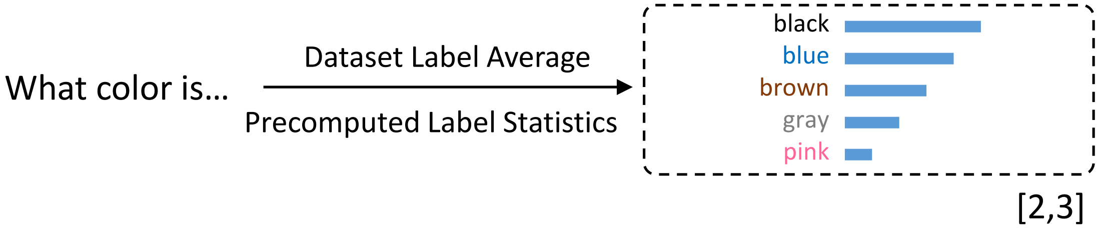
Generative Bias for Robust Visual Question Answering

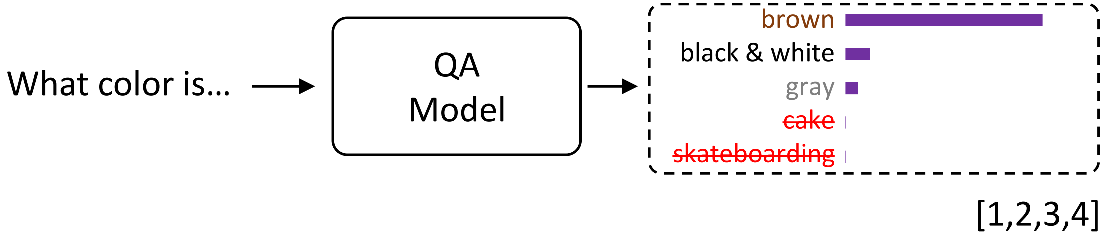
Jae Won Cho[1] Dong-Jin Kim[2] Hyeonggon Ryu[1] In So Kweon[1]

