Supplementary Materials for "Counterfactual VQA: A Cause-Effect Look at Language Bias"

This supplementary document is organized as follows:

- Section 1 introduces that RUBi [5] and Learned-Mixin [6] can be unified into our counterfactual inference framework.
- Section 2 provides an analysis of estimating NDE using the learnable parameter.
- Section 3 describes the implementation details.
- Section 4 describes the supplementary quantitative and qualitative results.

1. Revisiting RUBi and Learned-Mixin

As mentioned in Section 4.3, RUBi [5] and Learned-Mixin [6] can be unified into our counterfactual inference framework, which (1) follow a simplified causal graph without the direct path $V \rightarrow A$, and (2) use natural indirect effect (NIE) for inference. The detailed analysis is provided as follows.

1.1. Cause-Effect Look

Recent works RUBi [5] and Learned-Mixin [6] apply an ensemble architecture with a vision-language branch $\mathcal{F}_{\mathcal{VQ}}$ and a question-only branch $\mathcal{F}_{\mathcal{Q}}$, while the direct relation between vision and answer is not formulated. The architecture is shown in Figure 1 (a).

Note that total effect can be decomposed into natural direct effect (NDE) and total indirect effect (TIE). As introduced in the main paper, we remove language bias by subtracting the natural direct effect from the total effect. The TIE is calculated by:

$$\begin{aligned} TE &= Z_{q,k} - Z_{q^*,k^*}, \\ NDE &= Z_{q,k^*} - Z_{q^*,k^*}, \\ TIE &= TE - NDE = Z_{q,k} - Z_{q,k^*}, \end{aligned} \tag{1}$$

which corresponds to Eq. (4) in the main paper. An alternative option to reduce language bias is to substract the total direct effect (TDE) of questions on answers from total effect, which is formulated as:

$$\begin{aligned} TDE &= Z_{q,k} - Z_{q^*,k}, \\ NIE &= TE - TDE = Z_{q^*,k} - Z_{q^*,k^*}. \end{aligned} \tag{2}$$

Intuitively, both TIE and NIE reflect the increase of confidence for the answer given the visual knowledge, *i.e.*, from k^* to k. The difference between TIE and NIE is the existence of question q. The question q is block to calculate NIE (*i.e.*, q^*), while q is given to calculate TIE. We use TIE to reserve q as the language context. In addition, both TDE and NDE reflect the increase of confidence for the answer given the question, *i.e.*, from q^* to q. The difference between TDE and NDE is also the existence of question q. Note that we hope to exclude the effect directly caused by question. Therefore, the mediator knowledge should be blocked when estimating the pure language effect, which is captured by NDE.

1.2. Implementation

RUBi [5] and Learned-Mixin (LM) [6] use the following fusion strategies for ensemble-based training:

$$\text{(RUBi)} \qquad h(Z_q, Z_k) = Z_k \cdot \sigma(Z_q) \tag{3}$$

$$\text{(LM)} \quad h(Z_q, Z_k) = \log\sigma(Z_k) + g(k)\cdot\log\sigma(Z_q) \tag{4}$$

where $\sigma(\cdot)$ represents the sigmoid function, and $g(\cdot)$ is a learned function $\mathbb{R}^d \rightarrow \mathbb{R}^1$ with the knowledge representation $k \in \mathbb{R}^d$ as input and a scalar weight as output. During the test stage, they use Z_k for inference.

