SC-ML: SELF-SUPERVISED COUNTERFACTUAL METRIC LEARNING FOR DEBIASED VISUAL QUESTION ANSWERING

Xinyao Shu[1], Shiyang Yan[2], Xu Yang[3], Ziheng Wu[1], Zhongfeng Chen[1], Zhenyu Lu[1]*

[1]School of Artificial Intelligence, Nanjing University of Information Science and Technology
[2]Inria, Universite Paris-Saclay
[3]School of Computer Science and Engineering, Southeast University
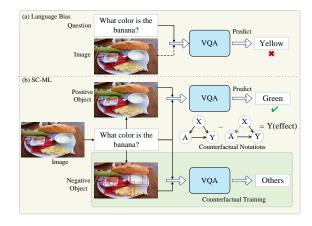
ABSTRACT

Visual question answering (VQA) is a critical multimodal task in which an agent must answer questions according to the visual cue. Unfortunately, language bias is a common problem in VQA, which refers to the model generating answers only by associating with the questions while ignoring the visual content, resulting in biased results. We tackle the language bias problem by proposing a self-supervised counterfactual metric learning (SC-ML) method to focus the image features better. SC-ML can adaptively select the question-relevant visual features to answer the question, reducing the negative influence of question-irrelevant visual features on inferring answers. In addition, question-irrelevant visual features can be seamlessly incorporated into counterfactual training schemes to further boost robustness. Extensive experiments have proved the effectiveness of our method with improved results on the VQA-CP dataset. Our code will be made publicly available.

***Index Terms*—** VQA, language bias, distance metric learning, self-supervised learning, counterfactual samples

1. INTRODUCTION

The research in computer vision, natural language processing, and multimodal processing has made remarkable progress recently. Among them, with the emergence of large-scale datasets [1, 2, 3], visual question answering (VQA) has been widely researched. Even though, VQA is still very challenging as it combines computer vision and natural language processing techniques. Moreover, VQA can bring significant social impacts, such as medical VQA, assistive devices for the blind, surveillance video queries, etc.

Visual question answering task requires the model to answer questions according to a given image. Both the image and the question are critical for the model to get the answer. However, recent research has shown that many of the current VQA models have language bias [4], i.e., the models only learn the associations between the questions and answers in

---
*Corresponding author

Fig. 1. (a) Language bias: The VQA model generates answers based on the question without going through the image. (b) Our motivations: The model divides the image into question-relevant visual objects and question-irrelevant visual objects based on the question. The question-relevant visual objects are used to infer correct answers, and the question-irrelevant visual objects are used for counterfactual training.

the training set without making reasonable use of visual information. As shown in Fig. 1(a), when faced with a problematic question-image pair, the model usually resorts to locking in the language prior knowledge in the training data and ignoring the images. Images are complex and rich, only a tiny region object (sometimes non-salient objects) helps answer the question. Therefore, during training, the model is more likely to reason directly based on the question alone or in combination with the background or salient objects in the image rather than based on the accurate relevant visual information, which causes the language bias problem. Language bias is very detrimental to the VQA model for practical applications in the real world. Because of language bias, the generalization ability and robustness of the model are limited [5, 6, 7].

Most current VQA models use a conventional supervised learning method, i.e., the models are simply supervised by a final loss function without a powerful causal reasoning capa-

bility. This likelihood-based method only supervises the final predicted answer but ignores the true causal links between the question-image contents and the answer. For example, when gives a green banana image and asks "What colour is the banana?" (In Fig. 1), the model is likely to predict the answer based on the question alone or in combination with the background (e.g. table or plate), ignoring the true cause (the association between "banana" and "yellow").

In this paper, to tackle the language bias and boost the generalization capacity of the VQA model, we propose a self-supervised counterfactual distance metric learning method. Specifically, as shown in Fig. 1(b), we design a new self-supervised metric learning method. The method has an adaptive feature selection module that adaptively classifies visual features into question-relevant and question-irrelevant visual features. The VQA model inferences answer directly based on question-relevant visual features, ensuring the actual cause of the answer. Secondly, we construct counterfactual samples based on the question-irrelevant visual features to provide a counterfactual supervised signal for the model training without manual labelling, further reducing language bias.

In summary, our contributions are threefold: 1) We propose an adaptive self-supervised distance metric learning method that can focus image features adaptively to answer questions, ensuring the actual cause of the answers and thus alleviating language bias. 2) We propose a counterfactual training method that further encourages the model to learn more accurate visual attention to reduce language bias. 3) Comprehensive experimental results have validated the effectiveness of our approach, and we achieve state-of-the-art results on publicly available benchmark datasets.