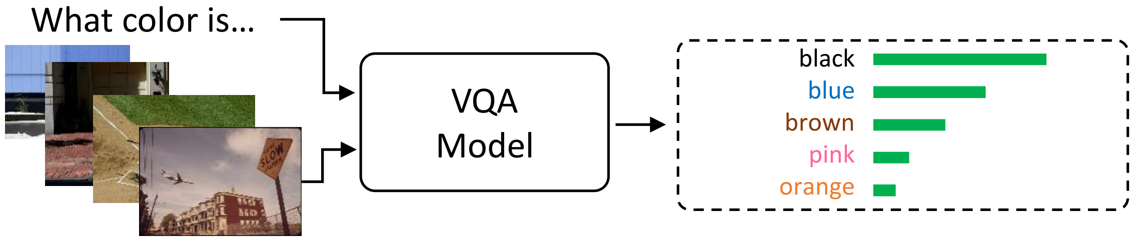
[1]KAIST, South Korea

[2]Hanyang University, South Korea

Issue of VQA Bias





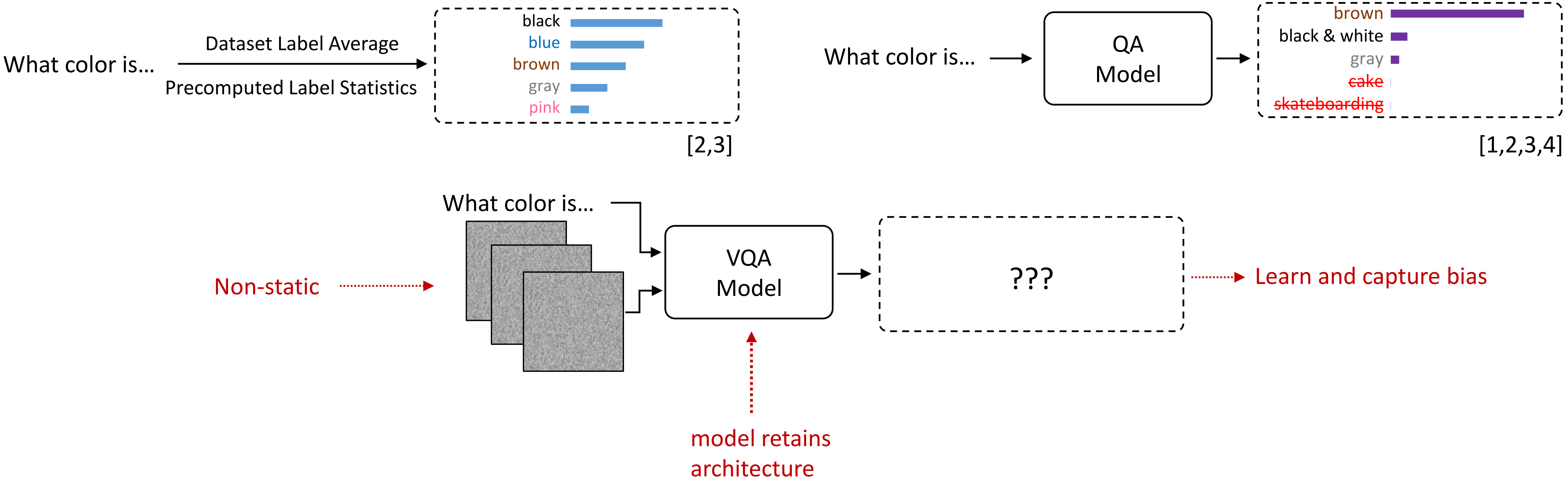


What the VQA model actually experiences

The better we can capture bias, the better we can debias

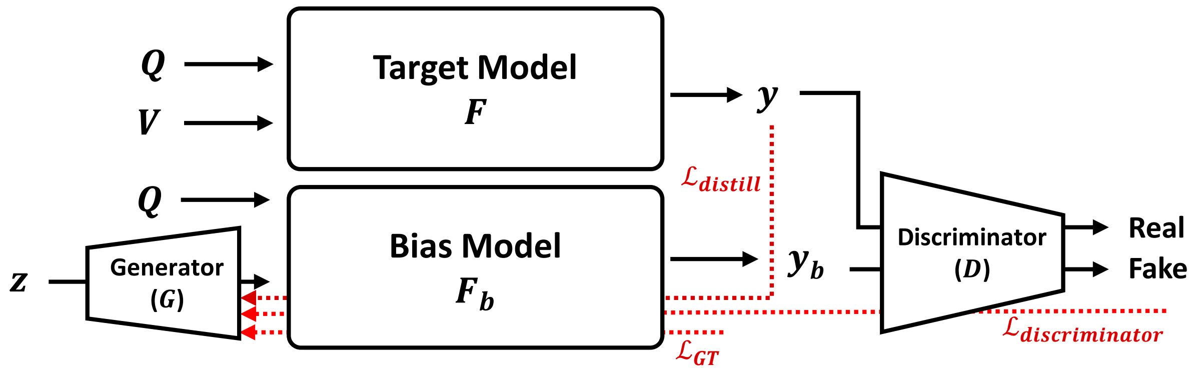
[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Generative Bias!

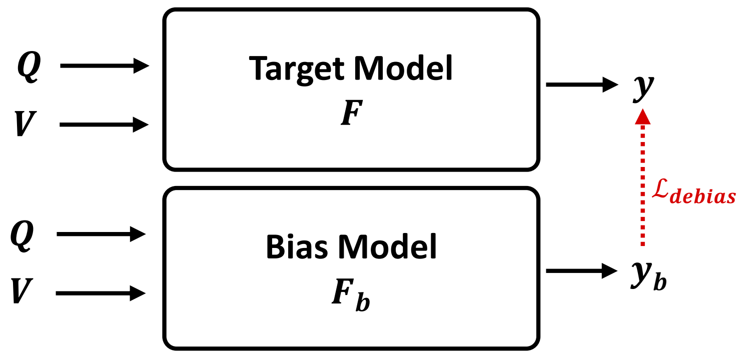


[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Generative Bias for Robust VQA

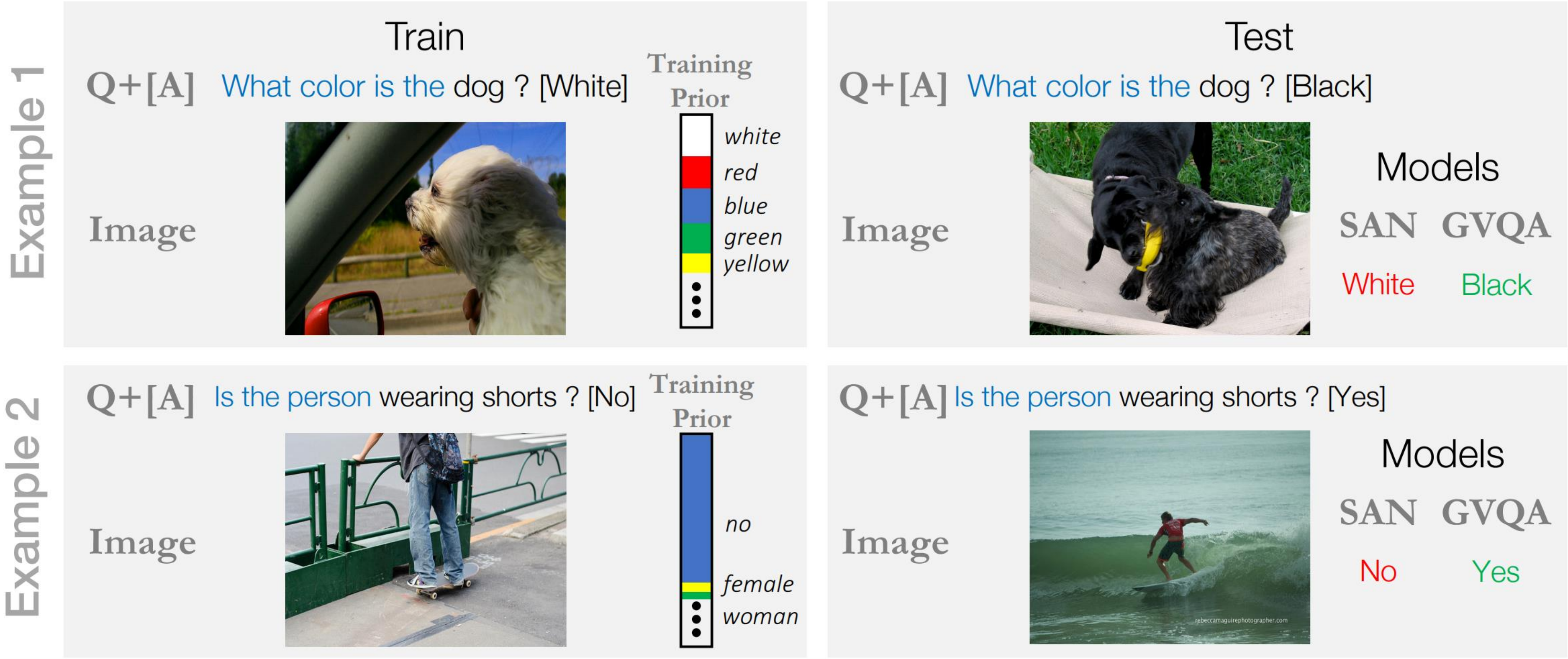


Full training of the bias model



Bias Issue

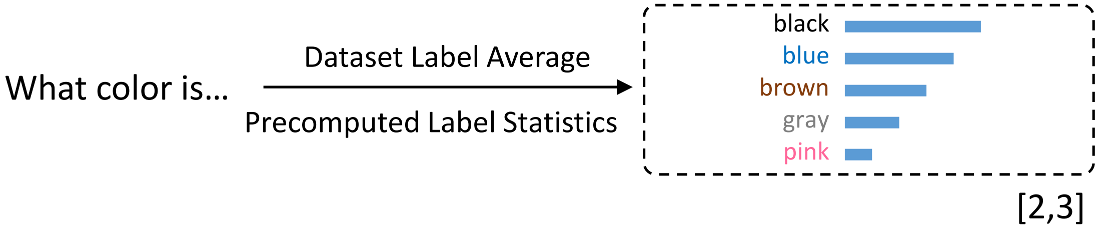
VQA models rely heavily on **language priors**!