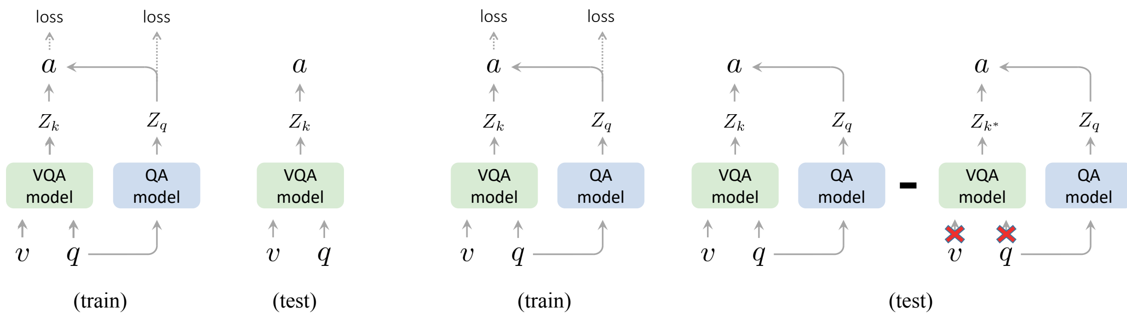
Perhaps supering As for RUBi, NIE is calculated as:

$$NIE = \underbrace{z_k \cdot \sigma(c)}_{Z_{q^*,k}} - \underbrace{c \cdot \sigma(c)}_{Z_{q^*,k^*}} \propto z_k \tag{5}$$

As for Learned-Mixin, NIE is calculated as:

$$\begin{aligned} NIE = &\underbrace{(\log\sigma(z_k) + g(k)\cdot\log\sigma(c))}_{Z_{q^*,k}} \\ &- \underbrace{(\log\sigma(c) + g(k^*)\cdot\log\sigma(c))}_{Z_{q^*,k^*}} \propto z_k \end{aligned} \tag{6}$$

where c, $g(k)$ and $g(k^*)$ are constants for the same sample. Therefore, we have $NIE \propto z_k$ for both RUBi and Learned-Mixin, which is exactly the output score of the



(a) language-prior based VQA

(b) CF-VQA

Figure 1: Comparison between our CF-VQA and language-prior based methods [5, 6] based on the simplified causal graph.

Algorithm 1 Improving RUBi [5] using CF-VQA

function RUBI(v, q, is_Training; θ, c)
  $z_q = \mathcal{F}_Q(q)$
  $z_k = \mathcal{F}_{VQ}(v,\ q)$
  if is_Training **then**
    $z = z_k \cdot \sigma(z_q)$
    updating θ according to $\mathcal{L}_{cls}$
    updating c according to $\mathcal{L}_{kl}$
  else
    $\cancel{z = z_k}$ $z = (z_k - c) \cdot \sigma(z_q)$
  end if
  return z
end function

vision-language branch $\mathcal{F}_{VQ}$. Note that RUBi and Learned-Mixin simply preserve the vision-language branch and uses z_k for inference. From our cause-effect view, *RUBi and Learned-Mixin use natural indirect effect for inference*.

1.3. Improving RUBi [5]

Thanks to our cause-effect look, RUBi [5] can be improved using CF-VQA, *i.e.*, using TIE for inference. Specifically, TIE for RUBi is calculated as:

$$TIE = \underbrace{z_k \cdot \sigma(z_q)}_{Z_{q,k}} - \underbrace{c \cdot \sigma(z_q)}_{Z_{q,k^*}} \tag{7}$$

where c denotes a learnable parameter. Table 5 in the main paper demonstrates that CF-VQA can outperform RUBi by 7% on VQA-CP v2. The red notes in Algorithm 1 show how RUBi is improved by changing several lines of code.

2. Analysis of Estimating NDE

In Section 4.2 in the main paper, we claimed that the learnable parameter c controls the sharpness of Z_{q,v^*,k^*} for estimating NDE. We give an intuitive analysis here.

