Debiasing Multimodal Models via Causal Information Minimization

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

Biases in VL tasks

What is the color of the toaster?

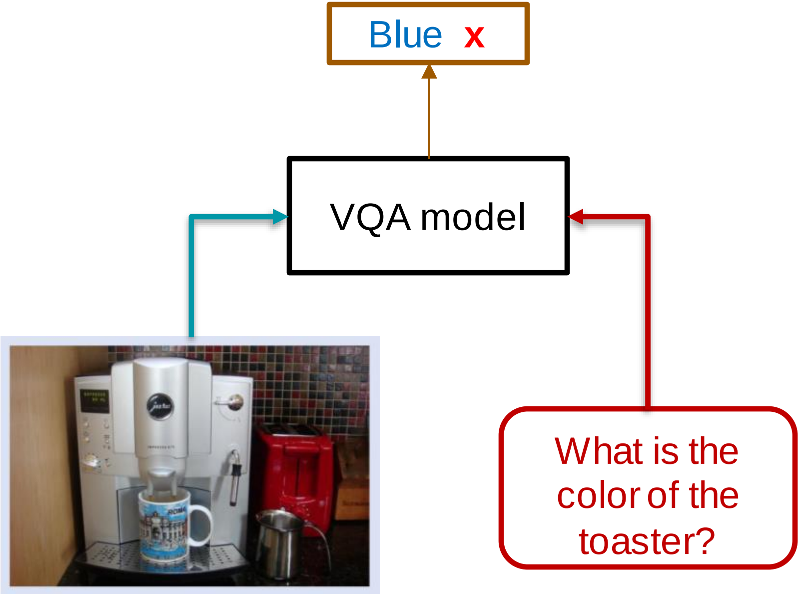
Biases in VL tasks

VQA model

What is the color of the toaster?

Biases in VL tasks

Vision bias



Biases in VL tasks

Vision bias

Blue x

VQA model

What is the color of the toaster?

What is the color of the banana?

https://cdancette.fr/2020/11/21/overview-bias-reductions-vqa/

Biases in VL tasks

Vision bias

Blue x

VQA model

What is the color of the toaster?

VQA model

What is the color of the banana?

https://cdancette.fr/2020/11/21/overview-bias-reductions-vqa/

Biases in VL tasks

Vision bias

Blue x

VQA model

What is the color of the toaster?

Language bias

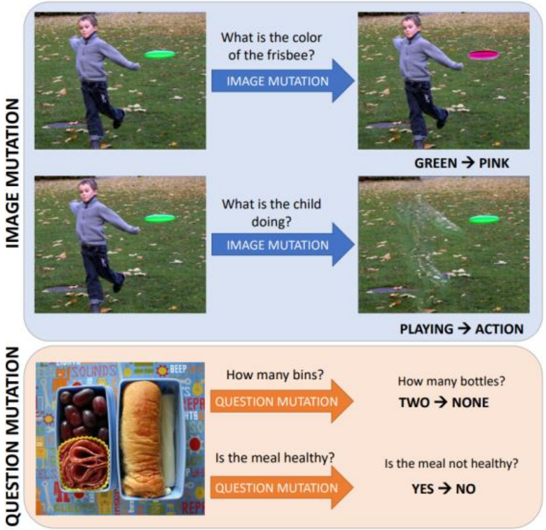
Yellow x

VQA model

What is the color of the banana?

https://cdancette.fr/2020/11/21/overview-bias-reductions-vqa/

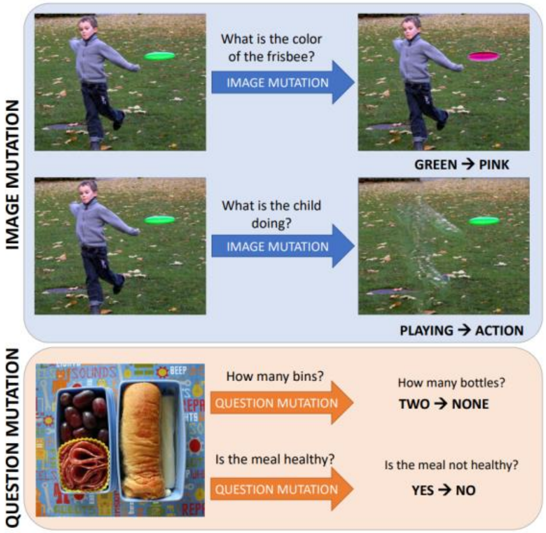
Previous works



Data augmentation

Gokhale, Tejas, et al. "MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering." *EMNLP*. 2020.