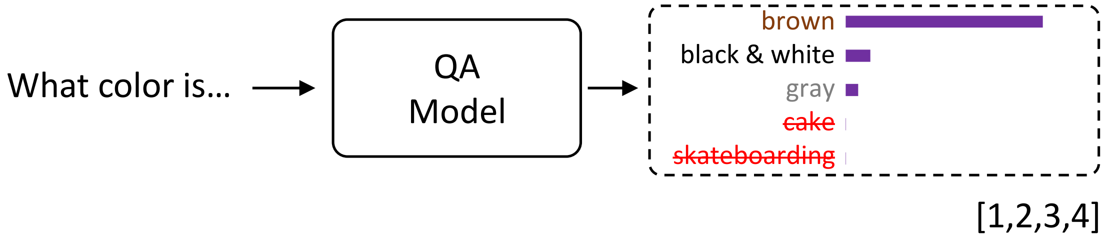


[1] Agrawal A., Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. CVPR 2018.

Bias?

Two commonly used statistics for debiasing in VQA

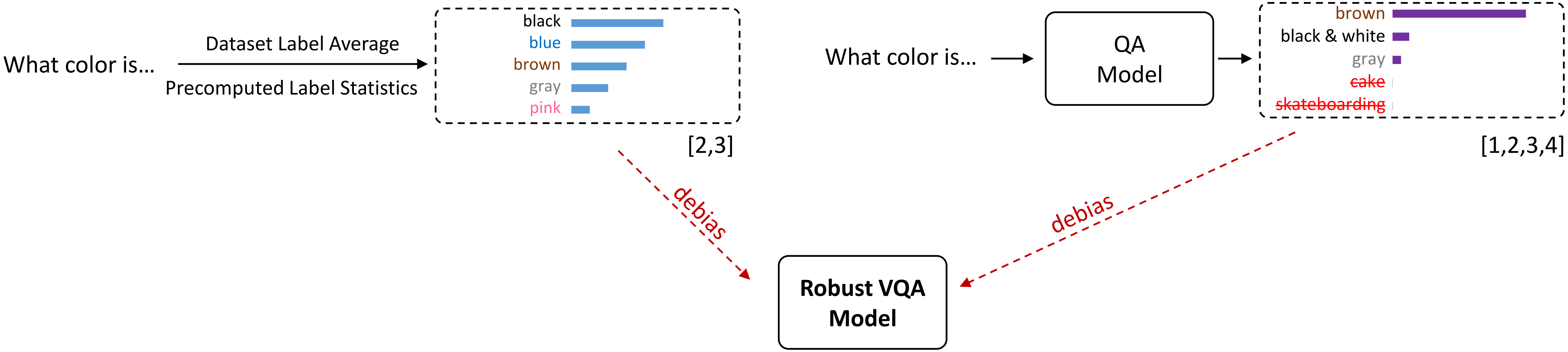




[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

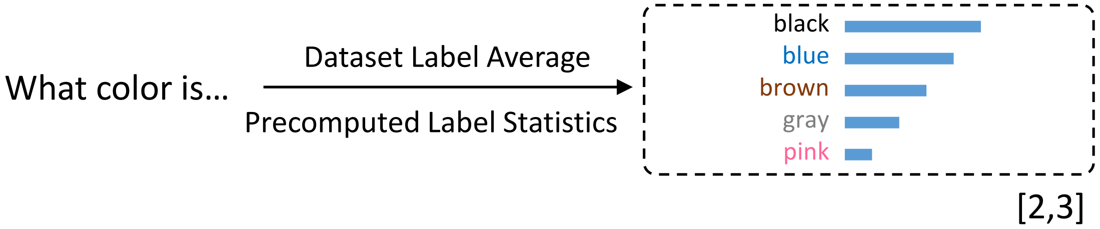
Bias?

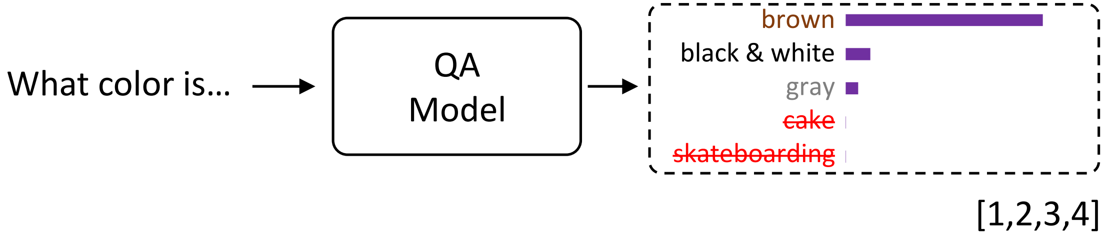
Two commonly used statistics for debiasing in VQA

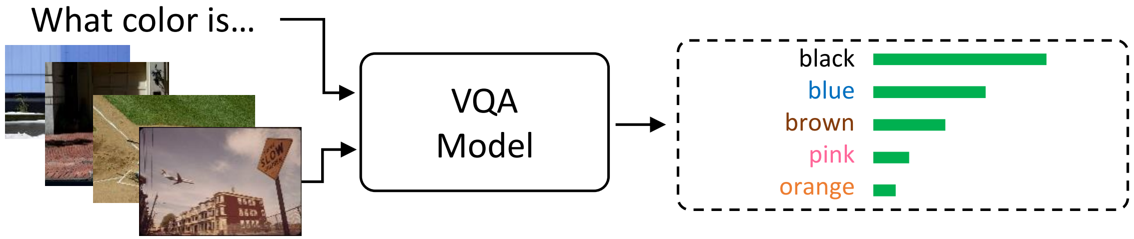


[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Bias?