For Harmonic (HM), we have:

$$\text{(HM)} \qquad Z_{q,v^*,k^*} = \log \frac{\sigma(z_q) \cdot c_{\text{HM}}}{1 + \sigma(z_q) \cdot c_{\text{HM}}}, \tag{8}$$

where $c_{\text{HM}} = (\sigma(c))^2 \in (0, 1)$. We approximate the limits of Z_{q,v^*,k^*} and $TIE = Z_{q,v,k} - Z_{q,v^*,k^*}$ as:

$$\text{(HM)} \qquad \begin{aligned} \lim_{c_{\text{HM}} \to 0} Z_{q,v^*,k^*} &= -\infty \\ \lim_{c_{\text{HM}} \to 0} TIE \quad &= z_{q,v,k} - C \\ &\propto z_{q,v,k}, \end{aligned} \tag{9}$$

where we use a extremely negative number C to replace $-\infty$ for valid estimation of TIE. In this case, NDE is estimated as the same constant for all the answers, and TIE is dominated by $z_{q,v,k}$, which means that the language bias is not reduced. For $c_{\text{HM}} \to 1$, we have

$$\text{(HM)} \qquad \begin{aligned} \lim_{c_{\text{HM}} \to 1} Z_{q,v^*,k^*} &= \log \frac{\sigma(z_q)}{1 + \sigma(z_q)} \\ \lim_{c_{\text{HM}} \to 1} TIE \quad &= \log \frac{\sigma(z_v) \cdot \sigma(z_k) \cdot (1 + \sigma(z_q))}{1 + \sigma(z_q) \cdot \sigma(z_v) \cdot \sigma(z_k)}. \end{aligned} \tag{10}$$

For SUM, we have

$$\text{(SUM)} \qquad Z_{q,v^*,k^*} = \log \sigma(z_q + 2c), \tag{11}$$

where $c \in (-\infty, +\infty)$. We approximate the limits of Z_{q,v^*,k^*} and $TIE = Z_{q,v,k} - Z_{q,v^*,k^*}$ as:

$$\text{(SUM)} \qquad \begin{aligned} \lim_{c \to \ \infty} Z_{q,v^*,k^*} &= -\infty \\ \lim_{c \to \ \infty} TIE \quad &= z_{q,v,k} - C \\ &\propto z_{q,v,k}. \end{aligned} \tag{12}$$

Similar to HM, TIE is dominated by $z_{q,v,k}$. For $c \to +\infty$, we have:

$$\text{(SUM)} \qquad \begin{aligned} \lim_{c \to +\infty} Z_{q,v^*,k^*} &= 0 \\ \lim_{c \to +\infty} TIE \quad &= z_{q,v,k}. \end{aligned} \tag{13}$$

Table 1: **Comparison on VQA-CP v2 val set**. “Base.” indicates the VQA base model.

		VQA-CP v2					VQA v2
		val (in-domain)				test (OOD)	val (in-domain)
	Base.	All	Y/N	Num.	Other	All	All
GRLS [8]	–	56.90	69.23	42.50	49.36	42.33	51.92
GradSup[10]	–	62.4	77.8	43.8	53.6	46.8	–
RandImg [12]	UpDn	54.24	64.22	34.40	50.46	55.37	57.24
CF-VQA (HM)	UpDn	65.47	79.09	45.86	57.86	49.74	63.73
CF-VQA (SUM)	UpDn	60.29	66.32	47.48	57.96	51.27	62.49
CF-VQA (HM)	S-MRL	63.08	75.76	44.88	55.99	53.55	63.54
CF-VQA (SUM)	S-MRL	57.86	66.24	44.98	53.38	55.05	60.94

Also, TIE is dominated by $z_{q,v,k}$. In both cases, the language bias cannot be excluded. This analysis shows that a extremely large or small c will fail to estimate NDE and TIE, and it is necessary to control the sharpness of NDE by selecting a optimal c. In the main paper, we use a KL-divergence in Eq. (17) to force the sharpness of NDE similar to that of TE.

3. Implementation Details

We use the same implementation of RUBi [5] for fair comparison, including feature representation, baseline architectures, and optimization.

Image Representation. Following the popular bottom-up attention mechanism [2], we use a Faster R-CNN based framework to extract visual features. We select top-K region proposals for each image, where K is fixed as 36.

Question Representation. Following [4, 5], we first lowercase all the questions and remove the punctuation, and then use the pretrained Skip-thought encoder [9] with fine-tuning. The size of final embedding is set as 4800.

