classifier head in TE-D in Fig. 5. As discussed in Sec. 4, the predictions from

	Acc.	Cons.
LXMERT (Tan and Bansal, 2019)	74.5	39.4
+ATE-D (ours)	74.9	39.9
+TE-D (ours)	75.0	39.6

Table 4: Accuracy (Acc.) and consistency (Cons.) results on NLVR2 ID test set. Higher is better.

bias classifier head closely track the distribution of answers in VQA-CP training set, even though the VQA-CP test set distribution is significantly different from VQA-CP train. This shows that the confounder representations indeed capture the strong priors present in training set.