2. RELATED WORK

2.1. Language Bias in VQA

Agrawal et al. [4] pioneer the research of language bias. They reclassified the biased VQA dataset into the VQA-CP dataset, which worked on alleviating language bias from the dataset. Most current approaches to alleviating language bias can be broadly classified into four categories: Adding branch structure method [8, 9], Answer-based method [4, 10, 11], Data-balanced method [12, 13, 14] and Other method: SAR [15] adopts answer re-ranking to address language biases. D-VQA [16] adopts two unimodal bias detection modules to recognize and remove the negative biases explicitly.

2.2. Distance Metric Learning

Distance metric learning plays a significant role in a variety of computer vision applications, such as image retrieval [17], cross-modal image-text matching [18], person re-ID [19], and transfer learning [20]. Current research on distance metric learning focuses on the loss functions, e.g., Triplet loss [21, 19], N-pair-mc [17]. There is also research work exploiting

the mining techniques to consider the relationships between data samples, e.g., lifted structured [20], ranked list loss [22], msloss [23].

2.3. Self-supervised Learning

Self-supervised learning improves the feature extraction capability of the model by designing proxy tasks to mine the representational properties of the data itself as supervised information. CLIP [24] uses contrastive self-supervised learning to learn multimodal representations of images and text. SSL [12] uses Self-Supervised Learning assisted tasks to help the model overcome language bias. In contrast to SSL, our efforts focus on learning self-supervised information from question-relevant visual information and counterfactual samples to alleviate language bias.

2.4. Counterfactual Learning

Counterfactual learning has inspired several pieces of research in computer vision [25]. Counterfactual learning has been exploited in recent VQA studies [12, 13, 26, 27]. In contrast to these efforts to generate counterfactual samples for debiased training, our efforts focus on adaptively mining the data samples for their self-biases and using these counterfactual information learning to improve the reasoning capability of the model.
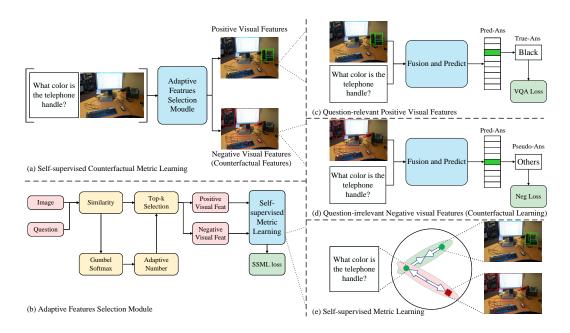
3. PROPOSED METHOD

3.1. Baseline Model

We adopt the LXMERT [6] model as the baseline model in our research. LXMERT is a dual-stream Transformer architecture consisting of two unimodal encoders for visual and language modalities, respectively, and a cross-modal encoder for aligning entities across both modalities. In this paper, we split the cross-modal encoder of the LXMERT into two parts. The first part is loaded with pre-training parameters for multimodal alignment, and the second is self-defined and without pre-training parameters and employs ReLU [28] as the activation function for multimodal fusion.

3.2. Feature Selection Module

For each image, the LXMERT uses an image encoder e_v to output a set of visual region features: $V = [v_1; \ldots \ldots; v_n] \in \mathrm{R}^{n \times d}$, where v_i is i-th object feature. For each question Q, the LXMERT uses a question encoder e_q to output a set of word features: $Q = [q_1; \ldots \ldots; q_m] \in \mathrm{R}^{m \times d}$, where q_j is j-th word.

Subsequently, we adopt cosine similarity to calculate the correlation between the visual region features V and the ques-

Fig. 2. A schematic diagram of the proposed method: We propose the SC-ML for adaptive feature selection, dividing the image features into relevant and irrelevant ones. The relevant features are applied in VQA training, while the irrelevant features assist the model training via counterfactual reasoning.

tion features Q.

$$Sim_k = \sum_{s=1}^{m} Cosine\left(V_k, Q_s\right),$$
$$k \in [1, n], s \in [1, m], \tag{1}$$

where Sim_k denotes the similarity representation of the k-th visual feature to the question feature sequence, the similarity vector Sim as a whole is represented as $Sim = [Sim_1, ; \ldots \ldots ; Sim_n]$.

As shown in Fig. 2(a), We select the top k visual features with higher Sim as the question-relevant visual features and the remaining $n - k$ visual features as the question-irrelevant visual features.

$$positive = \begin{cases} V[k], \text{if } k \text{ in } index(Top\text{-}k(Sim)) \\ 0, otherwise, \end{cases}$$
$$negative = \begin{cases} 0, \text{if } k \text{ in } index(Top\text{-}k(Sim)) \\ V[k], otherwise, \end{cases} \tag{2}$$

where positive denotes question-relevant visual features and negative denotes question-irrelevant visual features.

