Towards Unbiased Continual Learning:
Avoiding Forgetting in the Presence of Spurious Correlations
Supplementary Material

Giacomo Capitani, Lorenzo Bonicelli, Angelo Porrello,
Federico Bolelli, Simone Calderara, and Elisa Ficarra

Università degli Studi di Modena e Reggio Emilia, Italy

{name.surname}@unimore.it

1. Datasets Details

B-Celeba Splits [8 Tasks]. We created 8 splits based on the original dataset CelebA [5]. Since **B-CelebA1** Fig. [1-8] and **B-CelebA2** Fig. [9-16] contain multiple attributes, we bolded the target used during a specific task. The latent attribute z used to introduce spurious correlations is gender (blue *Male* and red *Female*, respectively, in the plots). In our experiments, p_{corr} is set to 0.95, indicating that 95% of images with a specific attribute y (e.g., *Blond Hair*) is of a particular latent attribute z (gender).

For each task, its training set comprises 4 480 images, with 2 240 labeled as $y = 0$ and 2 240 labeled as $y = 1$, as well as 2 240 labeled as $z = 0$ and 2 240 labeled as $z = 1$. The test sets, one for each task, are balanced in terms of the y label the task pertains to and, as a result, the label on which the model is evaluated. More in detail, in the test set, each group $g = (y, z)$ consists of 100 images. In cases where there are not enough elements in the dataset to ensure this allocation, we ensured the same ratio but with fewer elements (Heavy Makeup - *Task* 1 in B-CelebA1). For all datasets, an image can be selected for only one task.

Biased Camelyon [4 Tasks]. To make the splits of B-Camelyon Fig. 17, we based on Camelyon17 [1], employing the version present in WILDS benchmark [6]. During training, images are balanced with respect to tumor/no-tumor. In this case, we modeled that hospital 0 is correlated with the presence of a tumor (95% of tumoral images came from the hospital 0), and hospital 1 is correlated with the absence of a tumor. Also, hospitals 2 and 3 correlate with "no tumor" but are in the minority compared to hospital 1 (95% of no-tumoral images came from the hospital 1 + 2 + 3). Each hospital has an equal number of tumor and non-tumor images during testing. Among these, hospital 4, not present in training, serves as the *o.o.d.* test.

2. Training Procedure Details

In Algorithm 1, we describe our training procedure. We use the torchvision ResNet-18 [4] with pre-trained weights from ImageNet to initialize the feature extractor $\mathcal{F}_{pre} : X \rightarrow R^{512}$, employed for clustering at the start of each task. For the sake of simplicity, β is a function to get the bin of an element based on its loss $\mathcal{L}_{target}$ computed in step 12, and γ is a function to get the budget of a specific bin.

3. Additional Experiments

Using ViT as a Backbone. We conducted experiments on B-Celeba1 using a ViT-B/16 [3], freezing the backbone $\mathcal{F}_{pre}$ (ImageNet-21k weights) and training only the task-specific and cluster classifiers. For LwS, we observed an improvement compared to ResNet-18 of $+2.00\%$ on Acc_{worst} and $+1.87\%$ on Acc_{avg}. Conversely, for SGD, we noted a decrease of -5.14% on Acc_{worst} and an improvement of $+3.85\%$ on Acc_{avg}. Also, in Tab. 1 are shown experiments on B-Celeba1 and B-Celeba2 using backbone pre-trained on ImageNet21k with two different strategies: MoCo v3 [2] and supervised.

Algorithm 1 *Learning without Shortcuts (LwS)*

Require: learning rate η_θ, momentum m, tascs T, number of epochs E, number of batches B, number of clusters $|c|$, k-means cluster assignment $\mathcal{A}$, pre-trained feature extractor $\mathcal{F}_{pre}$, network active parameters θ, β function, γ function.

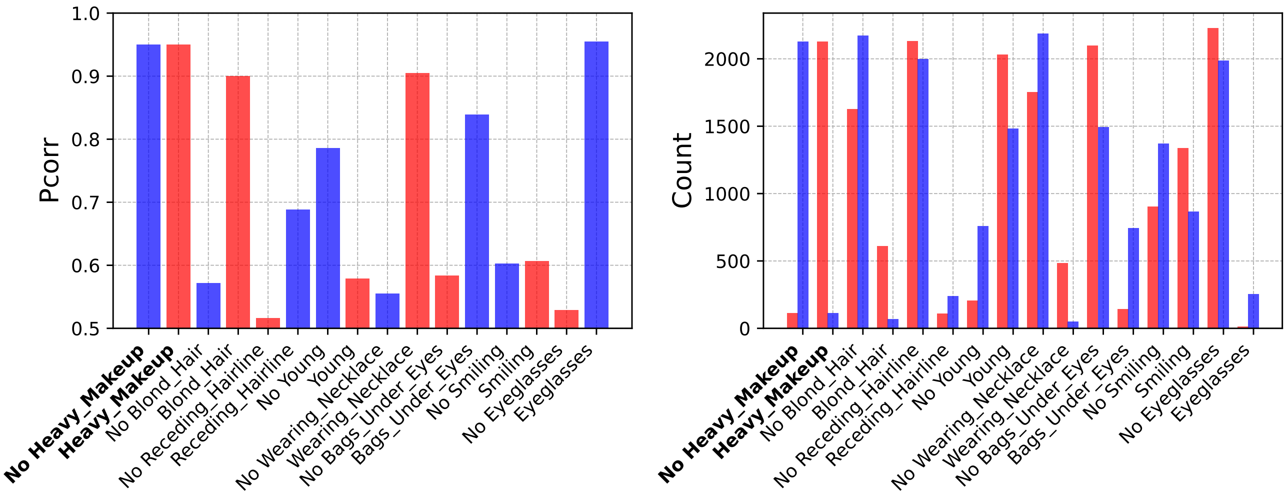
for $t = 1, ..., T$ **do**
    Step 1: Cluster Assignment ▷ Init task
    for $c = 1, ..., |c|$ **do**
        $P_c = \{\mathcal{A}(\mathcal{F}_{pre}(x_i)) = c\}$
        $N_c = |P_c|$

    Step 2: Debiased Training
    for $e = 1, ..., E$ **do**
        if $e == 5$ **then**
            $\mathcal{L}_{target}(D_t)$ ▷ Calculate loss for all elements
        for $b = 1, ..., B$ **do**
            $(x_i) \sim D_t$ ▷ Sample from current task
            $\alpha_i \leftarrow \omega_c$
            $\alpha \leftarrow (\alpha_1, ..., \alpha_{|b|})$
            $\alpha \leftarrow \frac{\alpha}{\sum_{i=1}^{|b|} \alpha_i}$
            $\mathcal{L}_{stream} \leftarrow \frac{1}{|b|} \sum_{i=1}^{|b|} \alpha_i \nabla \mathcal{L}_{target} + \nabla \mathcal{L}_{cluster}$
            for $c = 1, ..., c$ **do** ▷ Update weights w_c
                $\omega_c \leftarrow (1-m)\omega_c + \frac{m}{N_c} \sum_{(x) \in P_c} \mathcal{L}_{target} + \mathcal{L}_{cluster}$
            if $|\beta(x_i)| < \gamma(\beta(x_i))$ and $e >= 5$ **then**
                $\mathcal{M} \leftarrow x_i$ ▷ Memory insertion
            if $\mathcal{M}$ is not empty **then**
                $(x_m) \sim \mathcal{M}$ ▷ Sample from buffer $\mathcal{M}$
                $\mathcal{L}_{buffer} \leftarrow \frac{1}{|B|} \sum_{i=1}^{|B|} \nabla \mathcal{L}_{target} + \nabla \mathcal{L}_{cluster} + \nabla \mathcal{L}_{KD}$
                $\theta \leftarrow \theta - \eta_\theta(\mathcal{L}_{stream} + \mathcal{L}_{buffer})$
            else
                $\theta \leftarrow \theta - \eta_\theta \mathcal{L}_{stream}$

Table 1. Results of various approaches with varying pre-training strategies.

Pre-training	Method	B-CelebA1		B-CelebA2	
		Acc$_{worst}$[%]	Acc$_{avg}$[%]	Acc$_{worst}$[%]	Acc$_{avg}$[%]
ImageNet-21K [**Supervised**]	LwS	**54.44**	**72.97**	**36.0**	**68.53**
	CFIX + replay	17.25	61.25	14.12	60.31
	DER++	15.25	59.45	10.5	57.09
ImageNet-21K [**MoCoV3**]	LwS	**44.04**	**67.53**	**32.75**	**65.31**
	CFIX + replay	20.37	60.98	17.62	61.47
	DER++	18.5	59.71	18.0	61.22

Task 1 - Train Split



Task 1 - Test Split

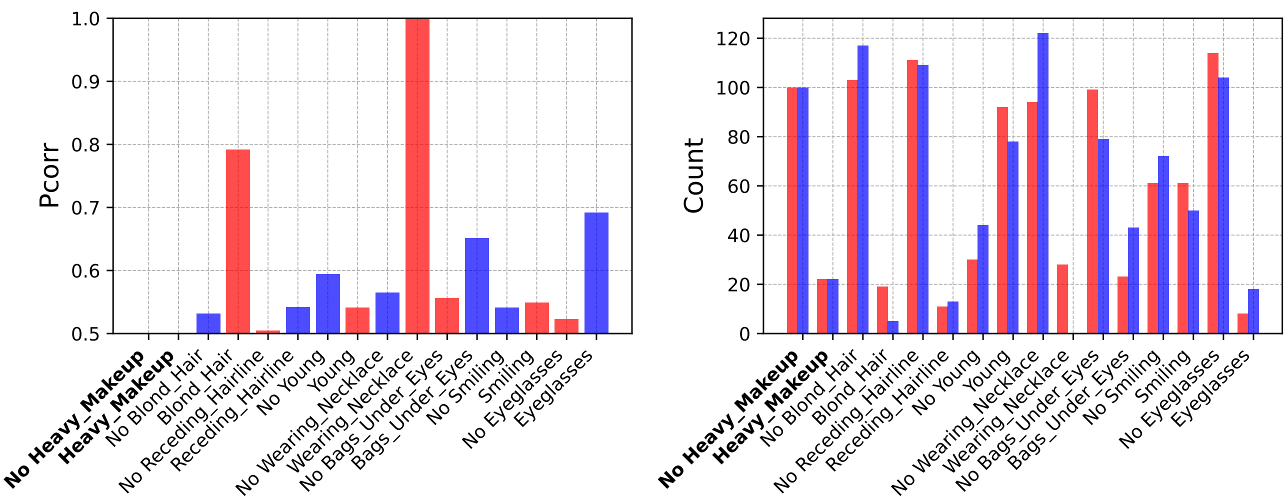
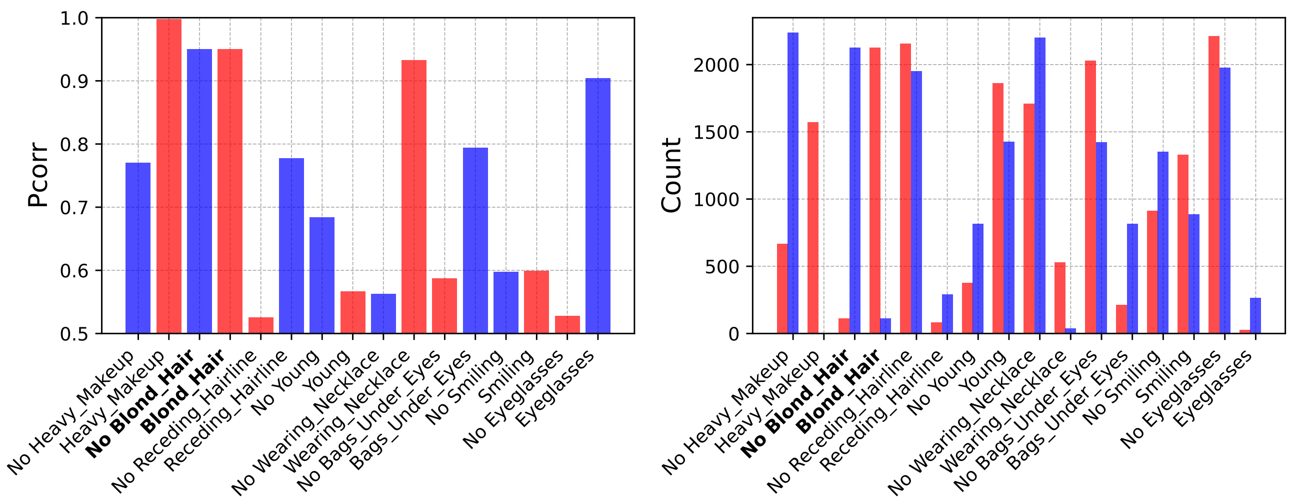


Figure 1. Task 1

***Task* 2 - Train Split**



***Task* 2 - Test Split**

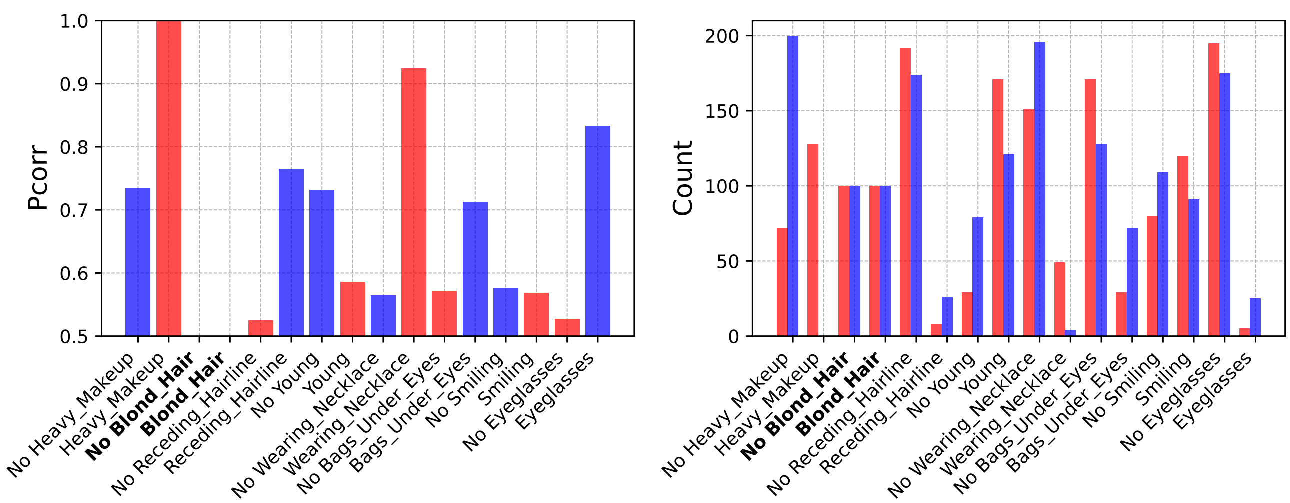
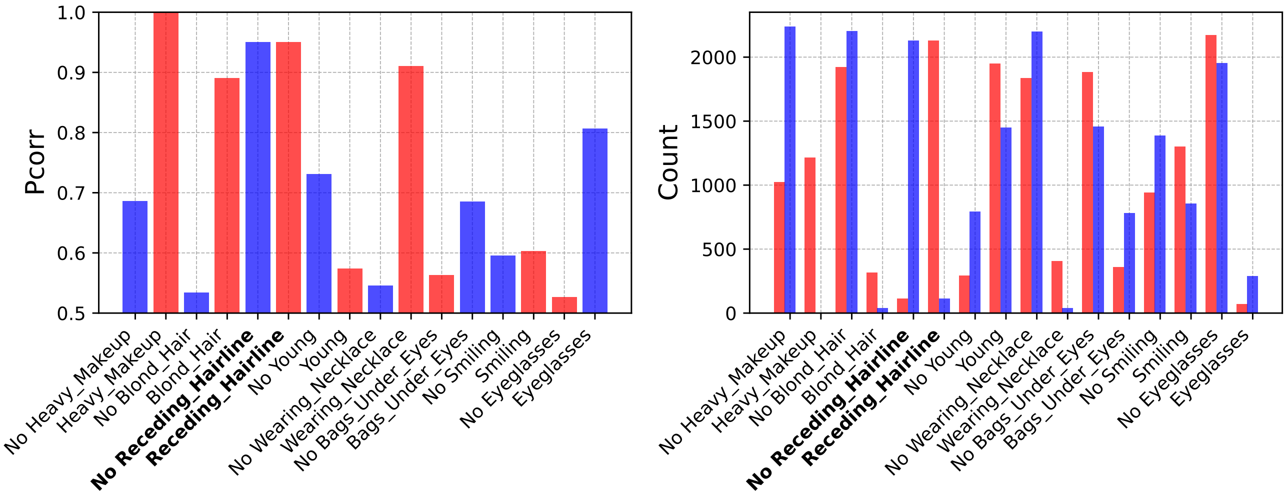