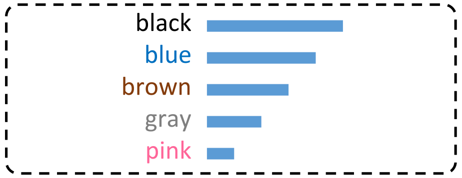


What the VQA model actually experiences

[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

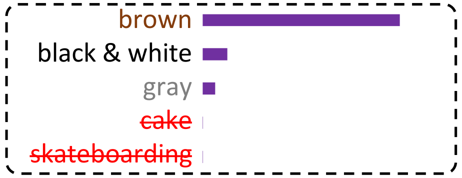
Bias?

What color is... Dataset Label Average / Precomputed Label Statistics



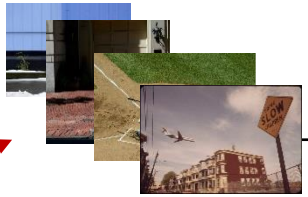
[2,3]

What color is... → QA Model →

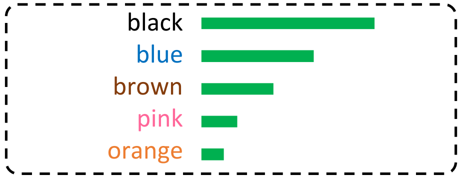


[1,2,3,4]

What color is...



→ VQA Model →



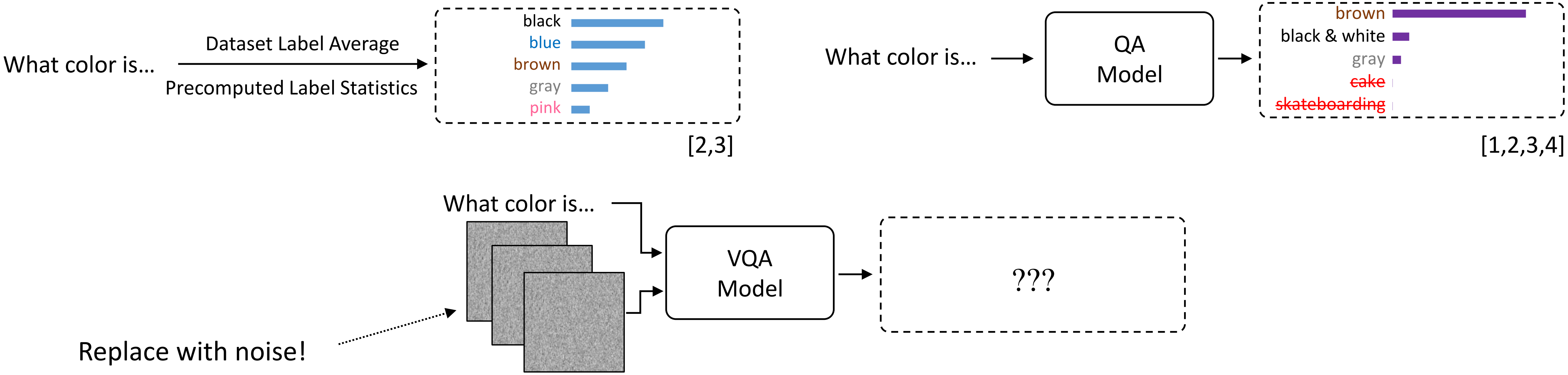
Limited by static from images

What the VQA model actually experiences

The better we can capture bias, the better we can debias

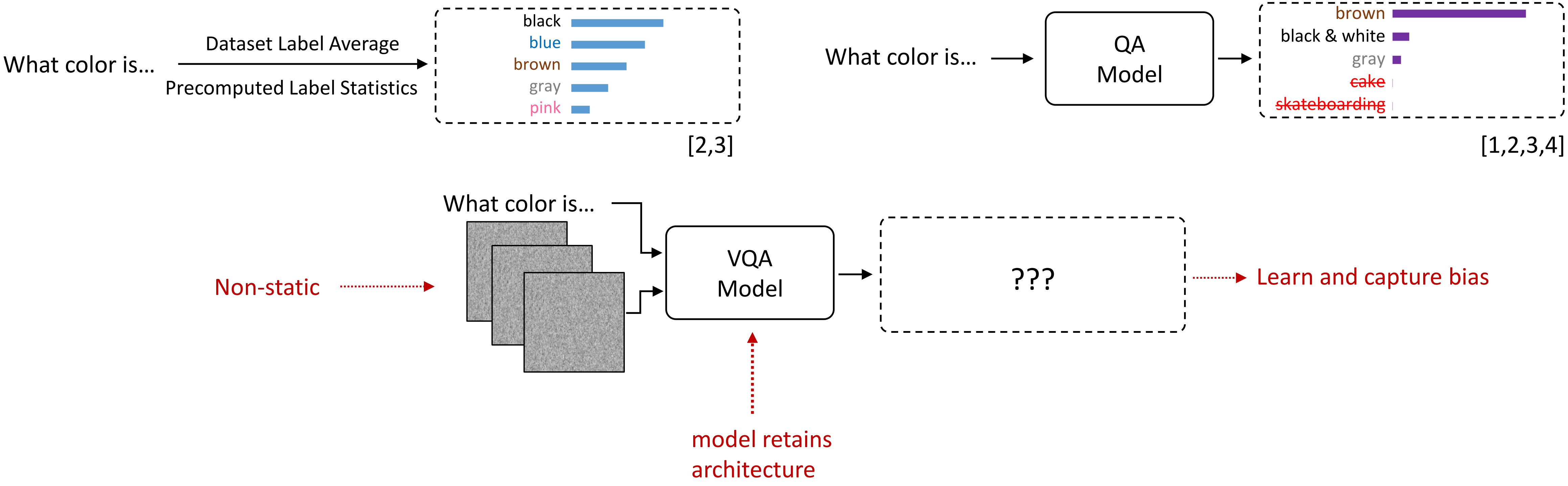
[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Generative Bias!