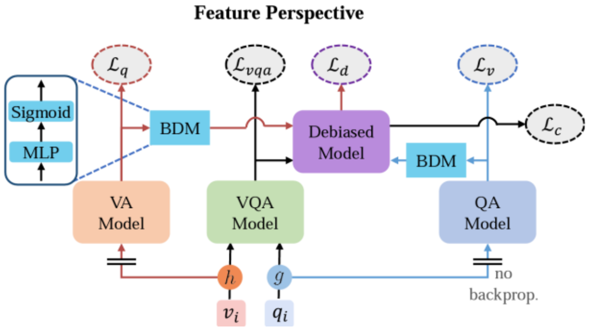
Previous works



Data augmentation

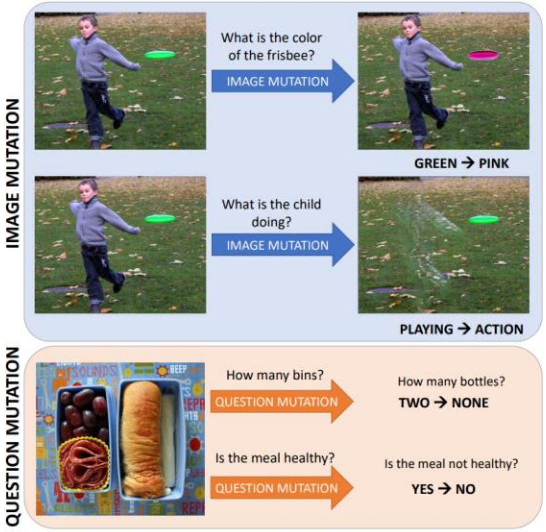
Gokhale, Tejas, et al. "MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering." *EMNLP*. 2020.

Inductive bias in model architecture



Wen, Zhiquan, et al. "Debiased visual question answering from feature and sample perspectives." Advances in Neural Information Processing Systems 34 (2021): 3784-3796.

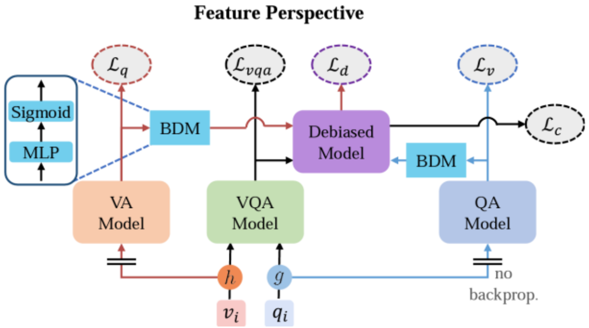
Previous works



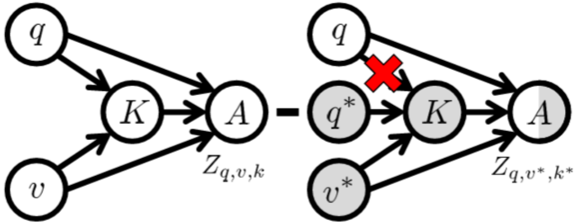
Data augmentation

Gokhale, Tejas, et al. "MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering." *EMNLP*. 2020.

Inductive bias in model architecture



Wen, Zhiquan, et al. "Debiased visual question answering from feature and sample perspectives." Advances in Neural Information Processing Systems 34 (2021): 3784-3796.



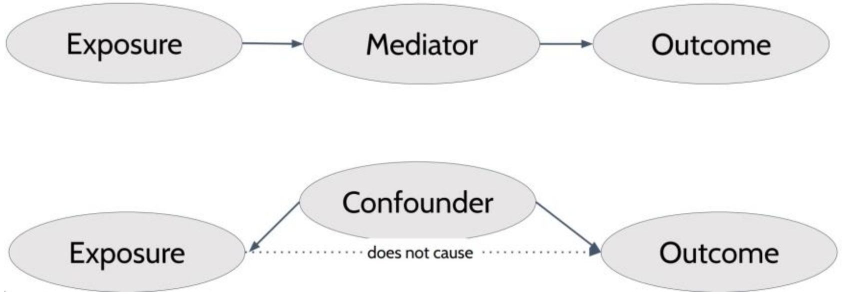
Causal debiasing

Niu, Yulei, et al. "Counterfactual vqa: A cause-effect look at language bias." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Causality background: Confounders