Vision-Language Branch. The vision-language branch consists of the image representation, question representation, and a visual knowledge encoder. The baseline models for encoding visual knowledge includes SAN [13], UpDn [2], and a simplified version of the recent architecture MUREL [4] (S-MUREL) proposed in [5]. In short, S-MUREL consists of a BLOCK [3] bilinear fusion between image and question representations for each region, and a MLP classifier composed of three fully connected layers with ReLU activations. The dimension are 2,048, 2,048, and 3,000. More details can be found in [5].

Language-Only Branch. The language-only branch consists of the question representation and a question-only classifier. The question-only classifier is implemented by a MLP with three fully connect layers with ReLU activations. Note that this MLP has the same structure with the classifier for vision-language branch with different parameters.

Vision-Only Branch. The vision-only branch is composed of the question representation and a vision-only classifier. The vision-only classifier has the same structure as the language-only classifier with different parameters.

Optimization. All the experiments are conducted with the Adam optimizer for 22 epochs. The learning rate linearly increases from $1.5 \times 10^{\ 4}$ to $6 \times 10^{\ 4}$ for the first 7 epochs, and decays after 14 epochs by multiplying 0.25 every two epochs. The batch size is set as 256.

Datasets. The experiments are conducted on VQA-CP [1] and VQA [7] datasets. VQA-CP v1 and v2 are created by re-organizing the train and val splits of the VQA v1 and v2 datasets, respectively [1].

4. Supplementary Experimental Results

We have conducted the ablation study and compared CF-VQA with state-of-the-art methods in the main paper. In this section, we show supplementary experimental results.

4.1. Quantitative Results

As suggested by [12, 8, 10, 11], we further hold out 8,000 instances from the training set (*i.e*., VQA-CP v2 val) to measure the in-domain performance. Note that the results on VQA v2 val set also measure the in-domain performance. The results are given in Table 1. Compared to GRLS [8], all of our variants outperform GRLS by large margins for both in-domain and out-of-distribution (OOD) settings. Compared tp GradSup [10], CF-VQA (HM) achieves better results on both VQA-CP val set and test set. Compared to RandImg [12], CF-VQA (SUM) achieves competitive results on VQA-CP v2 test set, and outperforms RandImge on in-domain settings by over 3%. These results demonstrate that CF-VQA not only effectively reduces language bias, but also performs robustly.

Table 2 shows the ablation study on VQA-CP v1 test split. As shown in Table 2, CF-VQA is general to both *baseline VQA architectures* and *fusion strategies*, which is also demonstrated by the results on VQA-CP v2. Table 3 shows the ablation study on VQA-CP v1 test split using the simplified causal graph. Similarly, CF-VQA achieves significant improvement for different baseline VQA architectures and fusion strategies.

Table 2: **Ablation of CF-VQA** on VQA-CP v1 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [5, 6]. * represents the reproduced results.

	All	Y/N	Num.	Other
SAN*	32.50	36.86	12.47	36.22
Harmonic	49.29	72.73	**20.57**	37.51
+ CF-VQA	**52.06**	**80.38**	16.88	**38.04**
SUM	38.34	49.88	**15.82**	35.91
+ CF-VQA	**52.87**	**84.94**	14.85	**36.26**

	All	Y/N	Num.	Other
UpDn*	37.08	42.46	12.76	41.50
Harmonic	**55.75**	80.65	**24.72**	43.46
+ CF-VQA	55.16	**82.27**	16.14	**43.87**
SUM	52.78	78.71	14.30	42.45
+ CF-VQA	**57.39**	**88.46**	**14.80**	**43.61**

	All	Y/N	Num.	Other
S-MRL*	36.68	42.72	12.59	40.35
Harmonic	53.55	79.38	17.39	42.38
+ CF-VQA	**55.26**	**82.13**	**18.03**	**43.49**
SUM	49.44	76.49	16.23	35.90
+ CF-VQA	**57.03**	**89.02**	**17.08**	**41.27**

Table 3: **Ablation of CF-VQA with the simplified causal graph** on VQA-CP v1 test set. "SAN/UpDn/S-MRL" denotes the baseline VQA model. "HM/SUM" represents the strategies that train the ensemble model and test with only the vision-language branch following ensemble-based method [5, 6]. * represents the reproduced results.

	All	Y/N	Num.	Other
SAN*	32.50	36.86	12.47	36.22
Harmonic	46.83	66.64	19.45	38.13
+ CF-VQA	**54.48**	**83.73**	**22.73**	**38.15**
SUM	40.08	54.15	15.53	**35.95**
+ CF-VQA	**52.73**	**84.64**	**16.02**	35.75

	All	Y/N	Num.	Other
UpDn*	37.08	42.46	12.76	41.50
Harmonic	54.13	80.60	15.75	43.24
+ CF-VQA	**56.19**	**85.08**	**16.00**	**43.61**
SUM	51.20	74.70	13.61	42.94
+ CF-VQA	**56.80**	**87.76**	**13.89**	**43.25**

	All	Y/N	Num.	Other
S-MRL*	36.68	42.72	12.59	40.35
Harmonic	54.51	80.82	17.30	43.29
+ CF-VQA	**56.82**	**86.01**	**17.38**	**43.63**
SUM	52.54	78.42	16.77	**41.18**
+ CF-VQA	**57.07**	**89.28**	**17.39**	41.00

4.2. Qualitative Results

Figure 2 illustrates examples to show how CF-VQA improves RUBi by simply replacing natural indirect effect with total indirect effect for inference following Algorithm 1. The examples show that CF-VQA benefits from language context, *e.g.*, "large or small", "deep or shallow", and "real or a statue" in the first row. Some failure cases are shown in the last two rows. First, CF-VQA may tend to generate broad answers, *e.g.*, "houses" v.s "church", and "vegetables" v.s "peas". Second, CF-VQA may ignore visual content like traditional likelihood strategy. Therefore, there remains the challenge about how to balance visual understanding and language context.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 3

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3

[3] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109, 2019. 3

[4] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019. 3

[5] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in Neural Information Processing Systems*, 32:841–852, 2019. 1, 2, 3, 4

[6] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 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 4060–4073, 2019. 1, 2, 4

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 3

[8] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13, 2019. 3

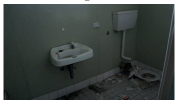
[9] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015. 3

[10] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020. 3

[11] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020. 3

[12] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On

Is this area large or small?

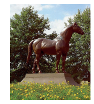


RUBi + CF-VQA		RUBi	
small	94.6%	old	36.3%
large	3.5%	yes	32.8%
big	1.5%	no	18.7%
medium	0.3%	both	5.8%
huge	0.1%	**small**	3.0%

Is this water deep or shallow?

RUBi + CF-VQA		RUBi	
deep	97.0%	no	40.3%
shallow	2.5%	yes	35.7%
yes	0.3%	unknown	12.1%
ascending	0.1%	expert	2.3%
choppy	0.1%	old	1.8%

Is this horse real or a statue?



RUBi + CF-VQA		RUBi	
statue	90.8%	horse	84.8%
toy	5.9%	no	13.0%
real	2.7%	none	1.0%
model	0.2%	neither	0.5%
big	0.1%	yes	0.4%

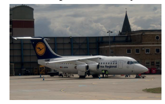
Is this natural or artificial light?

RUBi + CF-VQA		RUBi	
natural	50.4%	no	72.6%
yes	25.6%	yes	18.5%
both	7.7%	none	3.2%
curved	4.8%	old	1.3%
main	3.9%	unknown	1.3%

Is this bus going or coming?

RUBi + CF-VQA		RUBi	
going	78.4%	police	36.3%
stopped	8.6%	**going**	30.6%
coming	6.1%	forward	11.9%
leaving	1.1%	**coming**	6.2%
city	0.9%	no	4.3%

Is this airport in the city or country?