3.3. Self-supervised Metric Learning Module

Though by the feature selection module, we have grouped the visual features according to the question, some visual features may still be near the decision boundary, causing an incorrect division. To avoid this, we adopt Multi-Similarity Loss (ms loss) [23] to learn an embedding space (as Fig. 2(e)) that makes the distance between similar samples closer and the distance between different samples far away.

Compared with other metric learning loss, ms loss considers both the self-similarity with sample pairs and the relative similarity between sample pairs. It uses a weighting approach to obtain sample pairs with higher informativeness. The ms loss can be expressed in a specific way:

$$\mathcal{L}_{ms} = \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{k \in \mathcal{P}_i} e^{-\alpha(S_{ik} - \lambda)} \right] \right.$$
$$\left. + \frac{1}{\beta} \log \left[1 + \sum_{k \in \mathcal{N}_i} e^{\beta(S_{ik} - \lambda)} \right] \right\}, \tag{3}$$

where S_{ik} is the cosine similarity of the sample pair, λ is the similarity margin, α and β are hyperparameters.

3.4. Counterfactual Learning Module

According to the feature selection module, we classify the visual features into positive and negative features based on the question. In the answer prediction module, we use only positive features to predict answers (as Fig. 2(c)). Our intuition is that negative features should not contain question-relevant information. To better learn the relationship between positive and negative features, we propose a counterfactual learning module that takes negative features as counterfactual features to train the model in order to improve the robustness of the model. Specifically, we fuse the counterfactual features with the question features and use the same answer prediction

module to predict the answer (as Fig. 2(d)); we call the predicted answer $Pred_{neg}$. For the same reason, we call the answer predicted by the positive feature $Pred_{pos}$. As the counterfactual features do not contain features related to the question, $Pred_{neg}$ should get different answers from $Pred_{pos}$. We assign a pseudo-label to $Pred_{neg}$ that is different from the standard answer, denoted as Ans_{pseudo}. Ans_{pseudo} is to remove the first n answers predicted by $Pred_{pos}$ in the labeled answer Ans. If $Pred_{pos}$ is the same as Ans, then Ans_{pseudo} is all zeros. It can be specifically expressed as:

$$Ans_{pseudo} = \{Ans_i \mid Ans_i \in Ans,$$
$$Ans_i \notin Top\text{-}n(Pred_{pos})\}, \quad (4)$$

where Ans denotes the answer set, Ans_i denotes the i-th answer in the answer set, and $Top\text{-}n$ denotes the first top n samples.

Finally, the training loss of positive features and counterfactual features in the VQA task can be expressed as:

$$Loss_{VQA}(Pred, Ans) = -\frac{1}{N}\sum_{i=1}^{N} Ans_i \log(\sigma(Pred_i))$$
$$+ (1 - Ans_i)\log(1 - \sigma(Pred_i)),$$
$$Loss_{pos} = Loss_{VQA}(Pred_{pos}, Ans),$$
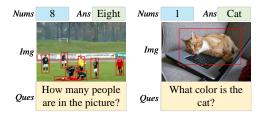$$Loss_{neg} = Loss_{VQA}(Pred_{neg}, Ans_{pseudo}),$$
$$(5)$$

Where $Loss_{VQA}$ is binary cross-entropy loss, σ is the sigmoid activation function, $Loss_{pos}$ is the training loss of positive features, and $Loss_{neg}$ is the training loss of counterfactual learning.

The overall loss function contains three parts:

$$Loss = Loss_{pos} + \gamma * (Loss_{neg} + Loss_{ms}), \quad (6)$$

where $Loss_{pos}$ is the optimization scheme for predicting correct answers by positive features; $Loss_{neg}$ is counterfactual reasoning loss; $Loss_{ms}$ is the metric learning loss, and γ is factor bwtween $Loss_{neg}$ and $Loss_{ms}$.

3.5. Adaptive Feature Selection Strategy

Fig. 3. Different pairs of questions and images require different numbers of features to be selected for the image regions related to the question.