Confounders:

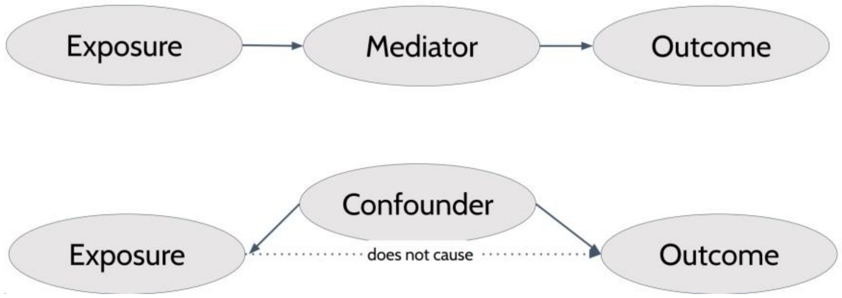
- create non-causal dependencies between inputs and output



Causality background: Confounders

Confounders:

- create non-causal dependencies between inputs and output



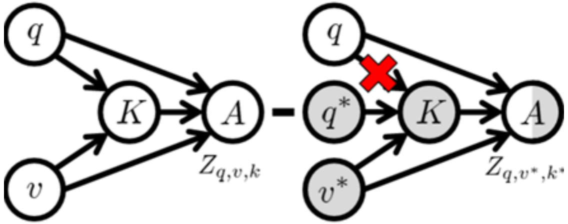
Biases in VQA:

- spurious correlations in the dataset

This work:

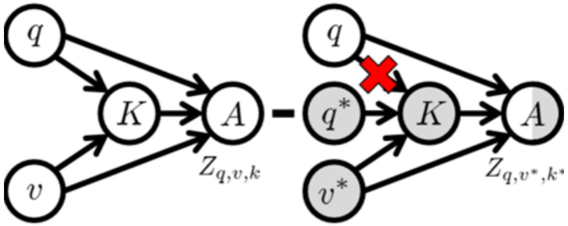
- Model biases as confounders

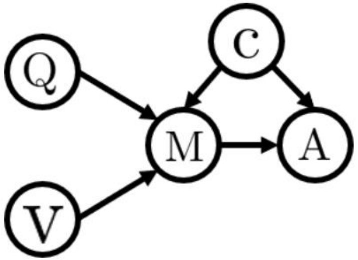
Causality background: Confounders



- Only model language bias through q
- Ignores vision biases
- Ignores multimodal biases too!

Causality background: Confounders



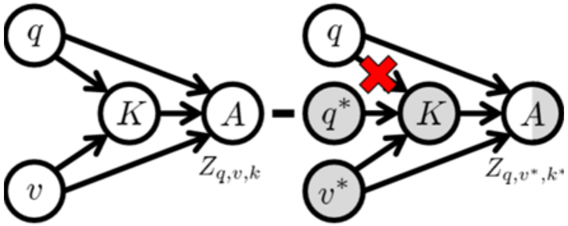


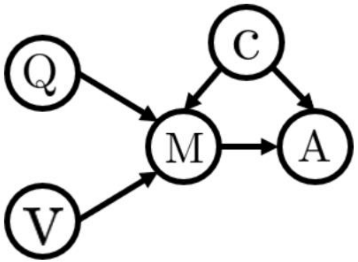
- Only model language bias through q
- Ignores vision biases
- Ignores multimodal biases too!

This work:

- Model multimodal biases

Causality background: Confounders





- Only model language bias through q
- Ignores vision biases
- Ignores multimodal biases too!

This work:

- Model multimodal biases

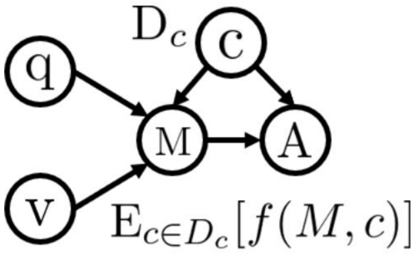
Causal inference

- Isolate the causal effect of M on A
- Free from the confounders c

Causal debiasing theories

Let's assume that the confounders C are known!

Average Treatment Effect



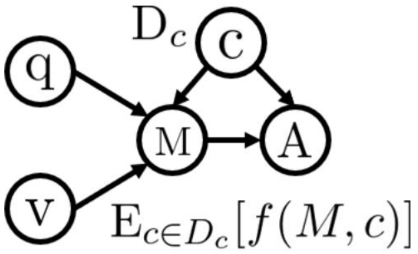
$$P(A|do(M)) = E_{c \sim C}[P(A|M, c)]$$

By taking the expected value over confounders, it eliminates the direct effect of C on M

Causal debiasing theories

Let's assume that the confounders C are known!

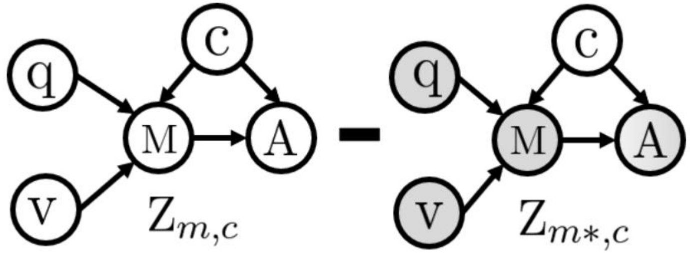
Average Treatment Effect



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

By taking the expected value over confounders, it eliminates the direct effect of C on M

Total Effect



$$TE = A_{m,C_m} - A_{m*,C_m}$$

By retaining the confounder in both sides of the difference, it eliminates the direct effect of Cm on M

How to model confounders?

Spurious correlations

- *simplest predictive features*
- explain biased datasets (Geirhos et al., 2020)

How to model confounders?

Spurious correlations

- *simplest predictive features*
- explain biased datasets (Geirhos et al., 2020)

- Deep models preferentially encode *dataset shortcuts* under limited representation capacity (Yang et al., 2022)

How to model confounders?

Spurious correlations

- *simplest predictive features*
- explain biased datasets (Geirhos et al., 2020)

- Deep models preferentially encode *dataset shortcuts* under limited representation capacity (Yang et al., 2022)

Neural nets tend to strike a balance between

- maximizing compression of learned representations
- fitting the labels (Shwartz-Ziv and Tishby, 2022)

How to model confounders?

Spurious correlations

- *simplest predictive features*
- explain biased datasets (Geirhos et al., 2020)

- Deep models preferentially encode *dataset shortcuts* under limited representation capacity (Yang et al., 2022)

Neural nets tend to strike a balance between

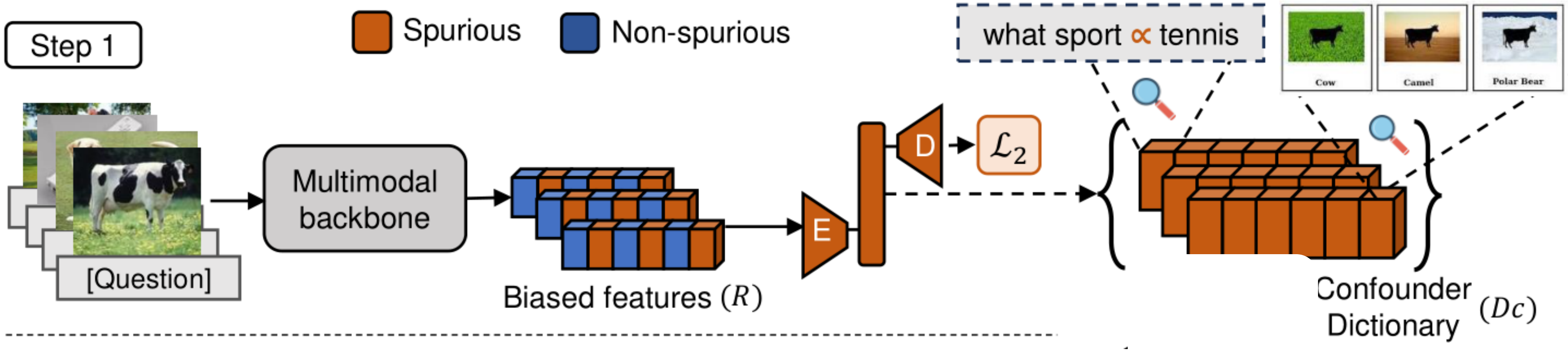
- maximizing compression of learned representations
- fitting the labels (Shwartz-Ziv and Tishby, 2022)

This work:

