# DEBIASING MULTIMODAL MODELS VIA CAUSAL INFORMATION MINIMIZATION

Vaidehi Patil, Adyasha Maharana and Mohit Bansal

UNC NLP

Code & Paper
vaidehi@cs.unc.edu

EMNLP 2023

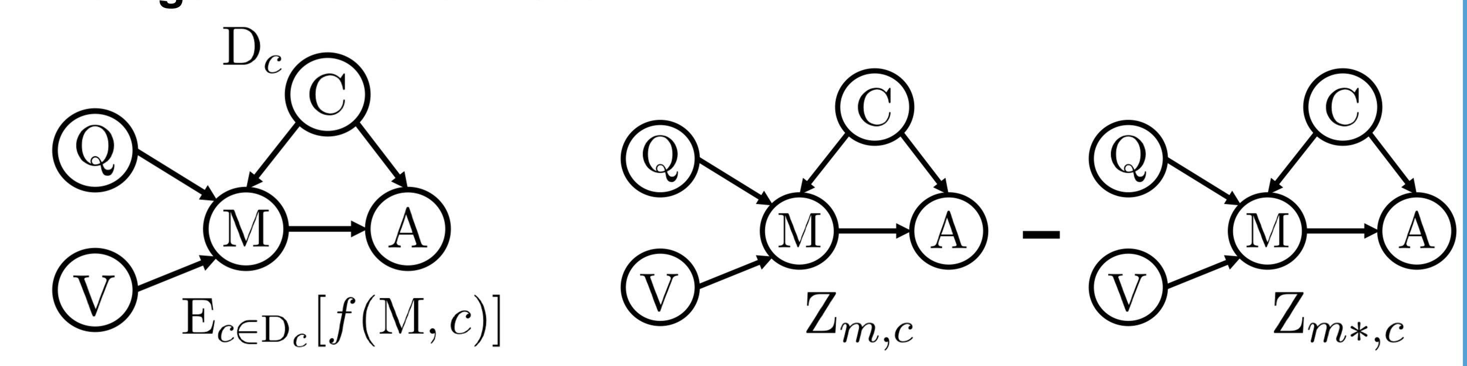
## Causal Mechanisms

**Our Contributions:**

We model the biases in multimodal datasets as confounders in causal graph.

We learn confounders by:
- minimizing information in biased representations &
- maximizing the task acc.

We propose two debiasing methods using these confounder to debias multimodal models.