RUBi + CF-VQA		RUBi	
city	93.1%	New York	25.3%
country	6.2%	yes	22.7%
yes	0.5%	unknown	18.4%
both	0.1%	not sure	7.9%
United States	0.1%	no	6.6%

Is this fruit fresh or frozen?

RUBi + CF-VQA		RUBi	
fresh	63.9%	orange	25.7%
frozen	8.3%	red	15.7%
half	3.8%	none	14.3%
salt	2.2%	white	13.7%
wheat	1.6%	yellow	8.7%

What brand is on the coffee cup?

RUBi + CF-VQA		RUBi	
starbucks	89.6%	none	37.7%
dunkin donuts	5.1%	unknown	29.6%
coca cola	4.7%	can't tell	7.8%
coke	0.5%	not sure	4.2%
jones	0.1%	not possible	3.5%

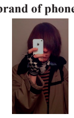
What brand of truck is shown?

RUBi + CF-VQA		RUBi	
ford	58.5%	no	29.2%
volvo	9.0%	yes	27.7%
toyota	8.8%	unknown	10.9%
dodge	5.4%	expert	6.2%
chevy	4.0%	old	5.0%

What brand is the man's shirt?

RUBi + CF-VQA		RUBi	
nike	60.1%	billabong	42.8%
adidas	12.9%	none	19.1%
billabong	10.4%	blue	12.1%
hurley	8.8%	white	6.6%
polo	3.3%	hurley	4.1%

What brand of phone is this?

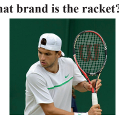


RUBi + CF-VQA		RUBi	
iphone	25.9%	nintendo	68.2%
motorola	24.5%	wii	26.3%
samsung	19.0%	none	4.3%
htc	10.2%	unknown	0.3%
apple	3.3%	**iphone**	0.3%

What brand of soda is on the nightstand?

RUBi + CF-VQA		RUBi	
coke	42.2%	none	45.2%
coca cola	38.9%	nothing	34.5%
coca-cola	8,9%	unknown	9.0%
starbucks	1.6%	not possible	3.3%
fanta	1.6%	white	1.5%

What brand is the racket?



RUBi + CF-VQA		RUBi	
wilson	50.2%	adidas	83.0%
nike	24.1%	**wilson**	6.5%
prince	9.1%	nike	4.8%
head	8.5%	w	2.8%
adidas	5.6%	white	0.8%

What brand is the box?

RUBi + CF-VQA		RUBi	
hp	32.2%	dell	64.2%
canon	21.5%	**hp**	26.5%
dell	15.1%	windows	6.9%
toshiba	13.4%	adidas	0.9%
head	8.0%	toshiba	0.6%

What brand is shown?

RUBi + CF-VQA		RUBi	
harley	55.7%	yamaha	34.8%
yamaha	20.6%	**harley**	33.9%
honda	7.8%	**harley davidson**	18.2%
toyota	6.8%	honda	4.6%
nike	2.8%	kawasaki	2.7%

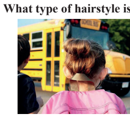
What type of herb is on the left?

RUBi + CF-VQA		RUBi	
parsley	52.1%	orange	73.7%
ginger	9.2%	red	18.6%
lily	6.0%	none	2.7%
pepper	5.7%	white	1.7%
maple	2.8%	yellow	0.9%

What type of seeds are stuck to the outside of the bun?

RUBi + CF-VQA		RUBi	
sesame	99.5%	unknown	41.7%
pepper	0.1%	none	30.1%
regular	0.1%	yellow	8.6%
sunflower	0.1%	**sesame**	6.6%
0	0.1%	white	2.9%

What type of hairstyle is the girl wearing?

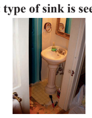


RUBi + CF-VQA		RUBi	
ponytail	19.5%	striped	26.2%
braid	18.7%	curly	14.1%
straight	11.4%	stripes	11.7%
long	11.0%	blue	10.0%
bob	8.6%	sweater	5.2%

What type of sneakers are the players playing in?