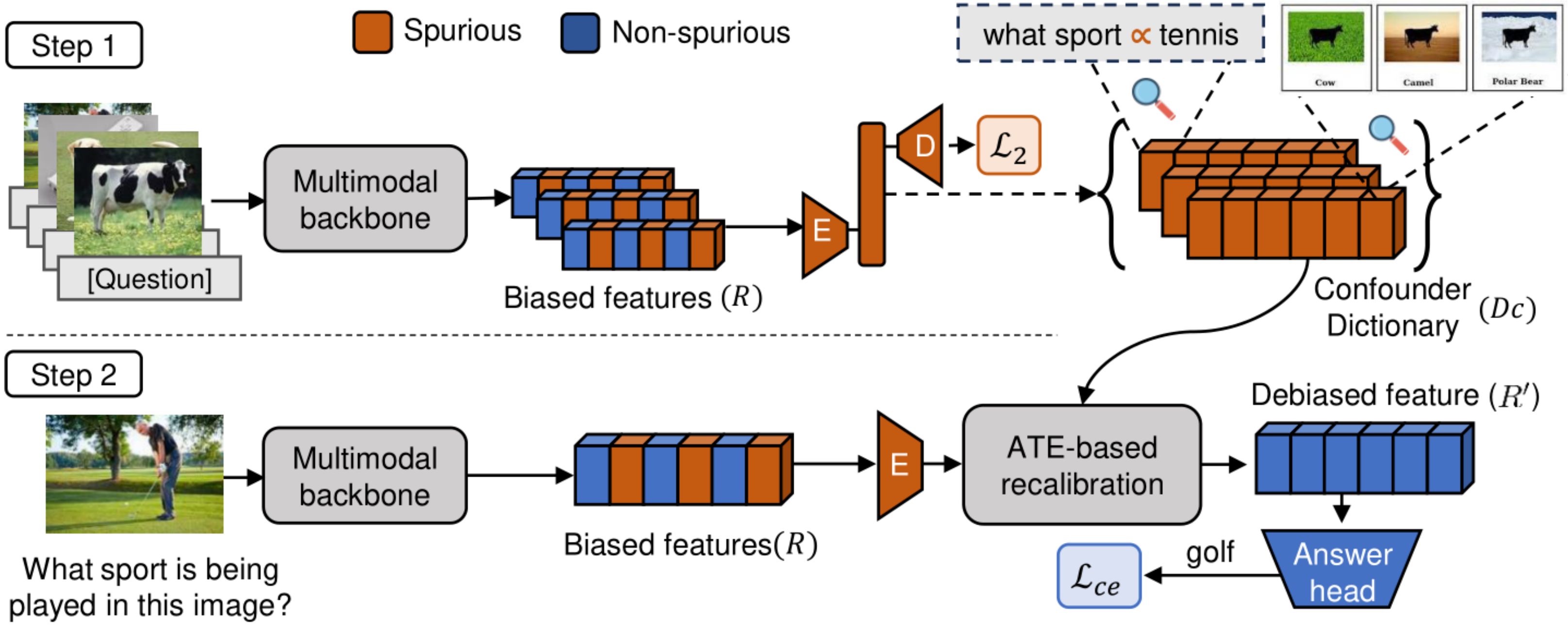
Modeling confounder

- Minimizing information in representations
- maximizing the task accuracy

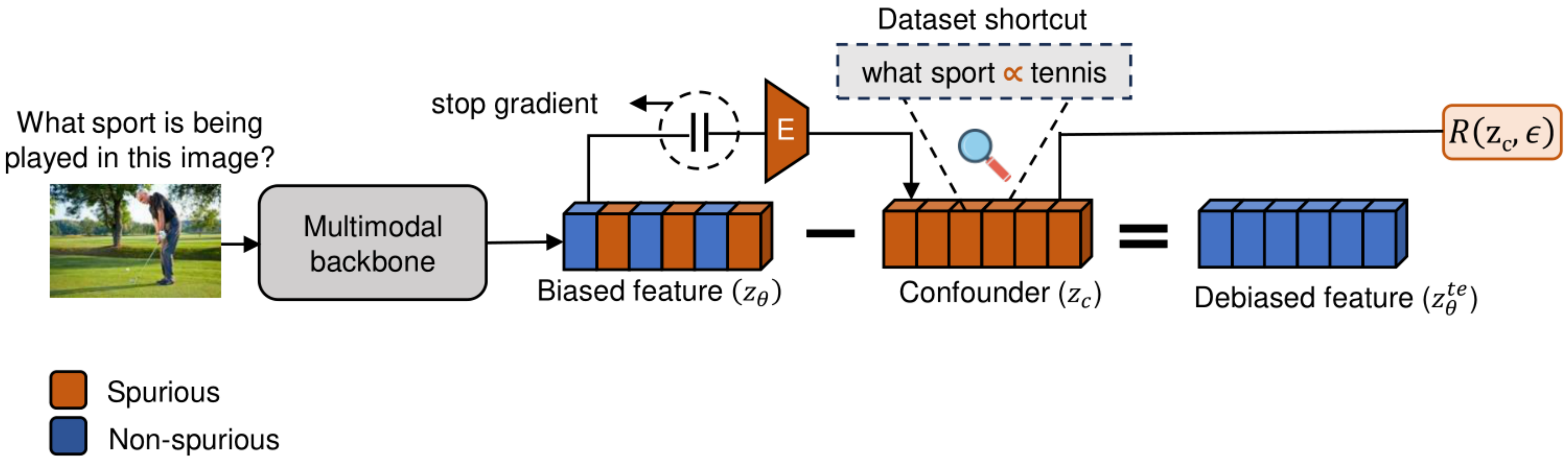
Average Treatment Effect-Debiasing (ATE-D)



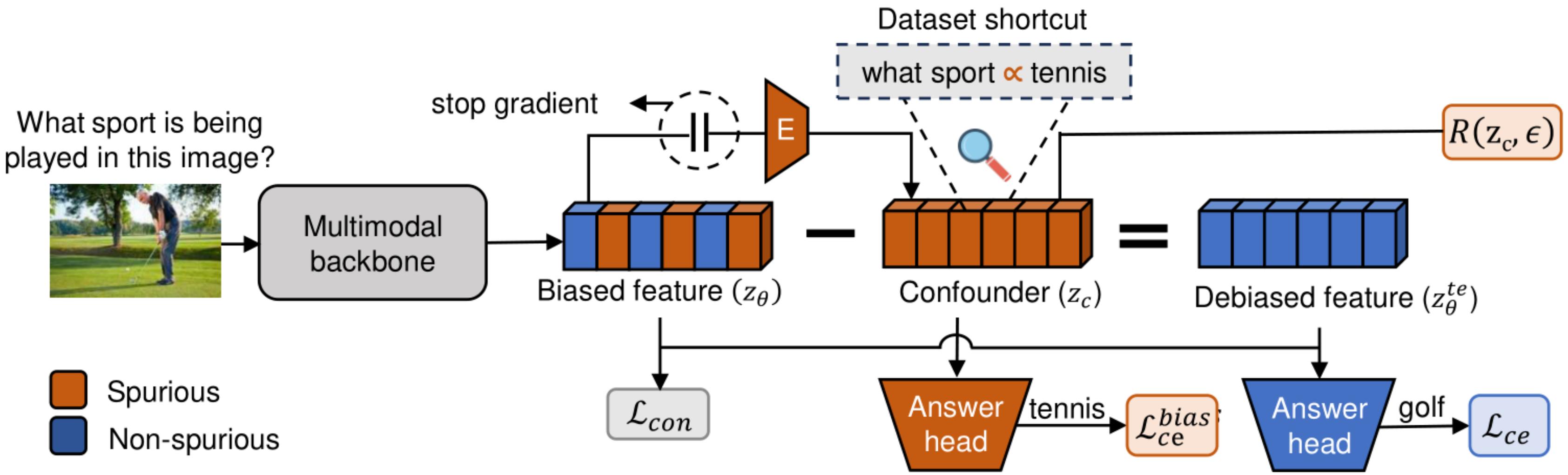
Average Treatment Effect-Debiasing (ATE-D)



Total Effect- Debiasing (TE-D)



Total Effect- Debiasing (TE-D)



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

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

TE-D improves the accuracy of Yes/No category by 4.2% which has higher bias presence

Does causal debiasing improve robustness to spurious features?