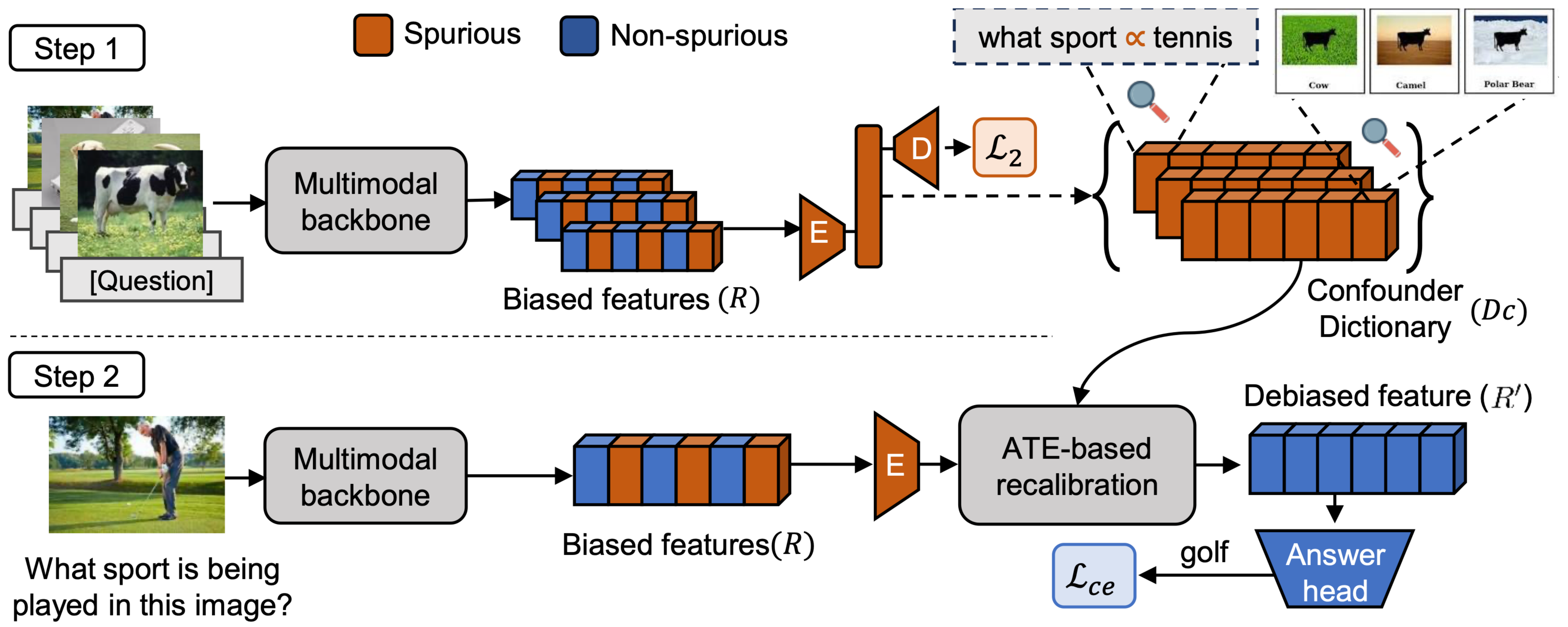

$P(A|do(M)) = E_{c \sim C}[P(A|M,c)]$

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

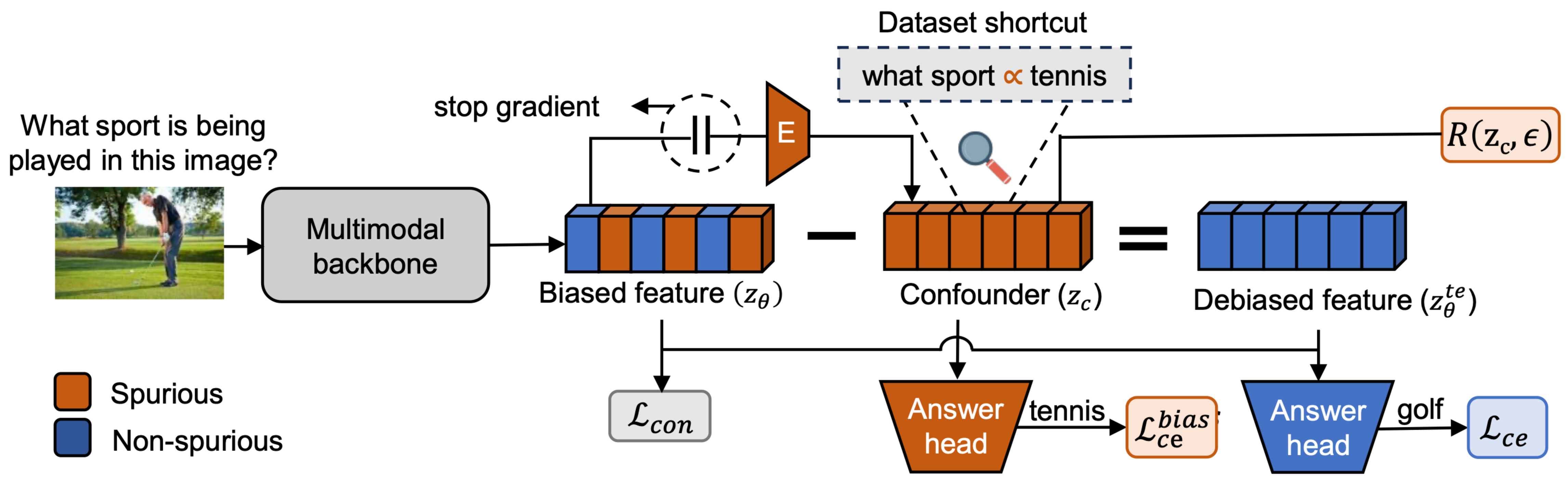
Average Treatment Effect computes the expected value over the distribution of confounders to eliminates the direct effect of C on M.

Total Effect eliminates the direct effect of C on M and A by taking the difference between biased A with and without treatment from M.

## Average Treatment Effect-Debiasing (ATE-D)




## Total Effect-Debiasing (TE-D)




## OOD Generalization

Our method based on causal info. minimization
- improves OOD acc. without hurting ID acc.
- removes biases arising from both unimodal and multimodal interaction

Data augmentation approaches are cumbersome but more effective than feature based debiasing.

