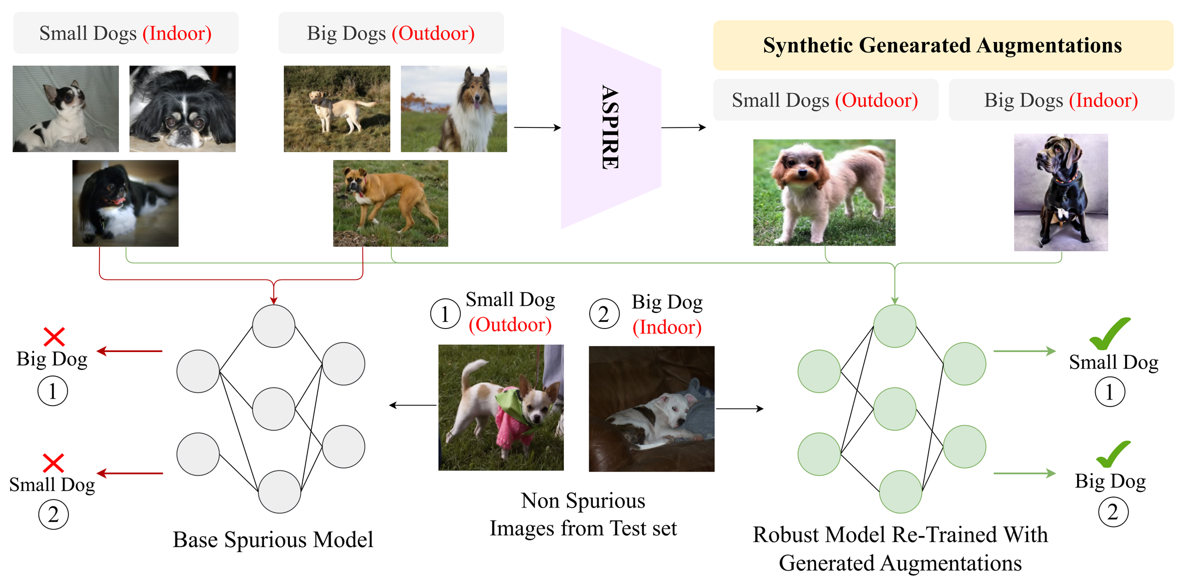
ASPIRE: Language-Guided Data Augmentation for Improving Robustness Against Spurious Correlations

ACL 2024
Bangkok, Thailand

Sreyan Ghosh[1], Chandra Kiran Everu[1], Sonal Kumar[1], Utkarsh Tyagi[1], S Sakshi[1], Sanjoy Chowdhury[1], Dinesh Manocha[1]
[1]University of Maryland, College Park, USA

Introduction & Motivation

What are spurious correlations? Image classifiers often rely on spurious correlations—nonpredictive image features that frequently occur together with class labels in the training data. This leads to poor performance in real-world situations where these spurious features are absent or different.

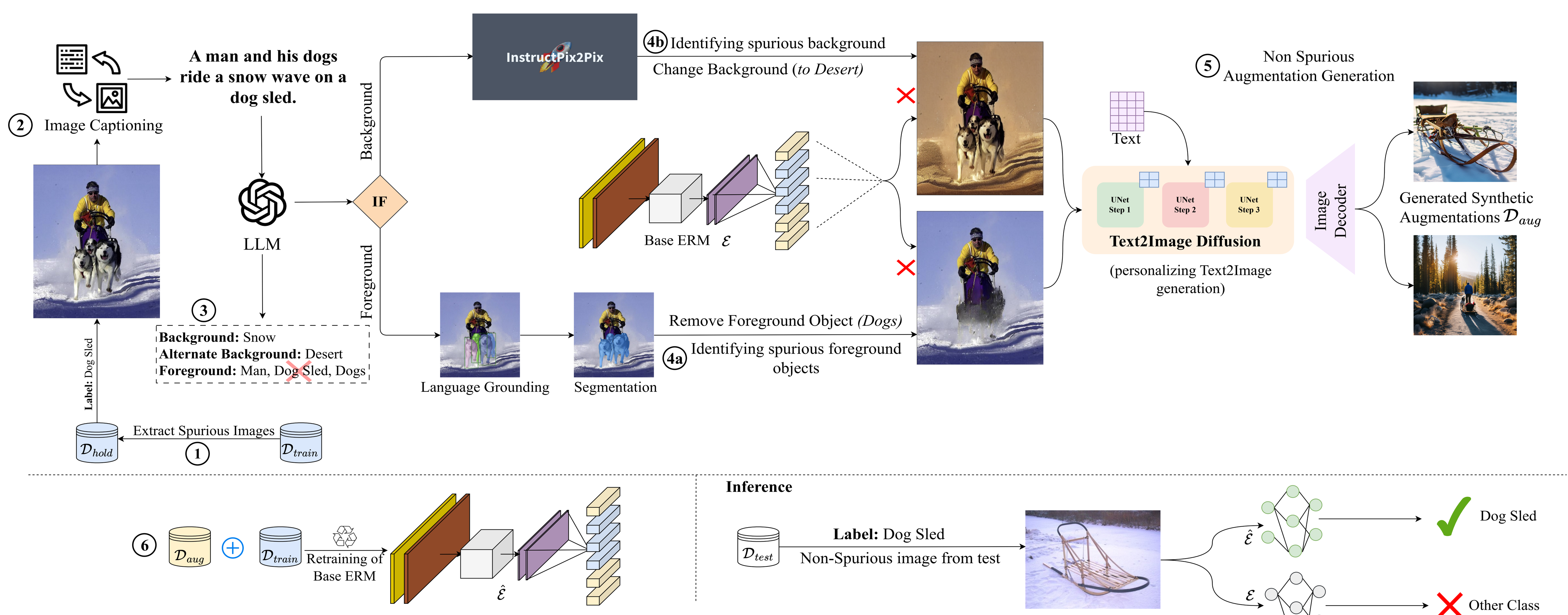


Contribution: We present ASPIRE (Language-Guided Data **A**ugmentation for **SP**ur**I**ous Correlation **RE**moval) that supplments the data with synthetic non-spurious images for learning a robust classifier. ASPIRE employs language-guidance at various steps and does not require existing non-spurious images or group labels for synthetic data generation.

Methodology

To augment existing datasets with non-spurious images, ASPIRE employs a 6-step pipeline to generate synthetic non-spurious images.

1. **Extracting D_{hold} from D using $\mathcal{E}$:** We identify the training examples correctly classified by E and randomly select a small percentage p% to form Dhold. These images contain spurious correlations.
2. **Image Captioning on D_{hold}:** We generate textual descriptions for each image in Dhold to capture foreground and background information.
3. **Extracting objects and backgrounds from captions:** We prompt an LLM to extract the phrases corresponding to foreground objects and the background in the generated captions.
4. **Identifying spurious foreground and background objects:** We remove the objects and the backgrounds one by one using image-editing tools and ask $\mathcal{E}$ to classify it. We collect the edited images that resulted in a wrong prediction.
5. **Non-spurious augmentation generation:** We first fine-tune an image generation model with textual inversion (Gal at al.) on the edited images from the previous step. We then prompt this model to generate non-spurious images.
6. **Re-training the base classifier $\mathcal{E}$:** We add the generated non-spurious images to the training dataset and re-train the image classifier to improve its robustness.

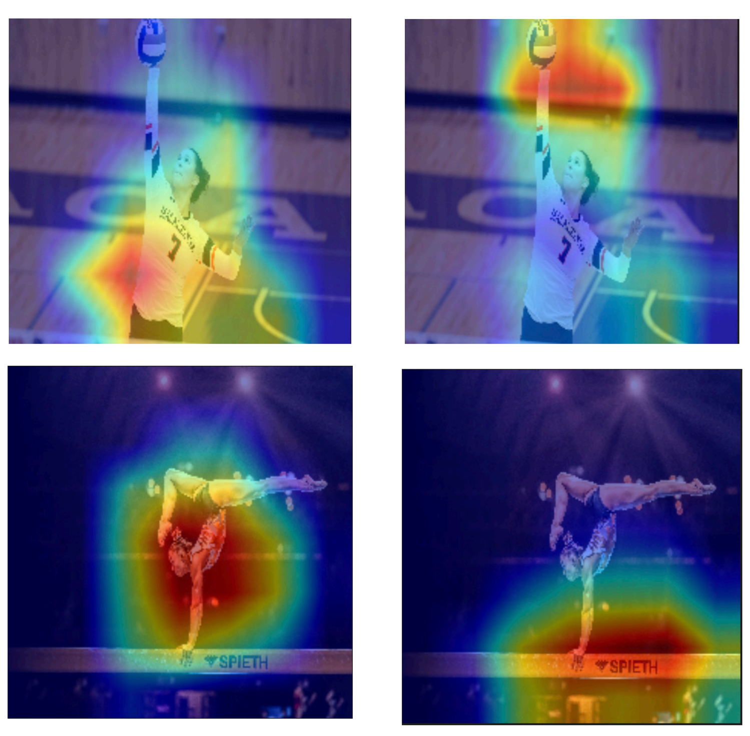


Quantitative Results & GradCam Visualizations

Method	Waterbirds		CelebA		SpucoDogs		Hard ImageNet	
	Worst-group Acc. (%)	Avg Acc. (%)	Worst-group Acc. (%)	Avg Acc. (%)	Worst-group Acc. (%)	Avg Acc. (%)	Worst-group Acc. (%)	Avg Acc. (%)
ERM	74.4	96.9	43.4	95.5	42.3	74.5	12.6	74.3
ERM + Azizi et al.	71.8	**97.1**	36.2	**96.7**	39.6	75.4	10.7	76.7
ERM + Gowal et al.	75.7	85.6	45.7	96.4	46.8	73.7	23.3	83.4
ERM + ASPIRE	**78.7**$_{\pm1.31}$ (+4.3)	89.6$_{\pm1.10}$	**50.5**$_{\pm0.79}$ (+7.1)	95.4$_{\pm1.08}$	**51.6**$_{\pm0.86}$ (+9.3)	**75.5**$_{\pm1.18}$	**50.1**$_{\pm1.26}$ (+37.5)	**96.5**$_{\pm1.32}$
LfF (Nam et al., 2020)	78.0	91.2	77.2	85.1	70.2	80.8	58.8	92.5
LfF + Azizi et al.	74.2	**92.3**	74.4	85.7	67.5	**81.6**	54.3	92.6
LfF + Gowal et al.	81.0	89.3	78.2	85.8	72.9	80.9	60.3	92.7
LfF + ASPIRE	**83.2**$_{\pm0.20}$ (+5.2)	91.4$_{\pm1.12}$	**81.7**$_{\pm0.43}$ (+4.5)	**86.3**$_{\pm1.25}$	**75.4**$_{\pm0.38}$ (+5.2)	80.9$_{\pm0.31}$	**63.8**$_{\pm0.30}$ (+5.0)	**92.7**$_{\pm0.21}$
Group DRO (Sagawa et al., 2019)	91.4	93.5	88.9	92.9	75.4	82.8	65.6	91.8
Group DRO + Azizi et al.	88.2	94.1	85.6	93.2	71.7	84.1	62.8	**92.9**
Group DRO + Gowal et al.	91.6	94.2	89.8	93.7	76.3	83.4	65.5	91.7
Group DRO + ASPIRE	**92.8**$_{\pm0.49}$ (+1.4)	**94.6**$_{\pm0.49}$	**90.1**$_{\pm1.08}$ (+1.2)	**94.3**$_{\pm0.92}$	**78.7**$_{\pm1.26}$ (+3.3)	**84.3**$_{\pm0.58}$	**67.4**$_{\pm1.01}$ (+1.8)	92.4$_{\pm0.59}$
JTT (Liu et al., 2021b)	86.7	93.3	81.1	88.0	73.0	80.4	63.5	90.6
JTT + Azizi et al.	83.2	**94.9**	78.3	90.2	71.8	**82.2**	61.4	92.4
JTT + Gowal et al.	87.5	94.2	83.8	89.6	74.1	81.1	64.1	91.9
JTT + ASPIRE	**90.2**$_{\pm1.16}$ (+3.5)	94.6$_{\pm1.24}$	**85.7**$_{\pm0.64}$ (+4.6)	**91.6**$_{\pm0.75}$	**75.5**$_{\pm1.33}$ (+2.5)	81.7$_{\pm1.12}$	**65.2**$_{\pm0.54}$ (+1.7)	**92.9**$_{\pm0.82}$
DivDis (Lee et al., 2022)	85.6	87.3	55.0	90.8	39.3	65.5	15.5	71.8
DivDis + Azizi et al.	84.2	**88.6**	53.7	**92.2**	37.5	66.4	13.7	77.2
DivDis + Gowal et al.	86.3	87.4	56.1	91.2	42.1	66.3	23.9	76.9
DivDis + ASPIRE	**87.2**$_{\pm0.49}$ (+1.6)	87.8$_{\pm0.84}$	**57.4**$_{\pm1.13}$ (+2.4)	91.6$_{\pm0.66}$	**43.6**$_{\pm1.48}$ (+4.3)	**67.1**$_{\pm1.22}$	**35.5**$_{\pm0.82}$ (+20.0)	**77.6**$_{\pm0.34}$
SUBG (Idrissi et al., 2022)	88.9	91.2	86.2	89.1	74.2	81.5	62.3	90.9
SUBG + Azizi et al.	86.5	91.8	85.4	**91.3**	72.3	81.6	60.5	**92.9**
SUBG + Gowal et al.	89.7	91.7	88.2	89.9	75.6	81.7	64.8	91.6
SUBG + ASPIRE	**90.7**$_{\pm0.62}$ (+1.8)	**92.1**$_{\pm0.88}$	**88.6**$_{\pm1.37}$ (+2.4)	90.1$_{\pm0.64}$	**77.5**$_{\pm0.73}$ (+3.3)	**83.5**$_{\pm0.92}$	**66.7**$_{\pm1.22}$ (+4.4)	92.4$_{\pm0.63}$
Correct-n-Contrast (Zhang et al., 2022)	88.7	90.6	88.1	89.4	73.7	81.2	60.5	91.7
Correct-n-Contrast + Azizi et al.	84.3	**93.4**	85.2	91.3	70.8	**85.6**	58.7	**93.3**
Correct-n-Contrast + Gowal et al.	89.1	91.7	88.7	90.6	74.9	82.6	63.2	92.1
Correct-n-Contrast + ASPIRE	**90.8**$_{\pm1.18}$ (+2.1)	92.6$_{\pm1.48}$	**89.9**$_{\pm1.45}$ (+1.8)	**91.3**$_{\pm0.28}$	**76.8**$_{\pm1.10}$ (+3.1)	83.1$_{\pm1.04}$	**65.9**$_{\pm0.94}$ (+5.4)	91.9$_{\pm1.11}$
MaskTune (Taghanaki et al., 2022)	78.0	91.2	77.9	92.5	31.6	59.2	33.0	58.5
MaskTune + Azizi et al.	75.8	**93.4**	73.3	**93.5**	26.3	**63.4**	28.9	**61.3**
MaskTune + Gowal et al.	79.3	85.2	78.8	88.1	35.2	60.7	35.3	55.8
MaskTune + ASPIRE	**81.6**$_{\pm1.28}$ (+3.6)	91.3$_{\pm0.54}$	**81.2**$_{\pm0.22}$ (+3.3)	92.8$_{\pm0.38}$	**37.5**$_{\pm0.33}$ (+5.9)	61.3$_{\pm1.05}$	**41.0**$_{\pm0.61}$ (+8.0)	60.2$_{\pm0.37}$
DFR (Kirichenko et al., 2023)	81.7	90.1	80.5	85.3	78.8	83.2	33.3	95.7
DFR + Azizi et al.	78.6	**92.7**	78.3	88.4	72.1	85.1	29.5	**96.3**
DFR + Gowal et al.	83.1	86.5	83.4	86.2	81.0	84.4	35.2	92.0
DFR + ASPIRE	**85.3**$_{\pm1.34}$ (+3.6)	91.7$_{\pm0.79}$	**85.5**$_{\pm0.64}$ (+5.0)	**89.5**$_{\pm0.51}$	**84.2**$_{\pm0.83}$ (+5.4)	**87.5**$_{\pm0.57}$	**37.5**$_{\pm0.39}$ (+4.2)	96.2$_{\pm0.91}$

ASPIRE substantially improves the worst-group accuracy of all baselines.

Paper and Code



GradCAM visualizations before (left) and after (right) augmentation for ImageNet classes VollyeBall (top) Balance Beam (bottom).

Qualitative Results

Examples of Original Images, Edited Images from the ASPIRE pipeline and Generated Augmentations.