Using a fixed selection of visual region features as positive features does not apply to all VQA questions. For example, a fixed selection of k visual features as positive features will present particular problems. As shown in the left of Fig. 3, when asked "how many people are in the picture?", at least 8 visual region features of the person are needed to answer the question altogether. However, as shown in the right of Fig. 3, when asked "what colour is the cat?", only 1 feature of the cat can answer the question altogether. Therefore, if a fixed selection strategy is adopted, 8 features are selected as the number of the question-relevant visual features for all questions. Unfortunately, for the right figure, the remaining 7 question-irrelevant visual features in Fig. 3 right may all be interference features, which still negatively impact the model's inference. If 1 feature is selected as the fixed number of the question-relevant visual feature, the other 7 question-relevant visual features in Fig. 3 will be used for counterfactual training, which is contrary to the fact.

Therefore, we propose an adaptive feature selection strategy, where we apply Gumbel-Softmax [29] to output an index topology from a similarity vector (as Fig. 2(b)). Then we apply the masking operation to select the visual features relevant to the question. Finally, by applying Gumbel-Softmax and the masking operation, we realize an adaptive trainable Top-k operation. The proposed adaptive trainable Top-k operation avoids the laborious tuning of k hyper-parameters in traditional schemes. It is a general algorithm that can be easily extended to many other applications. Specifically, the adaptive Top-k is implemented by a Gumbel-Softmax and a masking technique to achieve the back-propagation capability:

$$k = Gumbel\_softmax(Sim),$$
$$mask = Ones(k - 1) + One\_hot(k),$$
$$positive = V * mask,$$
$$negative = V * (1 - mask),$$
$$(7)$$

where $Ones(dim)$ denotes the generation of an all-1 vector, and One_hot means the one-hot embedding. The operator $+$ denotes the adding operation for vectors, and the operator $*$ is the element-wise product. We use Gumbel-Softmax to automatically generate Top-k values and integrate the k value into the training of the entire model. We make the Top-k scheme trainable, in terms of the k, via a masking operation simply because all the operations involved are continuous.

4. EXPERIMENTS

4.1. Implementation details

Datasets. To validate the effectiveness of our method, we evaluate our method on the VQA CP [4] dataset. The VQA CPv2 and VQA CPv1 datasets are two standard benchmarks for estimating the ability of models to overcome language bias problems in VQA, and they reorganize the VQA v1 [1] datasets and VQA v2 [2] so that the answers to each question category have different distributions in the training and test sets. We evaluate the model by adopting standard VQA evaluation metrics. **Training.** Our model is trained for 20 epochs with a batch size of 128, an optimizer of Adamax, and
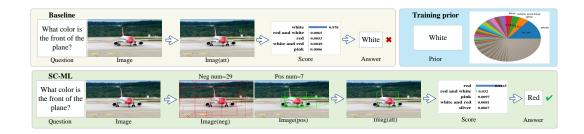
Fig. 4. Visualizations for mitigating the language bias. The blue area shows the language bias present in the current question.

Table 1. Ablation study on the VQA CPv2 dataset.

Methods	Overall	Yes/no	Number	Other
LMH-LXMERT	58.51	53.38	62.00	60.25
+ pos	67.10	84.08	60.52	60.01
+ pos + ms	67.16	84.40	60.36	59.99
+ pos + neg + ms	67.20	83.08	64.80	59.54
+ pos + neg + ms + adaptive	**68.42**	**87.57**	**63.07**	**59.86**

Table 2. The results on the VQA CPv2 dataset.

Methods	Base	Overall	Yes/No	Num	other
LMH [10]	BUTD	52.45	69.81	44.46	45.54
LMH-CSS [13]	BUTD	58.95	84.37	49.42	48.21
LMH-CSS-CL [26]	BUTD	59.18	86.99	49.89	47.16
LMH [10]	LXMERT	58.51	53.38	62.00	60.25
LMH-CSS [13]	LXMERT	63.63	84.70	62.12	53.00
LMH-CSST [30]	LXMERT	65.71	90.10	63.70	53.48
LMH-SAR [15]	LXMERT	66.73	86.00	62.34	57.84
LMH-Ours(SC-ML)	**LXMERT**	**68.42**	**87.57**	**63.07**	**59.86**

a learning rate of 5e-5. In ms loss (as Equation 3), λ is set to 0.5, α is set to 2 and β is set to 50.