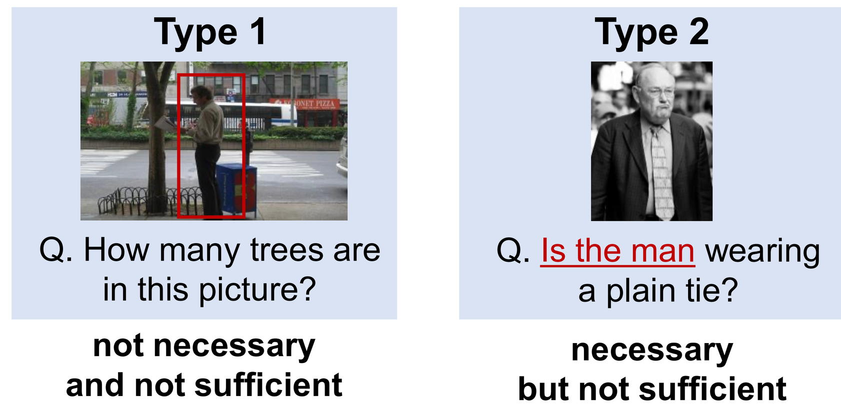
| | VQA-CP | | | | IVQA-CP | | | | Additional |
|---|---|---|---|---|---|---|---|---|---|
| | Overall | Yes/No | Num | other | Overall | Yes/No | Num | other | #MFLOPS |
| LXMERT Tan and Bansal (2019) | 41.2 | 44.1 | 13.9 | 47.2 | 35.0 | 43.3 | 12.7 | 36.8 | - |
| + IRM Peyrard et al. (2022) | 42.7 | 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 | <u>48.3</u> | 14.4 | 48.8 | 36.7 | <u>46.5</u> | 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** | <u>49.7</u> | **41.3** | **50.7** | **16.7** | **39.4** | 50.2 |
| D-VQA$_f$ Wen et al. (2021) | <u>43.9</u> | 47.5 | <u>15.7</u> | **49.8** | <u>37.3</u> | 45.8 | <u>13.9</u> | <u>39.2</u> | 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 |

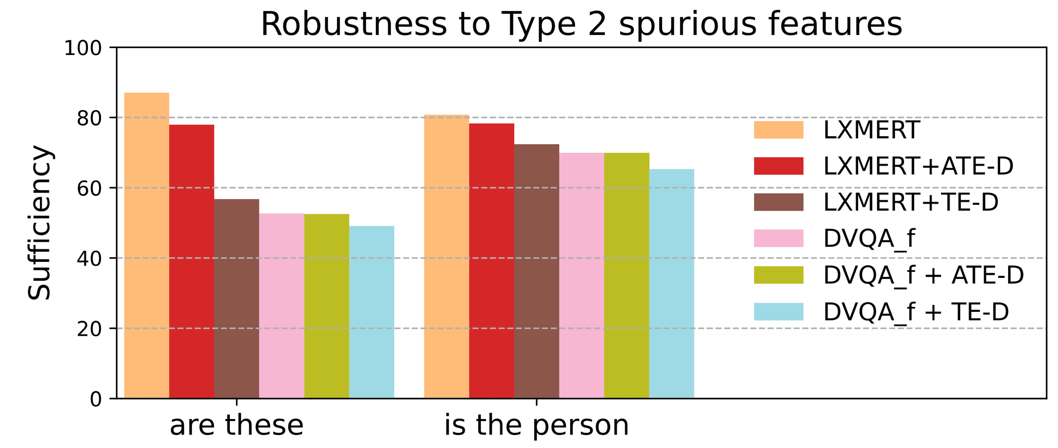
Results from evaluation of our methods and other debiasing methods on VQA-CP and IVQA-CP datasets.