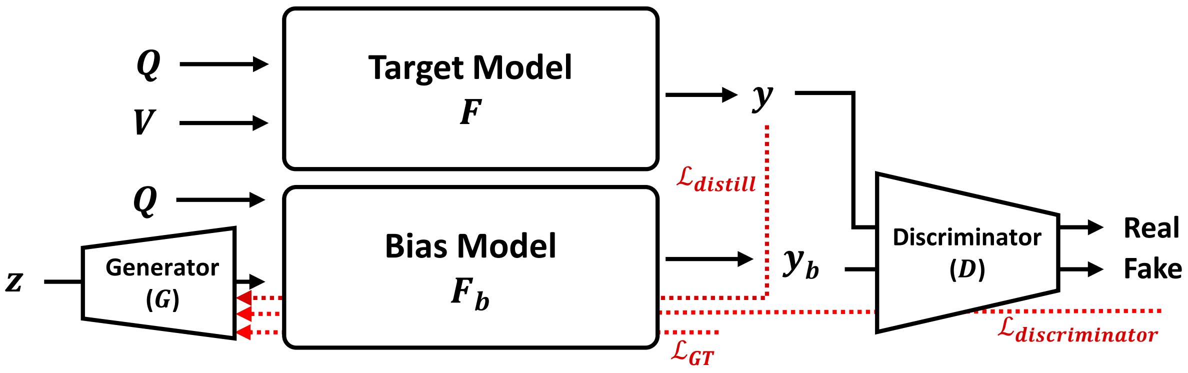
[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Generative Bias!

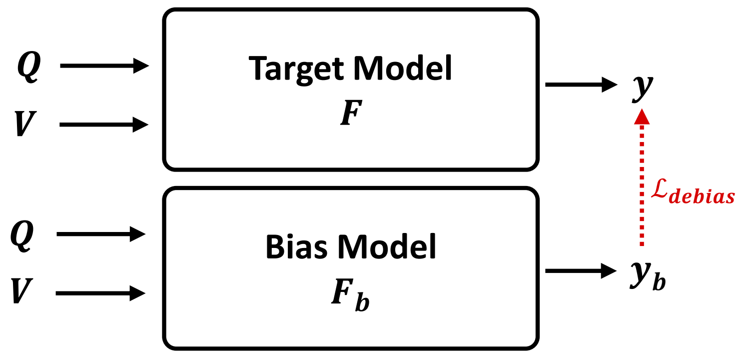


[1] Cadene R., RUBi: Reducing Unimodal Biases in Visual Question Answering. NeurIPS 2019.
[2] Clark C., Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. EMNLP 2019.
[3] Han X., Greedy Gradient Ensemble for Robust Visual Question Answering. ICCV 2021.
[4] Counterfactual VQA: A Cause-Effect Look at Language Bias. CVPR 2021.

Generative Bias for Robust VQA

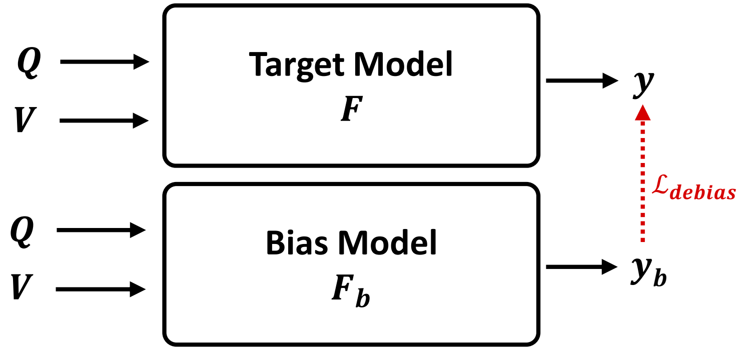


Full training of the bias model

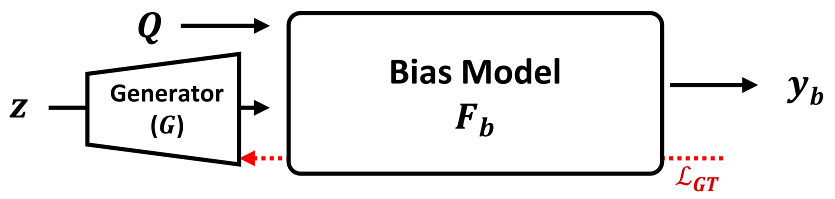


Ensemble Training

Bias Model captures ***bias*** and **Target Model** learns to ***debias*** from it

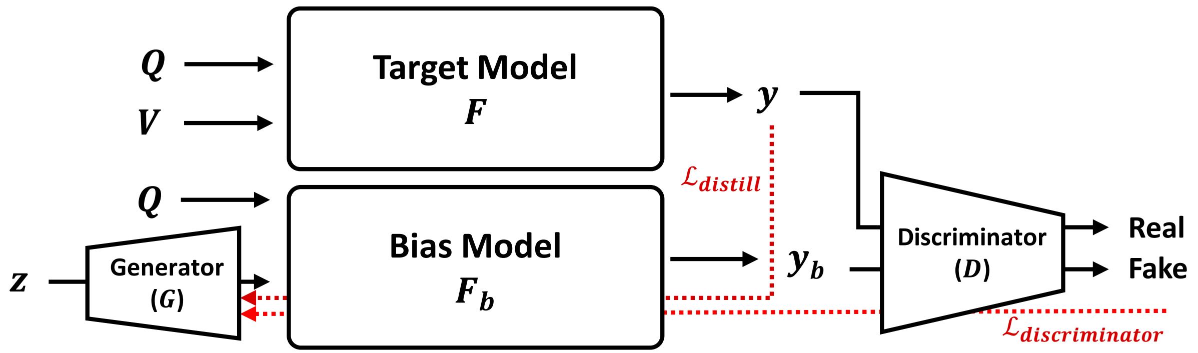


Bias Model Training



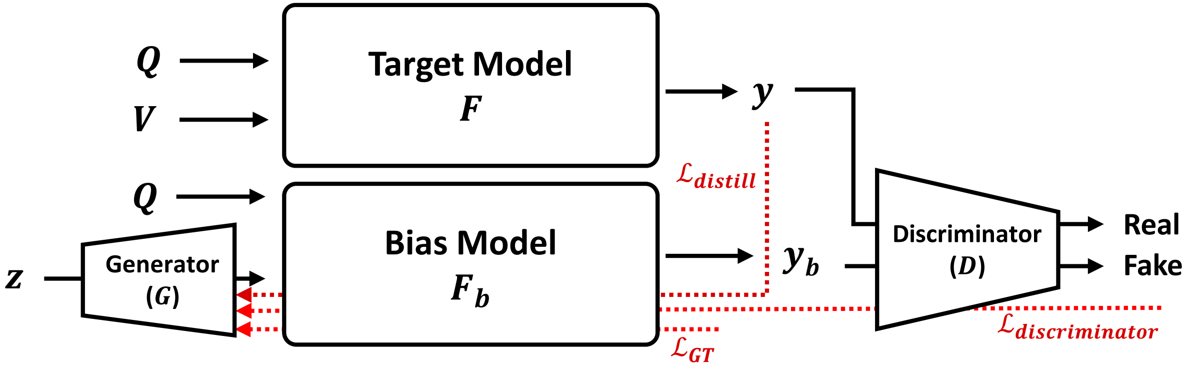
Learns **distribution Bias**

Bias Model Training



Learns the **Target Model's bias**

Bias Model Training



The bias model generates **stochastic bias representations**

Intuitively, Generator learns to "***hallucinates***" the "***visual input***"

Generative Bias for Robust VQA

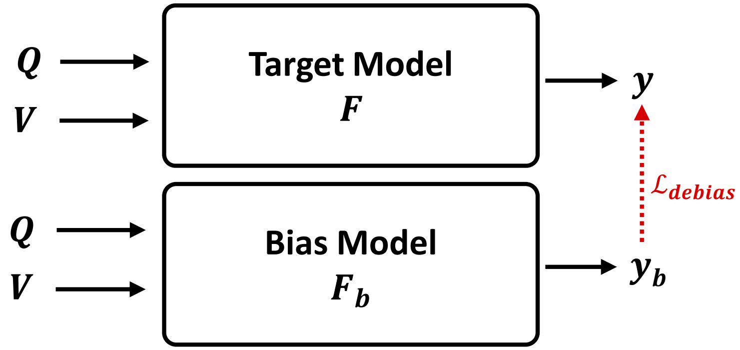
Target model debiasing

Bias Model’s output as negative gradient supervision

$$\mathcal{L}_{target}(F) = \mathcal{L}_{BCE}(\mathbf{y}, \mathbf{y}_{DL})$$

with,

$$\mathbf{y}^i_{DL} = \underline{\min\,(1}\ ,\ 2 \cdot \mathbf{y}^i_{gt} \cdot \sigma(-2 \cdot \mathbf{y}^i_{gt} \cdot \underline{\mathbf{y}^i_b}))$$



Using the raw unbounded output + clamping allows our loss to take into consideration the **intensity** of bias

Generative Bias for Robust VQA

Excellent performance

Method	Base	VQA-CP2 test				VQA-CP1 test			
		All	Yes/No	Num	Other	All	Yes/No	Num	Other
SAN [40]	-	24.96	38.35	11.14	21.74	32.50	36.86	12.47	36.22
GVQA [3]	-	31.30	57.99	13.68	22.14	39.23	64.72	11.87	24.86
S-MRL [7]	-	38.46	42.85	12.81	43.20	36.38	42.72	12.59	40.35
UpDn [4]	-	39.94	42.46	11.93	45.09	36.38	42.72	42.14	40.35
Methods based on modifying language modules									
DLR [22]	UpDn	48.87	70.99	18.72	45.57	–	–	–	–
VGQE [26]	UpDn	48.75	–	–	–	–	–	–	–
VGQE [26]	S-MRL	50.11	66.35	27.08	46.77	–	–	–	–
Methods based on strengthening visual attention									
HINT [32]	UpDn	46.73	67.27	10.61	45.88	–	–	–	–
SCR [38]	UpDn	49.45	72.36	10.93	48.02	–	–	–	–
Methods based on ensemble models									
AReg [31]	UpDn	41.17	65.49	15.48	35.48	43.43	74.16	12.44	25.32
RUBi [7]	UpDn	44.23	67.05	17.48	39.61	50.90	80.83	13.84	36.02
LMH [12]	UpDn	52.45	69.81	**44.46**	45.54	55.27	76.47	**26.66**	**45.68**
CF-VQA(SUM) [28]	UpDn	53.55	**91.15**	13.03	44.97	57.03	**89.02**	17.08	41.27
CF-VQA(SUM) [28]	S-MRL	55.05	90.61	21.50	45.61	**57.39**	**88.46**	14.80	43.61
CF-VQA(SUM) [28] + IntroD [29]	S-MRL	55.17	**90.79**	17.92	46.73	–	–	–	–
GGE [18]	UpDn	**57.32**	87.04	27.75	**49.59**	–	–	–	–
GenB (Ours)	UpDn	$\boxed{\textbf{59.15}}$	88.03	**40.05**	**49.25**	$\boxed{\textbf{62.74}}$	86.18	**43.85**	**47.03**

Method	GQA-OOD Test			
	All	Tail	Head	Avg
UpDn [4]	46.87	42.13	49.16	45.65
RUBi [7]	45.85	43.37	47.37	45.37
LMH [12]	43.96	40.73	45.93	43.33
CSS [9]	44.24	41.20	46.11	43.66
GenB (Ours)	$\boxed{\textbf{49.43}}$	**45.63**	**51.76**	**48.70**

Generative Bias for Robust VQA

Generative Bias works with other debiasing losses

Ensemble Debias Loss	Bias Model	VQA-CP2 test			
		All	Yes/No	Num	Other
–	UpDn	39.94	42.46	11.93	45.09
GGE [18]	UpDn	47.40	64.45	13.96	47.64
Our Loss	UpDn	52.47	88.20	30.09	40.38
RUBi [7]	GenB	30.77	72.78	12.15	13.87
LMH [12]	GenB	53.99	75.89	**44.62**	45.08
GGE [18]	GenB	49.51	70.63	14.08	48.16
Ours Loss	GenB	**59.15**	**88.03**	40.05	**49.25**

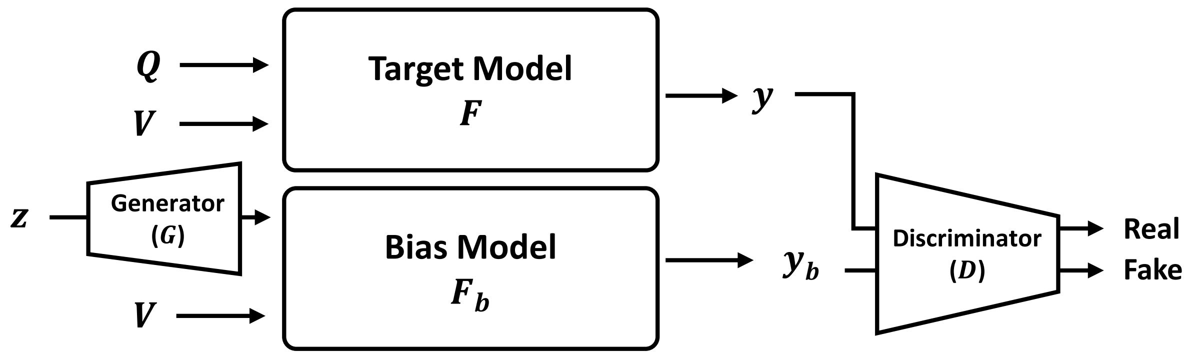
Architecture Agnostic

Architecture	VQA-CP2 test				Δ Gap
	All	Yes/No	Num	Other	
UpDn [4]	39.94	42.46	11.93	45.09	**+19.21**
UpDn [4] + GenB	**59.15**	**88.03**	**40.05**	**49.25**	
BAN† [25]	37.35	41.96	12.08	41.71	**+20.02**
BAN† [25] + GenB	**57.37**	**89.11**	**29.52**	**48.37**	
SAN† [40]	38.65	40.59	12.98	44.67	**+18.07**
SAN† [40] + GenB	**56.72**	**88.84**	**19.04**	**50.22**	
LXMERT [35]	46.23	42.84	18.91	55.51	**+24.93**
LXMERT [35] + GenB (**Ours Best**)	$\boxed{\textbf{71.16}}$	**92.24**	**64.71**	**61.89**	
Reported LXMERT Performance					
LXMERT [35] + MUTANT [14]	69.52	93.15	67.17	57.78	
LXMERT [35] + D-VQA [37]	69.75	80.43	58.57	67.23	
LXMERT [35] + SAR [33]	62.12	85.14	41.63	55.68	

State-of-the-art!

Generative Question Bias?

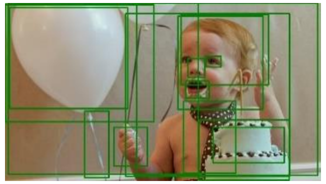
GenB Visual



Bias Model	VQA-CP2 test			
	All	Yes/No	Num	Other
UpDn	39.94	42.46	11.93	45.09
UpDn	52.47	88.20	30.09	40.38
Visual-Answer	41.03	42.69	12.66	47.93
Question-Answer	56.68	**89.30**	20.78	**49.43**
GenB Visual	49.54	72.05	12.58	47.89
GenB Question (Ours)	**59.15**	88.03	**40.05**	49.25

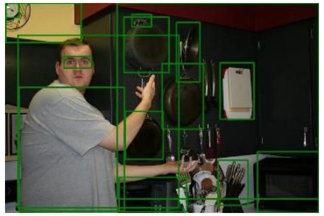
What does the model actually see?

Ground Truth



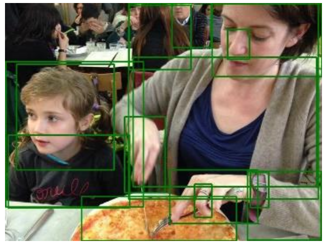
Q: What color is the balloon?

GT: white: 1.0



Q: What color is the man's shirt?

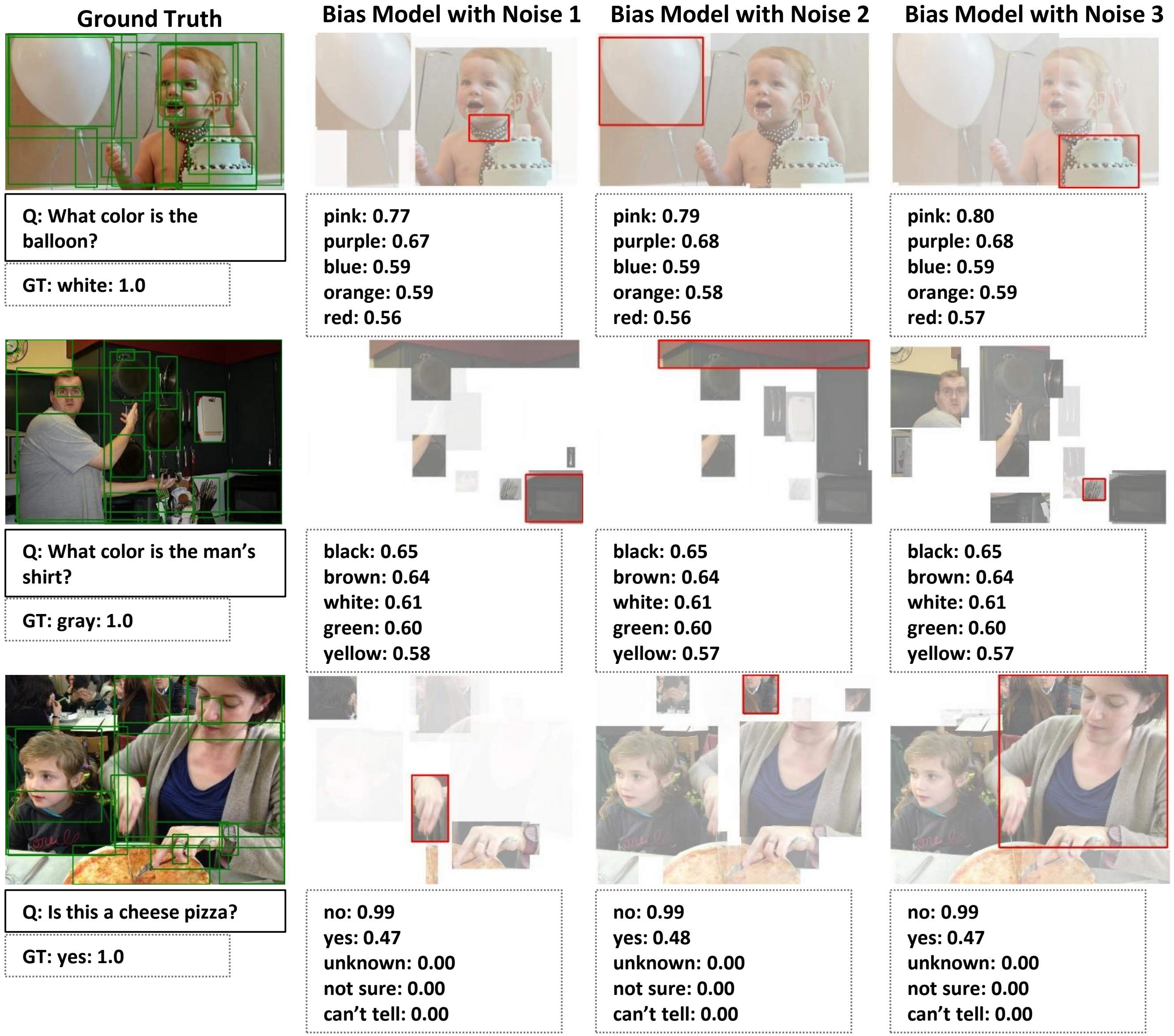
GT: gray: 1.0



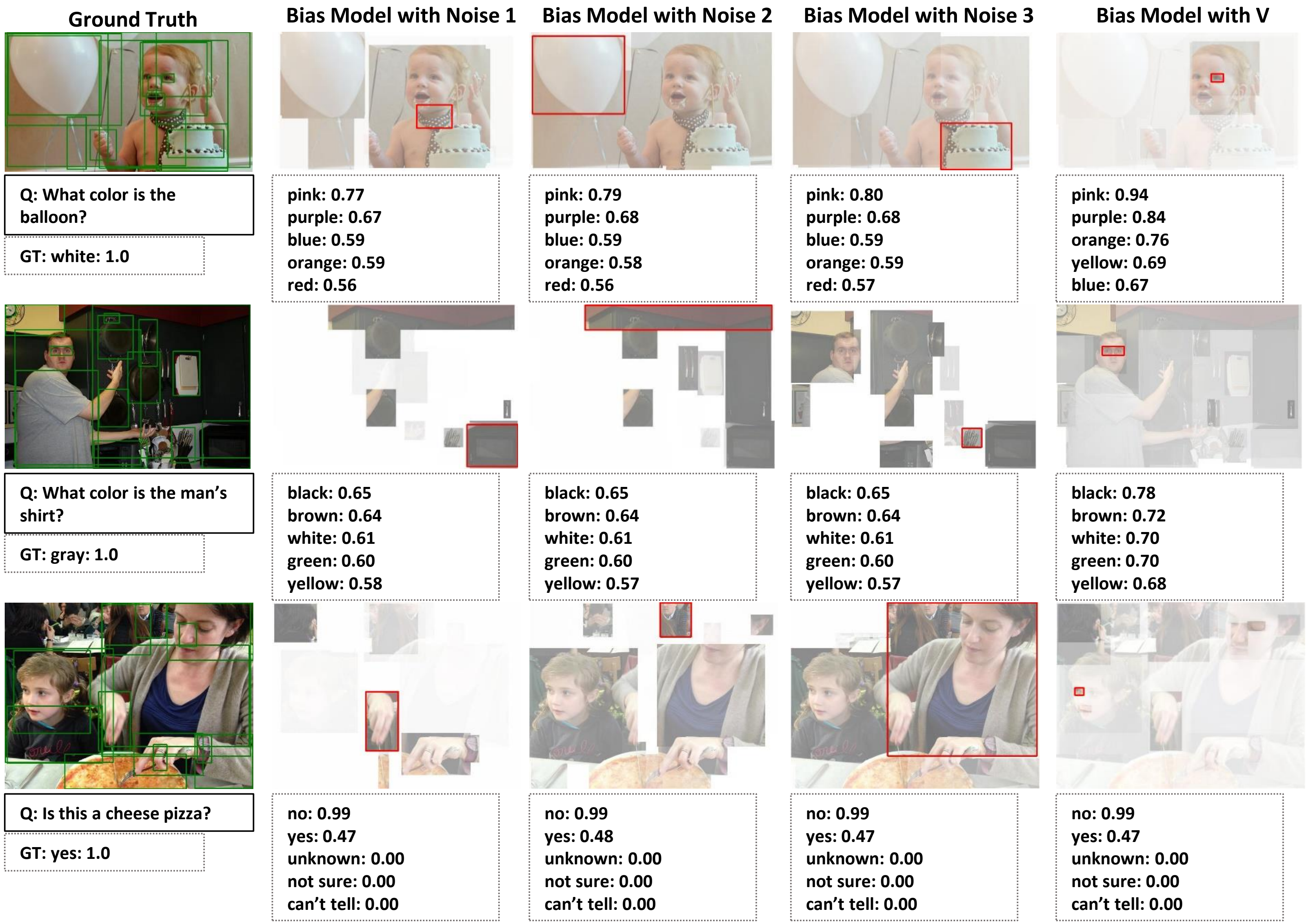
Q: Is this a cheese pizza?

GT: yes: 1.0

What does the model actually see?



What does the model actually see?



What does the model actually see?

Thank You!

Github: https://github.com/chojw/genb