A The effect of text prompt

Table 7: Under CelebA and CLIP (ViT-B/32), the average accuracy and worst-case accuracy over sub-populations with varying classification text prompt and debiasing text prompt. (%)

Classification text prompt And Debiasing text prompt	Method	Avg. Acc.	Worst-case Acc.
Input space sub-group: {female, male}			
"a photo of a {not blond, blond} hair people"	Zero-shot	85.2	70.6
And "a photo of a {female, male} people"[1]	L-DRO	83.6±0.3	**79.2±1.3**
Or "a photo of a {female, not female} people"	L-DRO	88.8±0.3	65.0±0.9
Or "a photo of a {male, not male} people"	L-DRO	89.4±0.3	37.8±2.2
Or "a photo of a {[female, not female], [male, not male]} people"	L-DRO	89.9±0.3	60.7±2.4
Input space sub-group: {old, young}			
"a photo of a {not blond, blond} hair people"	Zero-shot	85.1	73.5
And "a photo of a {old, young} people"	L-DRO	84.4±0.3	74.5±1.5
Or "a photo of a {old, not old} people"	L-DRO	82.2±0.05	78.2±0.6
Or "a photo of a {young, not young} people"	L-DRO	91.3±0.1	51.6±1.9
Or "a photo of a {[old, not old], [young, not young]} people"	L-DRO	88.0±0.7	**84.3±1.6**

[1] " " denotes default choice.

Table 8: Under Waterbirds and CLIP (ViT-B/32 and RN50), the Average Accuracy (Avg.Acc.) and Worst-Case Accuracy (W.C.Acc.) over sub-populations with varying classification text prompt and debiasing text prompt.(%)

Classification text prompt And Debiasing text prompt	Method	RN50 (Avg.Acc & W.C.Acc.)	ViT-B/32 (Avg.Acc & W.C.Acc.)
"a {landbird, waterbird}"	Zero-shot	68.1 & 43.4	74.8 & 56.8
And "{water, land}"	L-DRO	72.6±1.2 & 49.5±2.7	75.1±1.6 & 56.6±2.6
Or "{water, forest}"	L-DRO	74.9±1.2 & **57.6±2.6**	77.6±0.5 & **64.8±0.8**
"photo of {landbird, waterbird}"	Zero-shot	66.3 & **43.2**	66.1 & 39.6
And "photo of {water, land}"	L-DRO	63.3±1.4 & 41.0±3.3	76.0±0.7 & **61.9±1.4**
"photo of a {landbird, waterbird}"	Zero-shot	78.1 & 34.0	68.7 & 43.6
And "photo of a bird on {water, land}"	L-DRO	74.3±0.9 & **57.9±1.8**	71.8±2.5 & **49.7±4.7**
"photo of a {landbird, waterbird}"	Zero-shot	78.1& 34.0	68.7 & 43.6
And "photo of a bird on {water, land} background"	L-DRO	77.4±1.3 & **62.7±2.8**	70.0±3.2 & **46.9±4.8**
"a photo of a {landbird, waterbird}"	Zero-shot	76.8 & 40.8	69.7 & 45.5
And "a photo of a bird on {water, land}"	L-DRO	73.9±3.0 & **54.4±4.6**	71.4±3.4 & **50.2±5.2**
"a photo of a {landbird, waterbird}"	Zero-shot	76.8 & 40.8	69.7& **45.5**
And "a photo of a bird on {water, land} background"	L-DRO	75.3±0.8& **58.1±1.7**	67.5±2.9 & 43.9±4.2

B EFFECTS OF TWO-PHASE TRAINING ON DRO METHODS

Table 9: The average accuracy and worst-case accuracy over different datasets and methods.[1] (%)

Dataset	Architecture	Method	Average Acc.	Worst-case Acc.
CelebA	$I \triangleright A^2 \triangleright T$	ERM	95.3±0.1	44.2±2.5
	$I \triangleright A^2 \triangleright T$	CVaR DRO	86.6±1.0	11.7±9.7
	$I \triangleright A^2 \triangleright T$	χ^2-DRO	84.2±8.3	61.3±8.5
	$I \triangleright A^2 \triangleright T$	CVaR DRO*	84.8±4.9	67.1±10.4
	$I \triangleright A^2 \triangleright T$	χ^2-DRO*	87.4±4.5	72.0±9.6

[1] Keeping the same settings with Table 6. And $\star$ denotes using the same two-phase training strategy with JTT, and the method without $\star$ denotes the original version (mini-batch) of CVaR DRO and χ^2-DRO.

C TEXT PROMPT FOR CLIP (VIT-L/14)

Table 10 reveals that the effectiveness of text prompts on CLIP (ViT-B/32) does not consistently translate to high performance on CLIP (ViT-L/14). Employing "a photo of a { } people" as the prompt for CLIP (ViT-L/14) achieves a more reasonable performance, and the introduction of L-DRO further enhances the overall performance in this context.

Table 10: Under CelebA and CLIP (ViT-L/14), the average accuracy and worst-case accuracy over sub-populations with varying classification text prompt and debiasing text prompt [1].(%)

Classification text prompt And Debiasing text prompt	Method	Average Acc.	Worst-case Acc.
"a photo of {not blond, blond}"	Zero-shot	39.1	28.8
"photo of a {not blond, blond}"	Zero-shot	75.9	65.2
"a photo of a {not blond, blond}"	Zero-shot	64.0	39.7
"photo of a {not blond, blond} people"	Zero-shot	80.7	77.9
"a photo of a {not blond, blond} people"	Zero-shot	85.4	76.1
"photo of a {not blond, blond} hair people"	Zero-shot	78.5	70.7
"a photo of a {not blond, blond} hair people"	Zero-shot	75.6	64.5
And "a photo of a {male, female} people"[2]	L-DRO	85.9±0.9	**79.7±1.9**

[1] classification and debiasing text prompts use the same structure, e.g., "a photo of a { } people" will be used for both classification and debiasing text prompts.

[2] "_" denotes default choice.