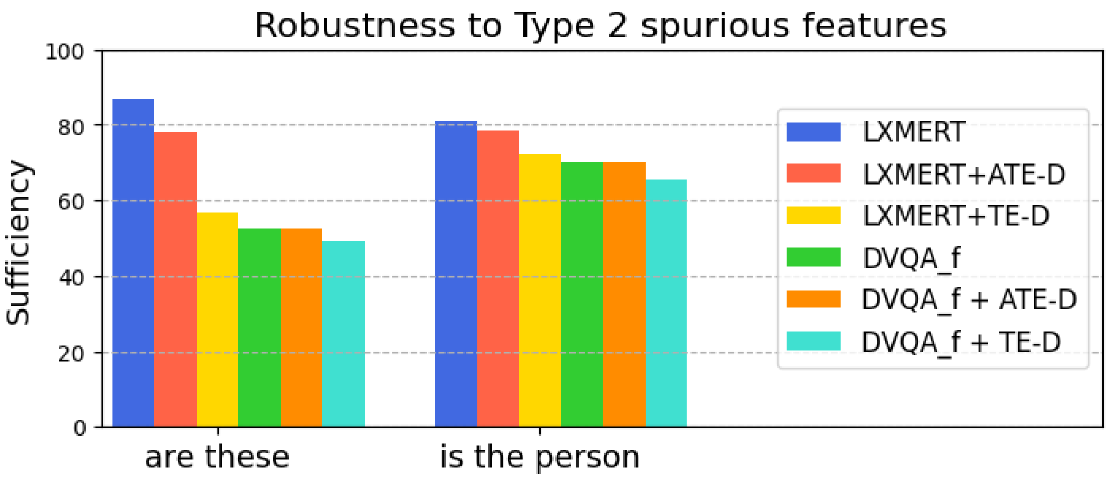
	VQA-CP				IVQA-CP				Additional
	Overall	Yes/No	Num	other	Overall	Yes/No	Num	other	#MFLOPS
LXMERT (Tan and Bansal, 2019)	41.2	44.1	13.9	47.2	35.0	43.3	12.7	36.8	-
+ IRM (Peyrard et al., 2022)	42.7	44.1	15.2	49.5	36.5	43.2	12.8	39.3	-
+ ATE-D (ours)	42.2	43.6	14.6	49.0	35.8	42.9	13.2	38.2	**0.7**
+ TE-D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

TE-D improves the accuracy of Yes/No category by 3.2% which has higher bias presence

Does causal debiasing improve robustness to spurious features?

- We propose sufficiency score (λ) as the percentage of the model's certainty attributed to the spurious input component in prediction.

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



Is cross-modal debiasing more effective than unimodal debiasing?

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

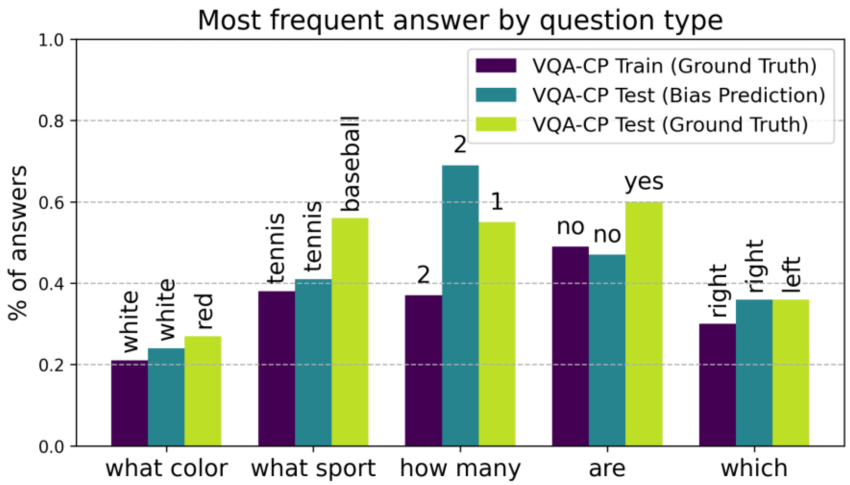
When D-VQAf is treated as the biased model in TE-D, additional improvements of 0.7 are achieved

What kind of biases are captured by confounder representations?

TE-D

- Answers from 0.34% of the vocabulary address 67% of training questions

- Most frequent answers obtained from biased representations align with those in train set, indicating effective representation of dataset biases



Conclusion

- *ATE-D and TE-D model and mitigate biases by imposing causally-driven information loss on biased features*
- *These methods effectively eliminate biases arising from both unimodal and multimodal interactions*
- *Data augmentation based approaches, although cumbersome, are more effective than feature-based debiasing*