4.2. Ablation Study

We provide SC-ML ablation experiments based on the VQA CPv2 dataset. We performed all the ablation experiments by building on top of the original LXMERT model [6] and LMH [10] based on LXMERT model (LHM-LXMERT). **Feature Selection Module.** From Table 1, 'pos' means inferring answers by the question-relevant visual positive features selected in Feature Selection Module. The performance of masking negative visual features to infer answers using only positive visual features is improved over the performance of using all visual features to infer answers. It proves that there is significant redundancy in visual information that affects the model's reasoning about visual information, which leads to the model inferring answers based on questions only (language bias). It also validates the effectiveness of our method in alleviating language bias. **Self-supervised Metric Learning.** From Table 1, 'ms' denotes the Self-supervised Metric Learning module. When the Self-supervised Metric Learning Module is added to the model, there is some improvement in accuracy. Because Self-supervised Metric Learning helps to mine the relationship between question-relevant positive visual features and question-irrelevant negative visual features during model training. **Counterfactual Learning Module.** From Table 1, 'neg' indicates the Counterfactual Learning Module. When the Self-supervised Learning Strategy is added, there is a certain improvement in model performance, which validates the effectiveness of the Counterfactual Learning Module on alleviating language bias. **Adaptive Feature Selection Strategy.** From Table 1, 'adaptive' indicates Adaptive Feature Selection Strategy. Other methods

use a fixed selection of 15 features. More specific results are shown in supplementary materials. The fixed feature selection strategy is likely to lead to wrongly viewing question-relevant positive visual features as counterfactual features, which is contrary to the facts. 'adaptive' indicates Adaptive Feature Selection Strategy, which is significantly better than other fixed feature selection methods. The adaptive Feature Selection Strategy avoids expensive tuning between different datasets and is suitable for generic scenarios.

4.3. Comparison with State-of-the-arts

We evaluated our method (SC-ML) on the VQA CPv2 benchmark (as shown in Table. 2). Our method significantly outperforms previous methods and achieves state-of-the-art performance on the VQA CPv2. We have visualized the SC-ML method and compared it with the baseline (LXMERT) method. As shown in Fig. 4, SC-ML can effectively alleviate language bias by adaptively masking the interfering visual features with counterfactual learning. We provide more results and visualizations in the supplementary.

5. CONCLUSION

This paper proposes a self-supervised counterfactual distance metric learning method (SC-ML) to alleviate the language bias problem in VQA tasks. SC-ML adaptively selects question-relevant image features to answer the questions effectively, alleviating language bias and improving model robustness. In addition, the negative features are seamlessly combined with counterfactual reasoning, further improving the final performance. Comprehensive experiments validate

the ability of the Self-supervised Distance Metric Learning method to alleviate language bias and achieve state-of-the-art results on the VQA-CP dataset.

6. REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *ICCV*, 2015.

[2] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017.

[3] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019.

[4] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *CVPR*, 2018.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[6] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *EMNLP*, 2019.

[7] J. Kim, J. Jun, and B. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018.

[8] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *NeurIPS*, 2018.

[9] R. Cadene, C. Dancette, M. Cord, D. Parikh, et al., "Rubi: Reducing unimodal biases for visual question answering," in *NeurIPS*, 2019.

[10] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," in *EMNLP*, 2019.

[11] Y. Guo, L. Nie, Z. Cheng, F. Ji, J. Zhang, and A. Del Bimbo, "Adavqa: Overcoming language priors with adapted margin cosine loss," in *IJCAI*, 2021.

[12] X. Zhu, Z. Mao, C. Liu, P. Zhang, B. Wang, and Y. Zhang, "Overcoming language priors with self-supervised learning for visual question answering," in *IJCAI*, 2021.

[13] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *CVPR*, 2020.

[14] T. Gokhale, P. Banerjee, C. Baral, and Y. Yang, "Mutant: A training paradigm for out-of-distribution generalization in visual question answering," in *EMNLP*, 2020.

[15] Q. Si, Z. Lin, M. yu Zheng, P. Fu, and W. Wang, "Check it again: Progressive visual question answering via visual entailment," in *ACL*, 2021.

[16] Z. Wen, G. Xu, M. Tan, Q. Wu, and Q. Wu, "Debiased visual question answering from feature and sample perspectives," in *NeurIPS*, 2021.

[17] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NeurIPS*, 2016.

[18] K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV*, 2018.

[19] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[20] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[22] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *CVPR*, 2019.

[23] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *CVPR*, 2019.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[25] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *CVPR*, 2021.

[26] Z. Liang, W. Jiang, H. Hu, and J. Zhu, "Learning to contrast the counterfactual samples for robust visual question answering," in *EMNLP*, 2020.

[27] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. Hengel, "Counterfactual vision and language learning," in *CVPR*, 2020.

[28] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[29] E. Jang, S. Gu, and B. Poole, "Categorical reparametrization with gumble-softmax," in *ICLR*, 2017.

[30] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples synthesizing and training for robust visual question answering," *arXiv preprint arXiv:2110.01013*, 2021.