## Robustness to Spurious Features

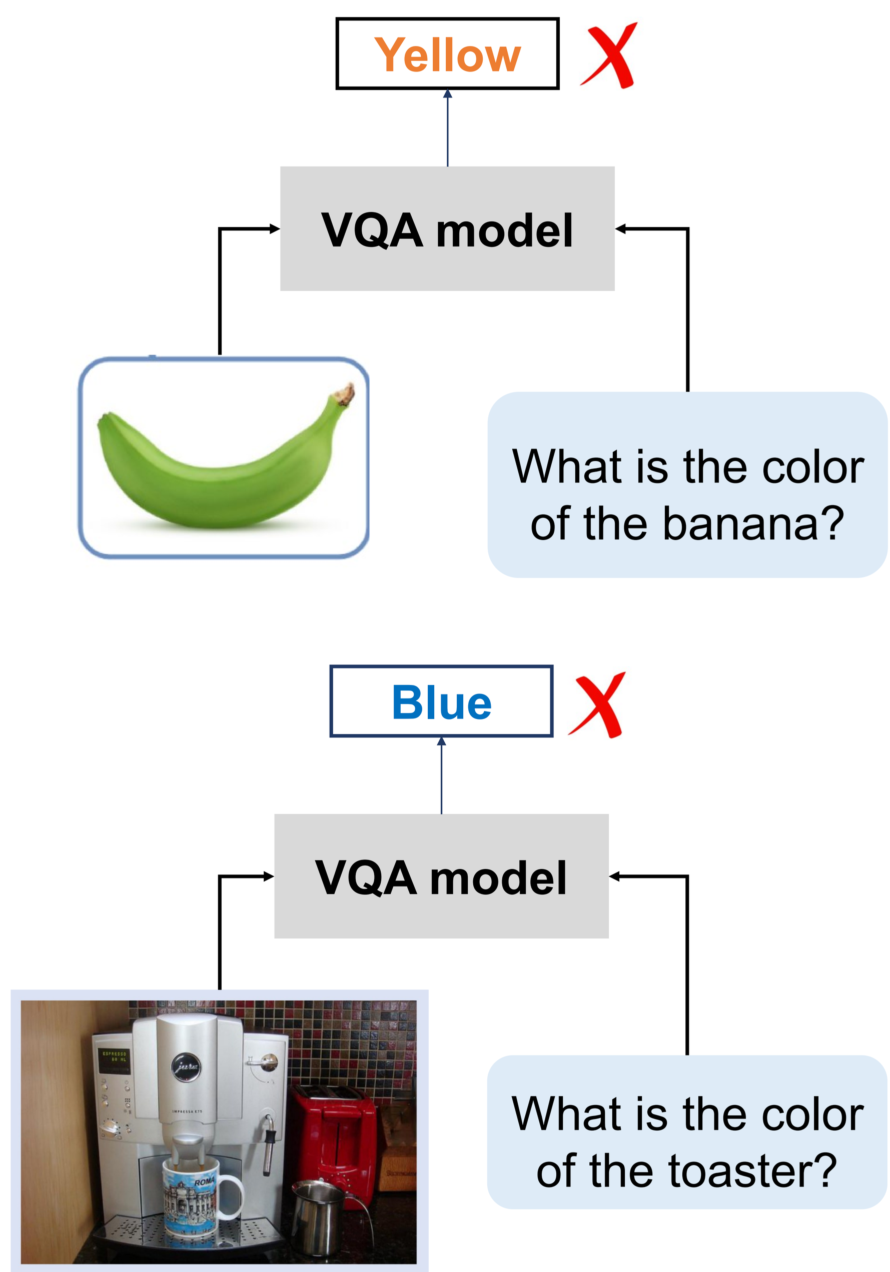
We propose sufficiency score (λ) as the percentage of the model's certainty attributed to the spurious input component in prediction.

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



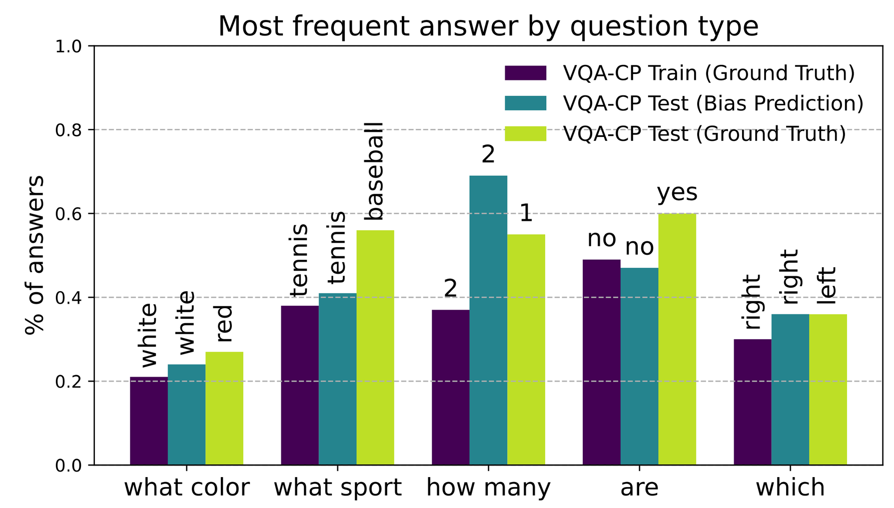




## VL Biases




## Confounder Analysis

**Total Effect-Debiasing (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




**Average Treatment Effect-Debiasing (ATE-D)**
- Boosting biased features hurts OOD accuracy
- We train a non-linear probe on confounder representations for the VQA task
  - Probe's accuracy is 25%
  - Probe's predicted answer distribution has lower entropy than unbiased features' predicted answer distribution

**Acknowledgements**

We thank Peter Hase, Zhuofan Ying, and Jaemin Cho for their useful insights about this work, and the reviewers of this paper for their helpful feedback. This work was supported by AROW911NF2110220, DARPAMCSN66001-19-24031, ONRN00014-23-1-2356, DARPA ECOLE Program No. HR00112390060. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency