Information Processing and Management 60 (2023) 103296

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Be flexible! learn to debias by sampling and prompting for robust visual question answering

Jin Liu [a], ChongFeng Fan [a], Fengyu Zhou [a,*], Huijuan Xu [b]

[a] Shandong University, JiNan, China
[b] Pennsylvania State University, Philadelphia, United States

ARTICLE INFO

Keywords:
Question debiasing
Visual question answering
Contrastive sampling
Question-type guided prompt

ABSTRACT

Recent studies point out that VQA models tend to rely on the language prior in the training data to answer the questions, which prevents the VQA model from generalization on the out-of-distribution test data. To address this problem, approaches are designed to reduce the language distribution prior effect by constructing negative image–question pairs, while they cannot provide the proper visual reason for answering the question. In this paper, we present a new debiasing framework for VQA by Learning to Sample paired image–question and Prompt for given question (LSP). Specifically, we construct the negative image–question pairs with certain sampling rate to prevent the model from overly relying on the visual shortcut content. Notably, question types provide a strong hint for answering the questions. We utilize question type to constrain the sampling process for negative question–image pairs, and further learn the question type-guided prompt for better question comprehension. Extensive experiments on two public benchmarks, VQA-CP v2 and VQA v2, demonstrate that our model achieves new state-of-the-art results in overall accuracy, i.e., 61.95% and 65.26%.

1. Introduction

Remarkable achievements have been made in various applications for visual question answering (VQA), such as personal assistants and intelligent service robots (Luo et al., 2022; Wang, Pradhan, & Gunasekaran, 2022). Most existing VQA methods tend to rely on question–answer language correlations without mining the correct visual information from the image to answer the question. For example, as shown in the first row of Fig. 1, some VQA models may favor the answer "blue" for the question type "What color" rather than providing the real color in the image, due to the fact that the most common answer for the question type "What color" in the training data is "blue". This phenomenon is related to the robustness of VQA models which typically contains two aspects introduced by Han, Wang, Su, Huang, and Tian (2021), the language distribution bias where the answers' distribution for certain question-type in train and test is different, and the visual shortcut bias which answers the question based on the improper visual grounding.

To overcome the language distribution bias, early contrastive learning-based method SSL-VQA (Zhu et al., 2020) mainly focuses on constructing negative image–question pairs through sampling irrelevant images from the training dataset. Unfortunately, exhaustive sampling for negative image–question pair where each question is paired with an irrelevant negative image, may result in the other type of bias, *i.e.*, the visual shortcut bias (Han et al., 2021), which is shown in the second row of Fig. 1. To balance the information of both modalities and avoid the visual shortcut bias, we incorporate the idea of dropout (Srivastava, Hinton,

* Corresponding author.
E-mail address: zhoufengyu@mail.edu.cn (F. Zhou).

https://doi.org/10.1016/j.ipm.2023.103296
Received 1 November 2022; Received in revised form 3 January 2023; Accepted 24 January 2023
Available online 2 February 2023

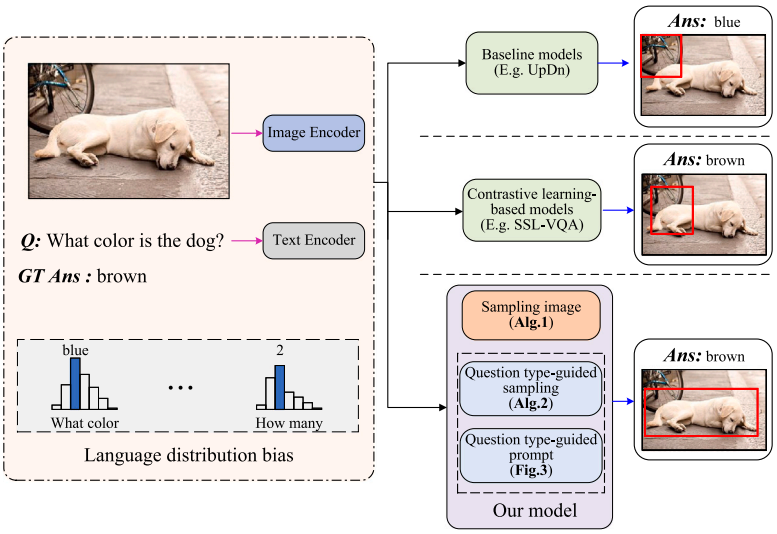

Fig. 1. Previous contrastive learning based VQA bias models alleviate language distribution bias while still suffer from the visual shortcut bias where the visual evidence does not fall on the right location. Our model with constraint sampling and question type-guided prompt learning helps alleviate both types of VQA bias.

Krizhevsky, Sutskever, & Salakhutdinov, 2014) in deep neural networks and construct the negative image–question pairs under certain probability rather than the exhaustive sampling as utilized in Zhu et al. (2020).

Also, to enhance the question contribution and alleviate visual shortcut, negative question–image pairs are also constructed via sampling irrelevant questions for each image or synthesizing negative question samples by mask (Kolling et al., 2022), in addition to the irrelevant image sampling per question (Zhu et al., 2020). On one hand, synthesized question samples (Kolling et al., 2022) increase the total number of questions and exacerbate the imbalance in the number of images and questions.

On the other hand, these methods ignore the abundant question type prior and obtain limited improvements. We introduce question type-guided sampling in the process of selective sampling, which helps constrain the negative sampling space to certain question type. We further design question type-guided prompt as language context to improve the question contribution in answering questions, inspired by the prompt learning (Jin, Cheng, Shen, Chen, & Ren, 2022; Zhou, Yang, Loy, & Liu, 2022). We are the first to utilize prompt learning to overcome the language bias in VQA task and we validate its effectiveness in debiasing problem by extensive experiments. Notably, our model is non-transformer architecture and utilizes the typical prompt design motivation (*i.e.*, retrieve potential knowledge context) to obtain certain context from the question itself.

In summary, we propose a model to Learn to Sample and Prompt for overcoming the visual shortcut and language distribution biases in VQA, namely LSP model,[1] which is shown in the third row of Fig. 1. The major contributions of this paper are listed as follows.

- We propose a debiasing model to overcome both the visual shortcut bias and the language distribution bias in VQA through sampling control. Specifically, image–question negative pairs are constructed with certain probability and question type is utilized to constrain the sampling space.
- We are the first to incorporate the prompt learning to augment the question context and alleviate the VQA bias problem by designing the learnable question type-guided prompt. We explore various variations of prompt modes and demonstrate the effectiveness of the prompt context in question comprehension and VQA bias alleviation.
- Extensive experimental results on two benchmark datasets, VQA-CP v2 and VQA v2, indicate that our LSP model achieves new state-of-the-art results and is able to provide the answer for given image and question with reasonable visual explanation.

2. Related work

2.1. Bias problem in VQA

Visual question answering (VQA) is a typical multi-modal task aiming to provide the answer for given image and question, which requires both visual perception and language reasoning (Antol et al., 2015; Yang, Miech, Sivic, Laptev, & Schmid, 2022; Zhu et al., 2020). Recent studies point out that most VQA models tend to exploit the superficial correlations between answers and questions to answer questions (Cadene, Dancette, Cord, Parikh, et al., 2019; Chen et al., 2020; Han et al., 2021; Wen, Xu, Tan, Wu, & Wu,

[1] Code will be available upon paper acceptance at https://github.com/LemonQC/LSPModel

2021). Although promising performance is achieved from utilizing such biases, these models typically obtain poor generalization on out-of-distribution datasets. The most straightforward methods to alleviate the bias problems in VQA are to construct a balanced dataset (Zhang, Goyal, Summers-Stay, Batra, & Parikh, 2016) in terms of question–image pairs. For example, Zhang et al. (2016) add the complementary abstract scenes with opposite answers for the corresponding binary question. Notably, Agrawal, Batra, Parikh, and Kembhavi (2018) design a new split from VQA dataset under Changing Prior, called VQA-CP, with various answer distributions between train and test dataset, to evaluate VQA bias problem.