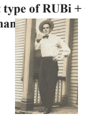
RUBi + CF-VQA		RUBi	
cleats	40.7%	baseball	92.5%
baseball	10.9%	**cleats**	6.1%
tennis shoes	8.6%	baseball cap	0.6%
converse	5.9%	don't know	0.1%
giants	5.7%	yes	0.1%

What type of sink is seen in the picture?



RUBi + CF-VQA		RUBi	
pedestal	49.6%	bathroom	48.5%
bathroom	10.3%	**pedestal**	40.0%
porcelain	8.0%	white	5.6%
ceramic	7.1%	ceramic	5.2%
white	4.1%	**porcelain**	1.0%

What type of RUBi + CF-VQA does this man



RUBi + CF-VQA		RUBi	
bow tie	74.7%	striped	50.6%
bow	15.1%	curly	21.1%
bowtie	8.0%	stripes	7.8%
regular	0.8%	blue	3.9%
horizontal	0.6%	sweater	3.5%

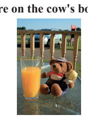
What type of flower is in the vase?

RUBi + CF-VQA		RUBi	
rose	31.0%	pink	69.2%
daisy	28.3%	**rose**	21.8%
carnation	16.0%	red	4.6%
lily	6.2%	**roses**	2.1%
lilly	3.0%	purple	0.8%

What type of building is pictured in the photo?

RUBi + CF-VQA		RUBi	
clock tower	28.8%	unknown	91.0%
church	15.1%	none	3.7%
apartment	9.5%	yellow	3.0%
school	7.9%	sesame	0.7%
tower	7.1%	white	0.5%

Where on the cow's body is there a tag?

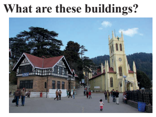


RUBi + CF-VQA		RUBi	
ear	48.8%	yes	76.8%
yes	17.3%	no	6.5%
back	9.3%	left	3.0%
head	5.0%	unknown	2.9%
legs	3.2%	bowl	2.5%

What are these machines used for?

RUBi + CF-VQA		RUBi	
money	39.2%	**parking**	90.0%
transportation	28.2%	money	8.6%
parking	13.4%	picture	0.6%
parking meter	3.6%	**parking meter**	0.2%
riding	2.0%	driving	0.1%

What are these buildings?

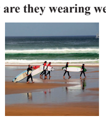


RUBi + CF-VQA		RUBi	
houses	26.8%	**church**	73.1%
skyscrapers	13.8%	city	8.4%
office	6.7%	building	5.5%
tower	6.4%	london	1.8%
church	6.2%	castle	1.6%

What are those round green things?

RUBi + CF-VQA		RUBi	
vegetables	27.3%	**peas**	97.3%
peas	19.5%	grapes	1.2%
grapes	10.2%	vegetables	1.2%
beets	6.6%	fruit	0.1%
fruit	5.9%	blueberries	0.1%

Why are they wearing wetsuits?



RUBi + CF-VQA		RUBi	
safety	67.9%	**surfing**	77.5%
surf	7.7%	surf	11.9%
yes	5.1%	yes	3.9%
protection	4.0%	sunny	2.1%
surfing	3.5%	walking	1.0%

Why are the people gathered?

RUBi + CF-VQA		RUBi	
parade	82.3%	**polo**	40.3%
riding horses	5.8%	rodeo	27.1%
jousting	2.3%	horseback riding	12.2%
race	2.1%	riding horses	9.3%
horseback riding	1.7%	jousting	7.2%

What kind of window covering is shown?

RUBi + CF-VQA		RUBi	
blinds	96.5%	**curtain**	16.8%
shade	2.9%	canopy	15.1%
sheet	0.3%	fan	14.9%
curtains	0.2%	**curtains**	7.9%
cloth	0.1%	sheet	7.7%

Figure 2: Qualitative comparison of RUBi and RUBi+CF-VQA on VQA-CP v2 test split. Red bold answer denotes the ground-truth one.

the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020. 3

[13] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 3