The methods to alleviate VQA bias can not only be from the data preparation of balancing question–image pairs but also from the VQA model design aspect (Agrawal et al., 2018; Han et al., 2021; Zhu et al., 2020). The typical ensemble-based models are equipped with one question-only branch to capture the language biases and further adjust the answer distributions (Cadene et al., 2019; Clark, Yatskar, & Zettlemoyer, 2019; Liang, Hu, & Zhu, 2021). For example, Cadene et al. (2019) reweight the answers' distributions based on the answer predictions from the question-only branch. Another line of works utilize human annotations as attention supervision maps for correct visual grounding (Selvaraju et al., 2019; Wu & Mooney, 2019) of answers. E.g., HINT (Selvaraju et al., 2019) encourage the utilization of the visual context through the alignment of object important scores with the human-attention maps. However, the human-attention maps are costly to collect. Recently, an effective solution for VQA bias is based on contrastive learning (Chen et al., 2020; Kolling et al., 2022; Wen et al., 2021; Zhu et al., 2020), which constructs negative image–question pairs to enforce the utilization of visual context and question information.

Recently, Han et al. (2021) categorize the language bias in VQA into two types, *i.e.*, language distribution bias with different answer distributions in training and test for certain question type, and visual shortcut bias with improper visual grounding for answers. Most of current VQA models focus on overcoming the language distribution bias while ignoring the visual shortcut bias. In this paper, we follow the general contrastive learning-based sampling pipeline, and further equip our model with selective sampling and question type-guided sampling when constructing negative pairs, to address both the visual shortcut bias and language distribution bias.

2.2. Prompt learning in VQA

Besides balancing image–question pairs to overcome VQA bias problem, improving the question comprehension ability is also critical for alleviating VQA bias problem. Prompt learning aims to provide effective context for language or visual understanding by designing fixed prompt or learnable prompt modes, and is applied in many natural language processing (NLP) tasks (Ding et al., 2021; Zheng & Huang, 2021), and vision-language pre-training models (Jin et al., 2022; Yao et al., 2021; Zhou et al., 2022). For example, Yao et al. (2021) propose a context optimization model for vision tasks with learnable prompt based on pre-trained CLIP weights (Radford et al., 2021). However, prompt Learning has not been applied in VQA debiasing task previously. Moreover, the essential motivation of the prompt is to learn a certain context. Following such thought, we are first to incorporate the prompt learning to tackle the VQA debiasing problem and utilize question type prior information in the design of learnable prompt to provide more language context and improve the question comprehension ability in the VQA task. Notably, compared with the existing prompt design approach, we just borrow the context-learning idea from the prompt design rather than using it based on pre-trained model weights since our model is non-transformer without any pretraining.

3. Approach

3.1. Problem definition

In this section, we phrase the open-ended VQA task as a multi-class classification problem. Specifically, given a dataset $\mathcal{D} = \{v_i, q_i, a_i\}_i^N$ consisting of N triplets with an image $v_i \in \mathcal{I}$, a question sentence $q_i \in \mathcal{Q}$ and a corresponding answer $a_i \in \mathcal{A}$, the VQA model aims to learn a mapping function $\mathcal{F} : \mathcal{I} \times \mathcal{Q} \rightarrow \mathbb{R}^{|\mathcal{A}|}$ to generate an answer distribution over the answer candidates $\mathcal{A}$. In general, the mapping function $\mathcal{F}$ is formulated in Eq. (1) and composed of three main parts: the feature extractors for images and questions e_v and e_q, the multimodal feature fusion module $m(\cdot)$, and the answer classifier $c(\cdot)$.

$$\mathcal{F}(\hat{a}_i|v_i, q_i) = c(m(e_v(v_i), e_q(q_i))), \tag{1}$$

where the image feature extractor satisfies $e_v : v_i \rightarrow \mathbb{R}^{n_v \times d_v}$ and the question feature extractor satisfies $e_q : q_i \rightarrow \mathbb{R}^{n_q \times d_q}$. The multimodal feature fusion module satisfies $m(\cdot) : \mathbb{R}^{n_v \times d_v} \times \mathbb{R}^{n_q \times d_q} \rightarrow \mathbb{R}^{d_n}$, and the answer classifier satisfies $c(\cdot) : \mathbb{R}^{d_n} \rightarrow \mathbb{R}^{|\mathcal{A}|}$. $\hat{a}$ denotes the predicted answer distribution.

Due to the imbalance between the number of questions and images, most VQA models are overly dependent on the correlations of the answers and questions, which prevent the model from being robust to the language distribution bias, and could not provide the right visual grounding for the answer prediction and suffer from the visual shortcut bias. In this paper, we propose a model to Learn to Sample paired image–question and Prompt for given question, called LSP model, to overcome both the language distribution bias and the visual shortcut bias.

3.2. Our proposed LSP model

To balance the image–question pairs and improve language and visual cross-modal understanding for robust VQA, our LSP model shown in Fig. 2 is designed with three main components, (1) selective sampling rate to prevent the model from overly relying on the visual shortcut content in the original exhaustive sampling pairs, (2) question type-guided sampling to constrain the sampling space for negative question–image pairs, and (3) the question type-guided prompt for better question semantic comprehension. The details of these three components in LSP model are introduced in the following sections.

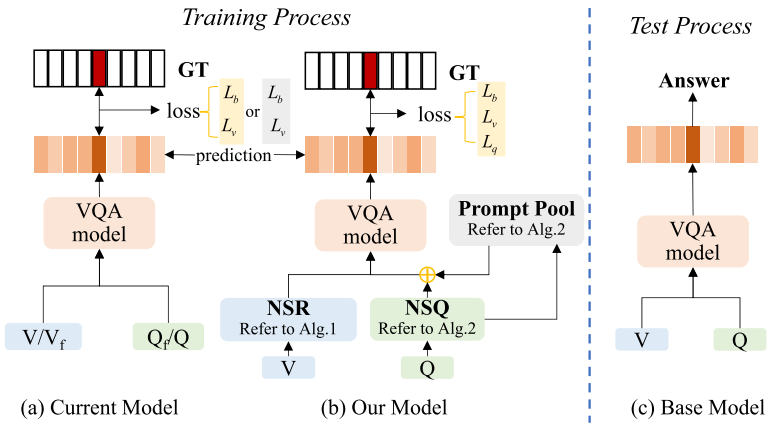

Fig. 2. Comparisons of previous VQA models (a) and our proposed LSP (b). V_f and Q_f denote the negative image and question samples. GT is the ground-truth answer. $\oplus$ represents the concatenation operation. After debiasing, we utilize the base model (e.g., UpDn (Anderson et al., 2018)) to predict the answer.

3.2.1. Selective sampling rate

Existing contrastive learning-based models, e.g., SSL-VQA (Zhu et al., 2020), propose to construct negative samples by sampling an irrelevant image for each question to address the language distribution bias, which may result in the model focusing more on the visual shortcut and excessively penalizing the prediction from the question branch (Shrestha, Kafle, & Kanan, 2020). Therefore, to alleviate the exhaustive sampling issue and balance the visual and language contribution in cross-model modeling, we introduce the sampling rate into the negative image–question sampling process to selectively sample negative pairs according to certain probability, as shown in Algorithm 1. In specific, there are three main steps of the Negative pair construction with Sampling Rate (NSR) module:

Algorithm 1: Selective negative sampling of images

Function *NSR* ($\mathcal{I}$, $\mathcal{Q}$, $\mathcal{A}$, α)**:**
- $\mathcal{I}^p, \mathcal{Q}^p, \mathcal{A}^p \leftarrow$ P_Sample($\mathcal{I}, \mathcal{Q}, \mathcal{A}$)
- $\delta \sim U[0, 1]$ ▷ Threshold for sampling
- **if** $\delta \leq \alpha$ **then**
  - $\mathcal{I}^n \leftarrow$ V_Sample($\mathcal{I}$)
  - $\mathcal{S}^v = 1$ ▷ Mark the selected negative sample
- **else**
  - $\mathcal{I}^n = \mathcal{I}^p$ ▷ Equivalent to not sampling
  - $\mathcal{S}^v = 0$
- **end**
- $\mathcal{I} \leftarrow \{\mathcal{I}^n \cup \mathcal{I}^p\}$
- $\mathcal{Q}, \mathcal{A} \leftarrow \{\mathcal{Q}^p, \mathcal{A}^p\}$
- return $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{A}$, $\mathcal{S}^v$

end

1. Randomly sample one instance, $\mathcal{I}^p$, $\mathcal{Q}^p$ and $\mathcal{A}^p$, from the training dataset and treat it as one positive instance. This triplet sampling process in Algorithm 1 is indicated as P_Sample.[2]
2. We sample a negative image $\mathcal{I}^n$ from image set $\mathcal{I}$ (indicated as V_Sample) relative to the positive instances sampled in previous step. We mark the negative images to be finally used in the training loss as 1 ($\mathcal{S}^v$ equals 1) with certain sampling rate α, and pair the selected negative images with the question–answer to construct negative triplets.
3. The positive triplet in first step and the selected negative triplets in the second step are merged to the training data. The same process is repeated to construct the training batch.

[2] The same process is applied in image sampling V_Sample, answer–question consistent sampling QI_Sample and answer–question inconsistent sampling QO_Sample.

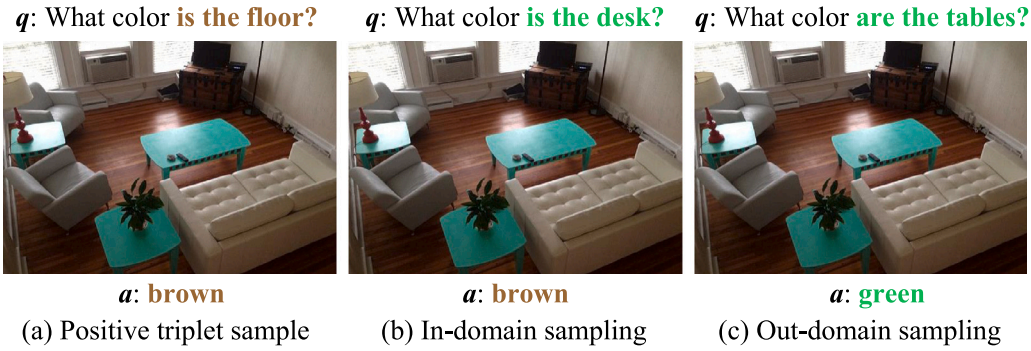

Fig. 3. Examples of in-domain and out-domain sampling in question type-guided sampling. Given the positive triplet in (a), in-domain negative sample (b) includes a sampled negative question with the same answer, while out-domain negative sample (c) includes a sampled negative question with a different answer.

3.2.2. Question type-guided sampling

Previously people try to increase the language contribution by sampling negative questions for each question–image pair, or synthesizing the negative question samples through masking the critical words (Chen et al., 2020; Kolling et al., 2022; Wen et al., 2021), with only slight improvement due to the fact that the imbalance in the number of images and questions is increased with extra question sampling. To keep the question contribution and alleviate the question imbalance, we design a question type-guided sampling to constrain the sampling space with question type prior and reduce the question sampling space. Notably when we sample negative question for given positive image-question-answer triplet, depending on whether the original answer is kept or not, the sampled negative triplets can be categorized into two types, *i.e.*, in-domain sampling and out-domain sampling. We keep both types of negative sampling to enhance question understanding which are shown to have better performance in our preliminary experiments. An example of in-domain and out-domain sampling is illustrated in Fig. 3 and the detailed question type-guided sampling process is listed in Algorithm 2 and described as follows.

- Randomly sample one instance, $\mathcal{I}^p$, $\mathcal{Q}^p$ and $\mathcal{A}^p$, from the training dataset and treat it as one positive instance.
- we construct in-domain negative samples and out-domain negative samples as illustrated in Fig. 3. We sample in-domain negative samples and out-domain negative samples with the ratio β, and pair the selected negative samples with the image to construct negative triplets.
- The positive triplets in the first step and the selected negative triplets in the second step are merged to training data. The same process is repeated to construct the training batch.

Algorithm 2: Negative sampling of questions

Function *NSQ ($\mathcal{I}$, $\mathcal{Q}$, $\mathcal{A}$, β)*:
- $\mathcal{I}^p, \mathcal{Q}^p, \mathcal{A}^p \leftarrow$ P_Sample($\mathcal{I}, \mathcal{Q}, \mathcal{A}$)
- $\xi \sim U[0,1]$ ▷ Threshold for sampling
- **if** $\xi \leq \beta$ **then**
  - $\mathcal{Q}^n \leftarrow$ QI_Sample($\mathcal{Q}$) //in-domain
  - $\mathcal{S}^{qi} = 1$ ▷ mark in-domain negative sample
  - $\mathcal{S}^{qo} = 0$
- **else**
  - $\mathcal{Q}^n \leftarrow$ QO_Sample($\mathcal{Q}$) //out-domain
  - $\mathcal{S}^{qi} = 0$
  - $\mathcal{S}^{qo} = 1$ ▷ mark out-domain negative sample
- **end**
- $\mathcal{Q} \leftarrow \{\mathcal{Q}^n \cup \mathcal{Q}^p\}$
- $\mathcal{I}, \mathcal{A} \leftarrow \{\mathcal{I}^p, \mathcal{A}^p\}$
- return $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{A}$, $\mathcal{S}^{qo}$, $\mathcal{S}^{qi}$

end

3.2.3. The learnable question type-guided prompt

Prompt learning aims to provide effective context for language or visual understanding by designing fixed prompt (e.g., adding additional fixed texts to the input) or learnable prompt modes (e.g., adding additional learnable vectors to the input), and has been

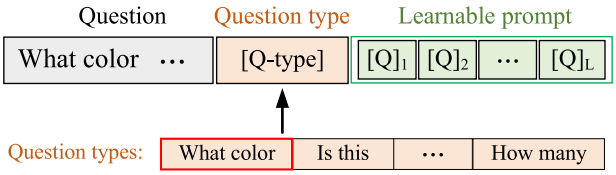

Fig. 4. The question type-guided prompt learning which consists of original question, question type and a learnable vector component.

proven useful in various natural language processing (NLP) tasks with the effect of context (Jin et al., 2022; Zhou et al., 2022). We incorporate the learnable prompt into the question representation learning to further enhance the question comprehension ability, shown in Fig. 4. Notably, our designed learnable prompt additionally incorporates the question type component and stimulates the learnable prompt to learn question type related context. Specifically, our prompt design consists of three concatenated components, the original question, the question type words and the learnable prompt component. Taking the question "What color is the car?" as an example, the question type is "What color" and the corresponding learnable prompt is $[Q]_{1\cdots L}$ with fixed length L and random initialization. Consequently, the final representation is "What color is the car? What color $[Q]_1[Q]_2 \cdots [Q]_L$". Our prompt design achieves significant performance improvement with question type prior context compared to other existing prompt modes.

3.3. Optimization and inference

We build our model on top of the typical VQA backbone UpDn (Anderson et al., 2018), with one binary cross-entropy (BCE) loss for answer classification:

$$\mathcal{L}_b = -\sum_{i}^{|\mathcal{A}|} y_i \log(\sigma(\hat{a}_i)) + (1 - y_i)(1 - \log(\sigma(\hat{a}_i))), \tag{2}$$

where $|\mathcal{A}|$ is the number of classes. $\hat{a}_i$ is the ith predicted class logit. σ is the sigmoid function. y_i denotes the target label and is computed by $y_i = min(\frac{\#votes}{3}, 1)$, where $\#votes$ represents the number of human annotated answers for each question.

With the sampled negative triplets in Sections 3.2.1 and 3.2.2, we construct two more learning losses for each sampled batch, $\mathcal{L}_v$ and $\mathcal{L}_q$. The loss for the selective negative image sampling $\mathcal{L}_v$ is shown as:

$$\mathcal{L}_v = \frac{1}{N}\sum_{i}^{N} \mathrm{SF}(\hat{a}_i)[k] \cdot s_i^v, \tag{3}$$

where N is the number of examples. SF denotes the softmax function. k is the index of the answer a_i in the answer set $\mathcal{A}$. s_i^v is the 0-1 indicator for the selected negative image.

Similarly, the loss for the negative question sampling $\mathcal{L}_q$ is formulated as:

$$\mathcal{L}_q = \frac{1}{N}\sum_{i}^{N} \mathrm{SF}(\hat{a}_i)[k] \cdot s_i^{qi} + \mathrm{SF}(\hat{a}_i)[k] \cdot s_i^{qo}, \tag{4}$$

where s_i^{qi} and s_i^{qo} are the 0-1 indicators for the selected in-domain and out-domain negative questions, respectively.

Finally, we optimize the whole LSP model via the total loss $\mathcal{L}$ in Eq. (5) consisting of the basic answer classification loss $\mathcal{L}_b$, the loss for negative image sampling $\mathcal{L}_v$, and the loss for negative question sampling $\mathcal{L}_q$.

$$\mathcal{L} = \mathcal{L}_b + w \cdot \mathcal{L}_v + \mathcal{L}_q, \tag{5}$$

where w denotes the loss weight hyper-parameter.

During the inference phase, these loss components are discarded. The testing image and question are loaded, and go through the base VQA backbone to obtain predicted answer.

4. Experiments

4.1. Datasets

In this section, following the dataset settings in previous works (Han et al., 2021; Wen et al., 2021), we evaluate our proposed LSP model on two benchmark datasets, *i.e.*, VQA-CP v2 (Agrawal et al., 2018) and VQA v2 (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017). In specific, VQA-CP v2 is constructed from VQA v2 through re-organizing to obtain different answer distributions between train and test sets. The VQA v2 consists of ~82K images and ~152K questions for the training subset, ~41K images and ~82K questions for the validation subset. In contrast, the VQA-CP v2 consists of ~121K images and ~178K questions for the training subset, ~98K images and ~110K questions for the test subset. Notably, we report the accuracy on the VQA-CP v2 test split and VQA v2 validation split.

Table 1
Result comparison in Accuracy (%) on VQA-CP v2 test set and VQA v2 val set. The **best** and second results are highlighted. Categories I-IV groups backbone models, annotation-based models, ensemble and contrastive learning-based models, and debiasing models with other backbones. Models with † indicate replicated results using released codes.

Category	Model	Base	Venue	VQA-CP v2 test (%)				VQA v2 val (%)			
				All	Yes/No	Num	Other	All	Yes/No	Num	Other
I	UpDn† (Anderson et al., 2018)	–	*CVPR'18*	40.63	41.27	13.63	47.70	64.21	81.72	43.54	56.36
	GVQA (Agrawal et al., 2018)	–	*CVPR'18*	31.30	57.99	13.68	22.14	48.24	72.03	31.17	34.65
	LXMERT (Tan & Bansal, 2019)	–	*EMNLP'19*	42.84	18.91	**55.51**	46.23	–	–	–	–
II	SCR (Wu & Mooney, 2019)	UpDn	*NIPS'19*	49.45	72.36	10.93	48.02	62.20	78.80	41.60	54.40
	CSS† (Chen et al., 2020)	UpDn	*CVPR'20*	41.16	43.96	12.78	47.48	59.21	72.97	40.00	55.13
	HINT (Selvaraju et al., 2019)	UpDn	*ICCV'19*	47.50	67.21	10.67	46.80	63.38	81.18	42.14	55.66
III	LMH (Clark et al., 2019)	UpDn	*EMNLP'19*	52.73	72.95	31.90	47.79	56.35	65.06	37.63	54.69
	DLR (Jing, Wu, Zhang, Jia, & Wu, 2020)	UpDn	*AAAI'20*	48.87	70.99	18.72	45.57	57.96	76.82	39.33	48.54
	SSL-VQA (Zhu et al., 2020)	UpDn	*IJCAI'20*	57.59	86.53	29.87	50.03	63.73	–	–	–
	LPF (Liang et al., 2021)	UpDn	*SIGIR'21*	55.34	88.61	23.78	46.57	55.01	64.87	37.45	52.08
	GGE (Han et al., 2021)	UpDn	*ICCV'21*	57.32	87.04	27.75	49.59	59.11	73.27	39.99	54.39
	D-VQA† (Wen et al., 2021)	UpDn	*NIPS'21*	60.64	88.51	46.75	49.85	64.04	81.32	43.65	56.30
	LRSV (Guo, Nie, Cheng, Tian, & Zhang, 2021)	UpDn	*TIP'22*	47.09	68.42	21.71	42.88	55.50	64.22	39.61	53.09
IV	AdvReg. (Ramakrishnan, Agrawal, & Lee, 2018)	SAN	*NIPS'18*	33.29	56.65	15.22	26.02	52.31	69.98	39.33	47.63
	RUBi (Cadene et al., 2019)	S-MRL	*NIPS'19*	47.11	68.65	20.28	43.18	61.16	–	–	–
	ECD (Kolling et al., 2022)	LMH	*WACV'22*	59.92	83.23	52.59	49.71	–	–	–	–
III	Ours	UpDn	–	**61.95**	**89.50**	52.44	**50.12**	**65.26**	**82.38**	**44.77**	**57.67**

4.2. Evaluation metrics

In terms of evaluation, we utilize the standard VQA evaluation metric, answer prediction accuracy (Antol et al., 2015), which is computed in Eq. (6).

$$Acc = \min(1, \frac{\#\text{Answer}}{3}), \tag{6}$$

where #Answer refers to the number of times the predicted answer is the same as the answer provided by ten annotators.

4.3. Implementation details

Following previous works (Han et al., 2021; Wen et al., 2021), Faster R-CNN (Ren, He, Girshick, & Sun, 2015) is utilized to extract 2048-d visual features for top 36 objects in the images. Each word is initialized with 300-d Glove embeddings (Pennington, Socher, & Manning, 2014) and processed by a GRU (Cho et al., 2014) of 1024-d hidden vector with maximum question length 14. Adam optimizer is adopted with initial learning rate of 1e-3 and fixed learning rate of 0.0005 since the epoch 10. The batch size of 256. The loss weight w in the total loss Eq. (5) is set to 3. The sampling rate α in negative image sampling Algorithm 1 is set to 0.85 and the rate β in negative question sampling Algorithm 2 is set to 0.66. All the experiments are conducted on a single 3090 GPU with pytorch 1.6. Each epoch occupies about 3 min, with a total of 20 epochs of training.

4.4. Main results on VQA-CP v2 and VQA v2

We compare our LSP model with recent state-of-the-art models on VQA-CP v2 test set and VQA v2 validation set under the standard VQA evaluation metric (Antol et al., 2015) in Accuracy (%). Results are reported in Table 1 and show the following conclusions.

On both datasets, our LSP model outperforms existing models in the overall accuracy computed on all question types and achieves new state-of-the-art without extra human annotations. In per-question type accuracy, our LSP model outperforms in most question types, with only one exception in "Num" question type achieving third place among all the models. We compare with models with different VQA backbones in category I. Among them, our LSP model surpasses its own backbone UpDn (Anderson et al., 2018) by a large margin with 21.63% higher in the overall accuracy "All", demonstrating the effectiveness of our model in reducing language bias and utilizing visual context. Notably, although LXMERT (Tan & Bansal, 2019) achieves very high accuracy in question type "Num", its accuracy is significantly lower on the rest of metrics. Category II lists the models with annotations. Compared with the VQA backbones in Category I, extra annotation can bring performance improvement in overall accuracy "All", "Yes/No" and "Other", while these models fail to outperform in question type "Num". Our LSP model outperforms all the models in Category II with extra annotations.

Category III list the typical ensemble-based and contrastive learning-based approaches under UpDn backbone, and category IV lists other VQA debiasing model with different backbones. The debiasing mechanisms in both Category III and Category IV all show effectiveness compared with models in category I without debiasing and category II with extra annotation. Comparing to other debiasing mechanisms in both Category III and Category IV, our LSP model outperforms in all metrics significantly, and achieves comparable result in the question type "Num" with ECD (Kolling et al., 2022) model. Compared the results on VQA v2 dataset and VQA-CP v2, unlike other debiasing models, our LSP model exhibits good generalization characteristics on VQA-CP v2, while maintaining promising results on the original VQA v2 dataset.

Table 2
Ablation study for components (NSR, NSQ, Prompt) in our LSP model on VQA-CP v2 dataset. ✓ denotes adding the component. The baseline model without all three components is UpDn (Anderson et al., 2018).

NSR	NSQ	Prompt	All	Yes/No	Num	Other
			40.63	41.27	13.63	47.70
✓			58.41	88.54	32.68	49.68
✓	✓		59.83	88.89	40.36	49.48
✓	✓	✓	61.95	89.50	52.44	50.12

Table 3
Ablation study for various backbones. LXMERT (Tan & Bansal, 2019) is a pre-trained model, and we report the results of the latest model SAR (Si et al., 2021) conditioned on LXMERT for comparison.

Models	All	Yes/No	Num	Other	Δ ↑
S-MRL (Cadene et al., 2019)	40.11	55.96	9.98	40.08	**+12.60**
+LSP	52.71	80.88	15.75	48.09	
SAN (Yang et al., 2016)	40.24	41.81	13.13	46.85	**+11.11**
+LSP	51.35	79.42	11.18	47.66	
BAN (Kim et al., 2018)	33.16	41.67	13.11	34.19	**+10.02**
+LSP	43.18	56.33	18.59	43.04	
LXMERT (Tan & Bansal, 2019)	42.84	18.91	55.51	46.23	-
+LSP	71.06	86.56	59.01	66.24	**+28.22**
+SAR(Top20) (Si et al., 2021)	65.44	83.13	54.52	59.16	**+22.60**
+SAR+LMH(Top20) (Si et al., 2021)	66.73	86.00	62.34	57.84	**+23.89**

4.5. Ablation study

4.5.1. Effect of each model component

In this section, we examine the effectiveness of each component in our LSP model by adding each component gradually. We conduct this ablation on VQA-CP v2 dataset and the results of four metrics are reported in Table 2. As can be observed from Table 2, selective sampling rate in NSR improve the performance over baseline model UpDn (Anderson et al., 2018) by a large margin in all the metrics. The question type constraint sampling in NSQ further improves all the metric performance, especially in question type "Num" with an increase by 7.68%. By adding the question type-guided prompt with designed learnable context, the model performance in all metrics can be further improved.

To further explore the details of our proposed model LSP, we conduct extra ablation experiments to answer the following questions:

- Q1: Can the proposed model improve the state-of-the-art VQA backbones?
- Q2: How do parameter settings affect the model performance?
- Q3: Does our model can perform on the OOD dataset without retraining on new data?

***Q1**: Can the proposed model improve the state-of-the-art VQA backbones?* To demonstrate the effectiveness of our proposed model on various VQA backbones, we conduct extensive experiments using extra three prevalent backbones, *i.e.*, S-MRL (Cadene et al., 2019), SAN (Yang, He, Gao, Deng, & Smola, 2016), BAN (Kim, Jun, & Zhang, 2018) and LXMERT (Tan & Bansal, 2019), on VQA-CP v2 (Agrawal et al., 2018), The detailed results on four metrics are reported in Table 3. Before delving into the details of the results in Table 3, we first recap the basic VQA backbones.

- S-MRL (Cadene et al., 2019) utilize an extra question-only branch to capture the language biases and then dynamically adjust the answer prediction process to compensate for biases.
- SAN (Yang et al., 2016) is a typical multi-layer stacked neural network with designed attention modules. Its main motivation is to reason on the question like human beings and provide the answers progressively.
- BAN (Kim et al., 2018) is built on bilinear attention distributions, considering the channel interactions between visual and linguistics information.
- LXMERT (Tan & Bansal, 2019) is built on the Transformer blocks and pre-trained on large-scale images and questions.

From the results reported in Table 3, we can observe that our proposed LSP is able to improve the overall performance ("All" metric) by at least about 10% regardless of the backbones. Besides, we can observe that our model significantly improves the performance of the pre-trained backbone LXMERT (Tan & Bansal, 2019). Surprisingly, we also find that our model brings more significant improvements compared to the recently proposed SAR (Si, Lin, yu Zheng, Fu, & Wang, 2021) with LXMERT backbone (28.22 *vs.* 23.69). These results demonstrate that our approach is model-agnostic and can improve the VQA backbones' performance substantially.

Table 4
Ablation study for the negative image sampling. The sampling rate is set to 0.85 in our LSP model. The baseline model without image sampling is UpDn (Anderson et al., 2018). Δ denotes the magnitude of the improvement (↑). The same notion as below.

Models	All	Yes/No	Num	Other	Δ ↑
Baseline	40.63	41.27	13.63	47.70	–
+ SSL-VQA (Zhu et al., 2020)	57.59	86.53	29.87	50.03	+16.96
+ NSR (Ours)	58.41	88.54	32.68	49.68	**+17.78**

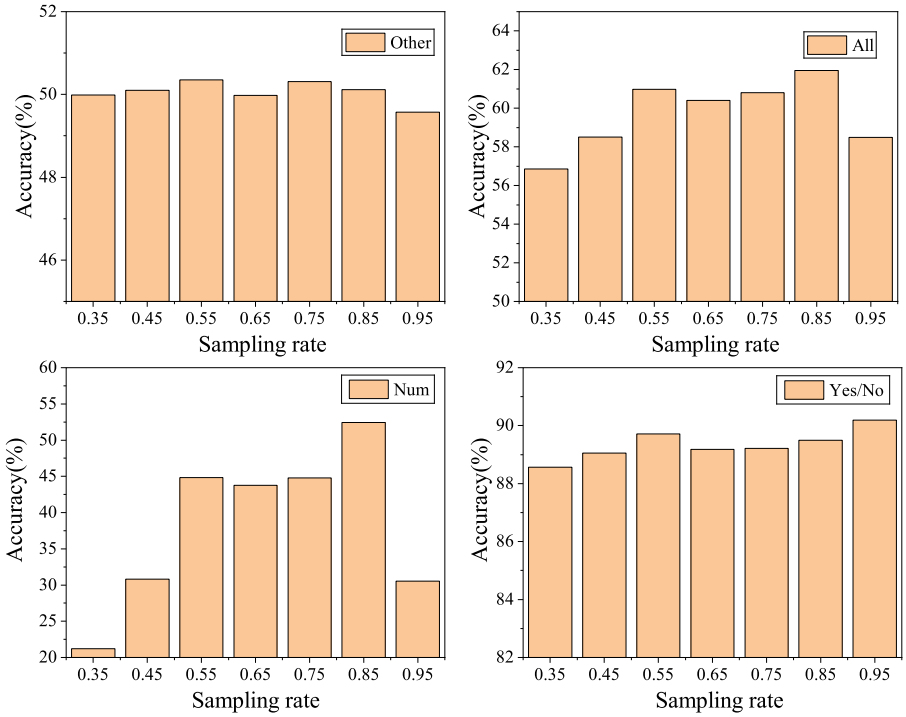

Fig. 5. Results of different sampling rate on "All" and "Other" metrics.

***Q2**: How do parameter settings affect the model performance?* In order to explore the effect of model performance on different parameter settings, we conduct extensive analysis of the selective sampling rate, question type-guided sampling, different prompt modes and prompt lengths.

(1) The effect of sampling rate in NSR. To demonstrate the effect of the selective sampling rate, we conduct the experiments on VQA-CP v2 dataset and set the sampling rate to 0.85 comparing with SSL-VQA (Zhu et al., 2020) which samples a negative image for each positive image–question pair in a exhaustive sampling. The results are reported in Table 4 and reveal that, our model with NSR obtains comparable performance in "Other" and outperforms SSL-VQA (Zhu et al., 2020) on the remaining metrics.

Further, we compare our model performance with different selective sampling rates, ranging from 0.35 to 0.95, with step 0.1. Two competitive results of "All" and "Other" metrics are selected and shown in Fig. 5. From the results, we can find that when the selective sampling rate is set to 0.85, the model can achieve the best performance on the "All" metric, while the best performance on the "Other" metric is with selective sampling rate of 0.55. Besides, to balance the performance, we set the selective sampling rate to 0.85 on all experiments.

(2) Effect of question type-guided sampling in NSQ. To further verify the effectiveness question type-guided sampling, we compare our model on VQA-CP v2 dataset with its random sampling version and several strong models, *i.e.*, ECD (Kolling et al., 2022), CSS (Chen et al., 2020) and D-VQA (Wen et al., 2021), which sample one negative question from the training dataset for each image. The results are reported in Table 5 and show that, our model with NSQ obtains 3.49% higher accuracy in "All" compared with baseline model UpDn (Anderson et al., 2018) and outperforms other models. Additionally, we can also find that question type-guided sampling is superior to the random question sampling by about 3% higher overall accuracy. The above results also embody the effectiveness of question type-guided sampling.

(3) The effect of different prompt mode. To demonstrate the effectiveness of the prompt learning, we investigate three prompt modes, *i.e.*, fixed prompt mode (prompts are fixed words), learnable prompt mode (prompts are the vectors that need to learn) and mixed

Table 5
Ablation study for the negative question sampling. Our LSP model adopts the question type-guided sampling in NSQ, while the other methods utilize the question sampling. The UpDn (Anderson et al., 2018) model is selected as the baseline.

Models	All	Yes/No	Num	Other	Δ ↑
Baseline	40.63	41.27	13.63	47.70	–
+ ECD (Kolling et al., 2022)	40.69	41.71	13.41	47.63	+0.06
+ CSS (Chen et al., 2020)	40.05	42.16	12.30	46.56	−0.58
+ D-VQA (Wen et al., 2021)	41.40	43.33	12.97	48.19	+0.77
+ NSQ (ours)	44.12	50.92	13.27	49.02	**+3.49**
+ NSQ (random)	41.00	42.81	13.29	47.44	+0.37

Table 6
Ablation study for various fixed prompt mode. The model without prompt is selected as the baseline. The type is replaced by the corresponding question type text and the answer is a fixed text prompt. The same as below.

Models	All	Yes/No	Num	Other
baseline†	59.83	88.89	40.36	49.48
[Q]+type	60.13	88.60	43.15	49.88
type+[Q]	58.27	87.63	35.06	49.26
[Q]+answer:	59.74	90.08	35.78	50.41

Table 7
Ablation study for various learnable prompt mode. The context denotes the designed learnable prompt context. The same as below.

Models	All	Yes/No	Num	Other
baseline†	59.83	88.89	40.36	49.48
[Q]+contex	60.49	87.77	48.09	49.60
context+[Q]	60.26	89.10	48.05	48.50

Table 8
Ablation study for various mixed prompt mode.

Models	All	Yes/No	Num	Other
baseline†	59.83	88.89	40.36	49.48
[Q]+context+type	61.03	89.99	46.77	49.76
[Q]+type+context	61.95	89.50	52.44	50.12
context+[Q]+context+type	60.62	89.86	48.58	48.61
[Q]+answer+context	60.85	89.52	48.52	49.21

prompt mode (prompts are mixed by fixed words and learnable context). The results are reported in Table 6, Table 7 and Table 8, respectively.

As can be observed from Table 6, only the model with the fixed prompt mode "[Q]+type" outperforms the baseline method without prompt, while the mode "type+[Q]" obtains the worst results. As for the results of learnable prompt in Table 7, both modes can improve the model performance of the baseline. Moreover, we can observe the fact from Table 8 that all the mixed prompt mode can further improve the model performance compared with the results in Tables 6 and 7. In this paper, we select the best mode "[Q]+type+context" as the prompt mode in LSP model on the whole experiments.

(4) The effect of different length of prompt. Prompt design in our model with different length can acquire different useful context from the question. Thus, to investigate the performance of various lengths of the prompt, we set the length from 3 to 8 conditioned on the prompt mode "[Q]+type+context". The results are reported in Fig. 6. We can observe that, when the length is 5, the model achieves the best performances in terms of the question type "All", "Num" and "Other" and obtains the second best results in the question type "Yes/No". In this paper, we set the prompt length to 5 on all experiments.

(5) The effect of different loss weight w values. The loss weight w is utilized to balance the model and loss function. Therefore, to investigate the performance of various values of the loss weight w, we set it from 1 to 5. Two competitive results of All and Other metrics are selected and shown in Figure Fig. 7. From the results, we can find that when the loss weight w is set to 3 the model can achieve the best performance on the both metrics. Besides, we set the loss weight w to 3 on all experiments.

***Q3**: Does our model can perform on the OOD dataset without retraining on new data?* Recently, several approaches have pointed out that existing methods overuse the knowledge of how the dataset is constructed and drops significantly on the out-of-distribution (OOD) dataset without retraining on new data (Jiang, Liu, Liu, Nan, & Zheng, 2021; Teney et al., 2020). Therefore, following the

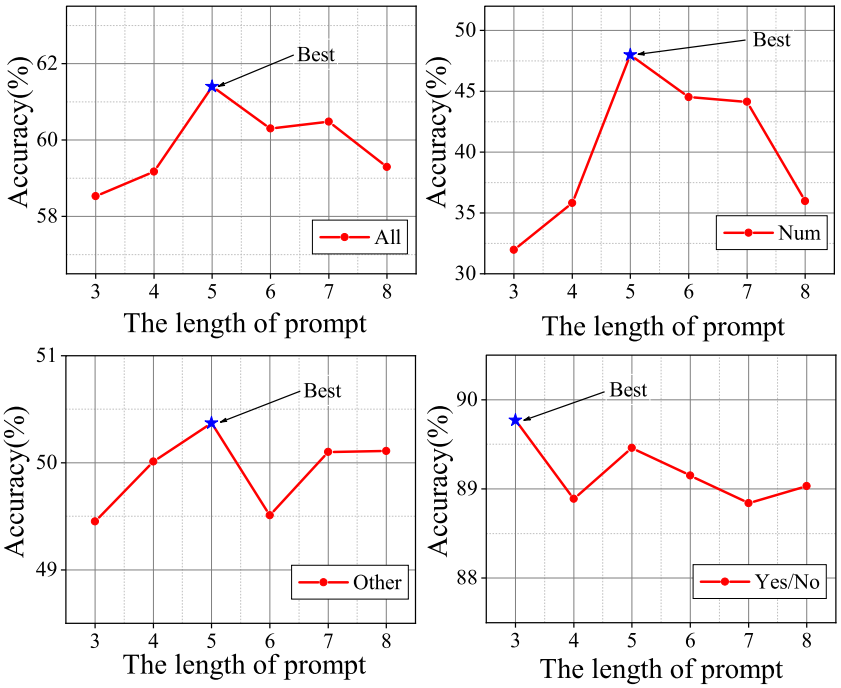

Fig. 6. Ablation study for different prompt length conditioned on [Q]+type+context mode. The blue star mark denotes the best results.

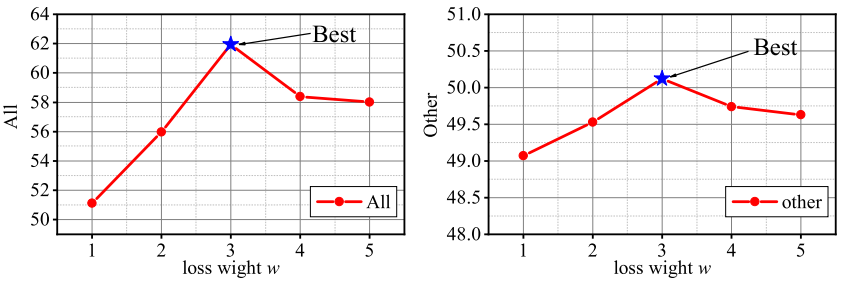

Fig. 7. Ablation study for Different Loss Weight w Values. The blue star mark denotes the best results.

same dataset setting on the re-splited VQA-CP v2 in (Jiang et al., 2021), we verify the model generalization performance on the OOD split as well as the in-distribution (ID) split. The results are reported in Table 9.

As can be observed from Table 9, our proposed model can outperform previous state-of-the-art models on ID split and maintain promising performance on the OOD split. In particular, our model drops slightly on the “All" metric on the OOD split compared with the LXMERT (Tan & Bansal, 2019), which is pre-trained by large image and question pairs. In contrast, our model outperforms other debiasing models (e.g., GGE (Han et al., 2021)) by a large margin. These results also verify that our model generalized to other different OOD settings as well as maintaining ID performance.

4.5.2. Qualitative analysis

In this section, we illustrate qualitative examples from visual shortcut bias, language distribution bias and failure cases in Fig. 8. From the example of visual shortcut bias to the left of the blue line, we can observe that, for left most column, state-of-the-art model SSL-VQA (Zhu et al., 2020) provide the wrong answer “no" due to the visual shortcut bias, which also embodies the exhaustive image sampling can result in wrong grounding and answer. In addition, although both models provide the correct answers for the question as shown in the second left-most column, our model can utilize the right visual grounding (the pack animal) for predicting the correct answer, while SSL-VQA (Zhu et al., 2020) tends to focus on the pack area (*i.e.*, visual shortcut area). The second example comes from the language distribution bias between the orange line and the blue line. UpDn (Anderson et al., 2018) correctly grounds the birds and chairs, while they still tend to answer the questions based on language distribution bias. In contrast, our model can correctly answer the questions with the right visual grounding.However, our model still suffers failure, as shown in the example to the right of the orange line shown in the figure. We can observe that, although LSP obtains reasonable visual groundings for “snow"

Table 9
The comparison results on VQA-CP v2 OOD test set and ID test set between our model and other debiasing models in terms of Accuracy (%). The **best** and second results are highlighted in bold and underline. Models with † represents the replicated results using released codes. Other results are reported in the original papers.

Models	VQA-CP v2 (OOD Split)				VQA-CP v2 (ID Split)			
	All	Yes/No	Num	Other	All	Yes/No	Num	Other
Unshuffling (Teney, Abbasnejad, & van den Hengel, 2021)	42.39	47.72	14.43	47.24	–	–	–	–
UpDn (Anderson et al., 2018)	38.82	42.98	12.18	43.95	64.73	79.45	49.59	55.66
AdvReg. (Grand & Belinkov, 2019)	36.33	59.33	14.01	30.41	50.63	67.39	38.81	38.37
GRL. (Grand & Belinkov, 2019)	42.33	59.74	14.78	40.76	56.90	69.23	42.50	49.36
RandImg (Teney et al., 2020)	51.15	75.06	24.30	45.99	59.28	70.66	43.06	53.40
X-GGM (Jiang et al., 2021)	45.71	43.48	27.65	52.34	67.16	83.74	48.26	56.91
GGE† (Han et al., 2021)	51.98	81.66	12.91	47.14	65.29	64.65	52.11	69.19
LXMERT (Tan & Bansal, 2019)	**63.90**	80.45	46.58	**59.98**	75.57	91.39	**59.83**	65.21
ours	61.91	**89.56**	**50.05**	52.28	**77.86**	**92.00**	57.49	**70.19**

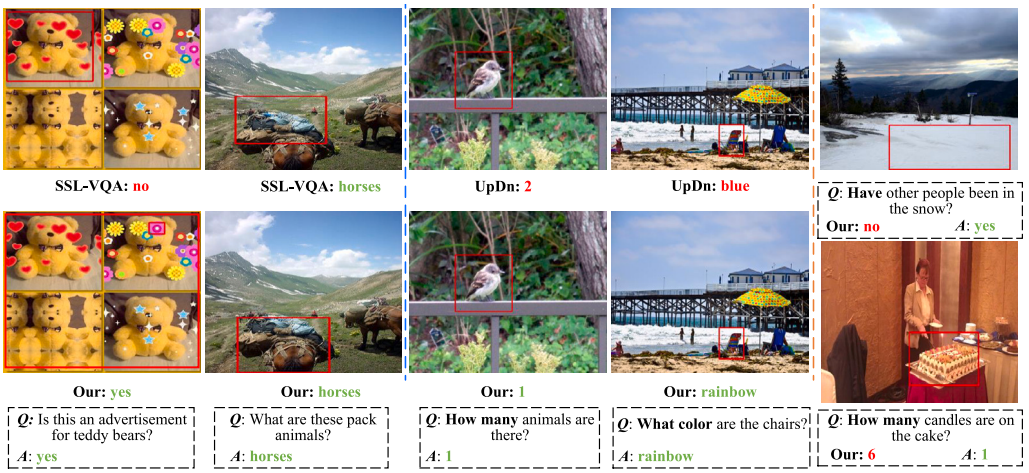

Fig. 8. Qualitative comparison between state-of-the-art methods and our LSP Model in two biases, *i.e.*, visual shortcut bias (To the left of the blue line), language distribution bias (Between the orange line and the blue line), and failure cases (To the right of the orange line). The ground-truth and correct answers are highlighted in green, while the wrong in red. The predicted visual region is shown in red rectangle on the image.

and "cake", it failed to answer the question. This may be due to LSP cannot disentangle more fine-grained grounding information, *i.e.*, footprints on the snow and one thin candle on the cake. The results also reveal that our model relies on the extracted features, leading to losing attention on small objects easily, and we leave it for our future work.

5. Model's pros and cons

As can be observed from the experimental results, we can find that our proposed LSP achieves new state-of-the-art performance on debiasing visual question answering (VQA) tasks and overcomes the visual shortcut and language distribution biases. Unlike existing approaches, we consider removing two biases on the out-of-distribution dataset (VQA-CP v2 (Agrawal et al., 2018)) and maintaining the performance on the in-distribution dataset (VQA v2 (Goyal et al., 2017)) without human annotations. Further, our LSP is a non-transformer model that utilizes the prompt without any pretraining compared with different pre-trained VQA models, e.g., LXMERT (Tan & Bansal, 2019). However, there still exist some potential limitations to our model. Firstly, the proposed LSP is built upon the extracted image and question features, which may need more works on the good quality of pre-extracted features from other visual and language models. Secondly, due to the size of pre-extracted features are indeed huge, the model will take a certain amount of time to load the features before the training process. We will study the above two problems in future work. One possible solution for the quality of the extracted features is that we can introduce a multi-modal fusion framework to fuse the

extracted features from multi-sources, e.g., Fast-RCNN (Ren et al., 2015) and YoloV5 (Tan, Lu, Jiang, & Huang, 2021). As for the second issue, we can reform the code structure and develop more compression approaches for data-efficient data loading.

6. Conclusion

In this paper, we propose a debiasing model for robust VQA by Learning to Sample and Prompt to overcome visual shortcut bias and language distribution bias, namely LSP. In specific, we introduce the selective sampling rate in the process of negative image sampling to balance the modality utilization of images and questions. Besides, we further design question type-guided sampling to reduce the question sampling space. Then we incorporate the learnable question type-guided prompt into the question for better question comprehension and explore the effectiveness of various prompt settings. Extensive experiments on two benchmark VQA-CP v2 and VQA v2 demonstrate that our LSP model achieves new state-of-the-art and provide the correct answer with right visual grounding. In the future, we will extend our model to more biased datasets in other domain, such as image classification and face recognition.

CRediT authorship contribution statement

Jin Liu: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **ChongFeng Fan:** Data curation, Software. **Fengyu Zhou:** Project administration. **Huijuan Xu:** Supervision, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by The National Key R and D Program of China (Grant No. 2017YFB1302400), Jinan '20 New Colleges and Universities' Funded Scientific Research Leader Studio, China (2021GXRC079), Major Agricultural Applied Technological Innovation Projects of Shandong Province, China (SD2019NJ014), Shandong Natural Science Foundation, China (ZR2019MF064), Beijing Advanced Innovation Center for Intelligent Robots and Systems, China (2019IRS19). In addition, the authors thank the anonymous reviewers for providing valuable comments to improve this paper.

References

Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. (2018). 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* (pp. 4971–4980).

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077–6086).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., et al. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).

Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. (2019). Rubi: Reducing unimodal biases for visual question answering. (p. 1).

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., & Zhuang, Y. (2020). Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10800–10809).

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734).

Clark, C., Yatskar, M., & Zettlemoyer, L. (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*.

Ding, N., Chen, Y., Han, X., Xu, G., Xie, P., Zheng, H.-T., et al. (2021). Prompt-learning for fine-grained entity typing. (p. 1). arXiv preprint arXiv:2108.10604.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). 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* (pp. 6904–6913).

Grand, G., & Belinkov, Y. (2019). Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the second workshop on shortcomings in vision and language* (pp. 1–13).

Guo, Y., Nie, L., Cheng, Z., Tian, Q., & Zhang, M. (2021). Loss re-scaling VQA: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing*, 227–238.

Han, X., Wang, S., Su, C., Huang, Q., & Tian, Q. (2021). Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1584–1593).

Jiang, J., Liu, Z., Liu, Y., Nan, Z., & Zheng, N. (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* (pp. 199–208).

Jin, W., Cheng, Y., Shen, Y., Chen, W., & Ren, X. (2022). A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 2763–2775).

Jing, C., Wu, Y., Zhang, X., Jia, Y., & Wu, Q. (2020). Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11181–11188).

Kim, J. H., Jun, J., & Zhang, B. T. (2018). Bilinear attention networks. (p. 1).

Kolling, C., More, M., Gavenski, N., Pooch, E., Parraga, O., & Barros, R. C. (2022). Efficient counterfactual debiasing for visual question answering. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3001–3010).

Liang, Z., Hu, H., & Zhu, J. (2021). LPF: A language-prior feedback objective function for de-biased visual question answering. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1955–1959).

Luo, H., Lin, G., Yao, Y., Liu, F., Liu, Z., & Tang, Z. (2022). Depth and video segmentation based visual attention for embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Ramakrishnan, S., Agrawal, A., & Lee, S. (2018). Overcoming language priors in visual question answering with adversarial regularization. (p. 1).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. (p. 1).

Selvaraju, R. R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., et al. (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* (pp. 2591–2600).

Shrestha, R., Kafle, K., & Kanan, C. (2020). A negative case analysis of visual grounding methods for VQA. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8172–8181).

Si, Q., Lin, Z., yu Zheng, M., Fu, P., & Wang, W. (2021). Check it again: Progressive visual question answering via visual entailment. 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)* (pp. 4101–4110).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 1929–1958.

Tan, H., & Bansal, M. (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* (pp. 5100–5111).

Tan, S., Lu, G., Jiang, Z., & Huang, L. (2021). Improved YOLOv5 network model and application in safety helmet detection. In *2021 IEEE international conference on intelligence and safety for robotics* (pp. 330–333).

Teney, D., Abbasnejad, E., & van den Hengel, A. (2021). Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1417–1427).

Teney, D., Abbasnejad, E., Kafle, K., Shrestha, R., Kanan, C., & Van Den Hengel, A. (2020). On the value of out-of-distribution testing: An example of Goodhart's law. (pp. 407–417).

Wang, J., Pradhan, M. R., & Gunasekaran, N. (2022). Machine learning-based human-robot interaction in ITS. *Information Processing & Management*, Article 102750.

Wen, Z., Xu, G., Tan, M., Wu, Q., & Wu, Q. (2021). Debiased visual question answering from feature and sample perspectives. (p. 1).

Wu, J., & Mooney, R. (2019). Self-critical reasoning for robust visual question answering. (p. 1).

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21–29).

Yang, A., Miech, A., Sivic, J., Laptev, I., & Schmid, C. (2022). Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1.

Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T. S., & Sun, M. (2021). Cpt: Colorful prompt tuning for pre-trained vision-language models. (p. 1). arXiv preprint arXiv:2109.11797.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and Yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5014–5022).

Zheng, C., & Huang, M. (2021). Exploring prompt-based few-shot learning for grounded dialog generation. (p. 1). arXiv preprint arXiv:2109.06513.

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2337–2348.

Zhu, X., Mao, Z., Liu, C., Zhang, P., Wang, B., & Zhang, Y. (2020). Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 1083–1089).