Figure 2. Task 2

Task 3 - Train Split



Task 3 - Test Split

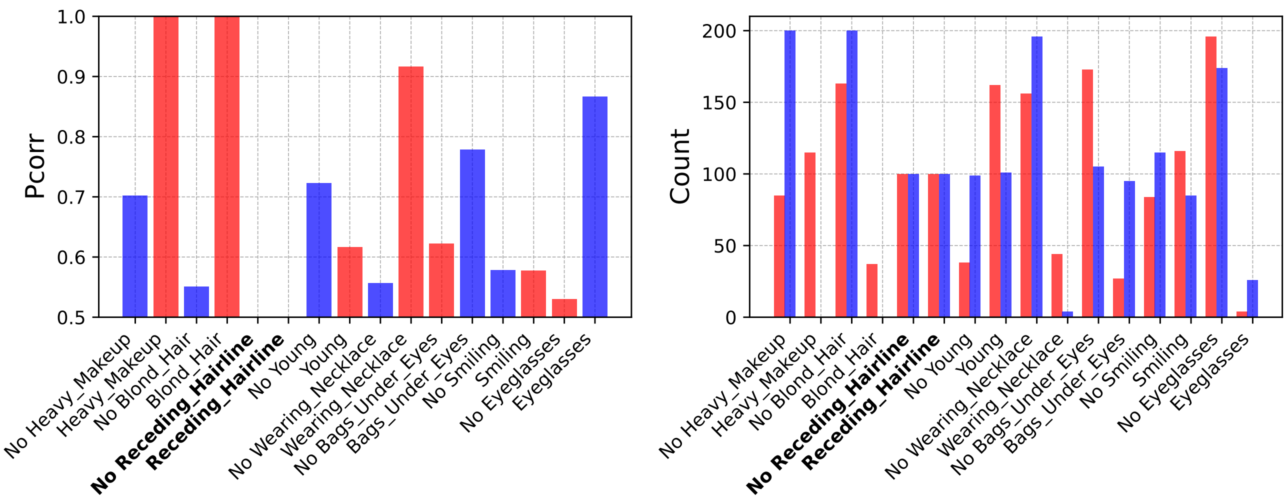
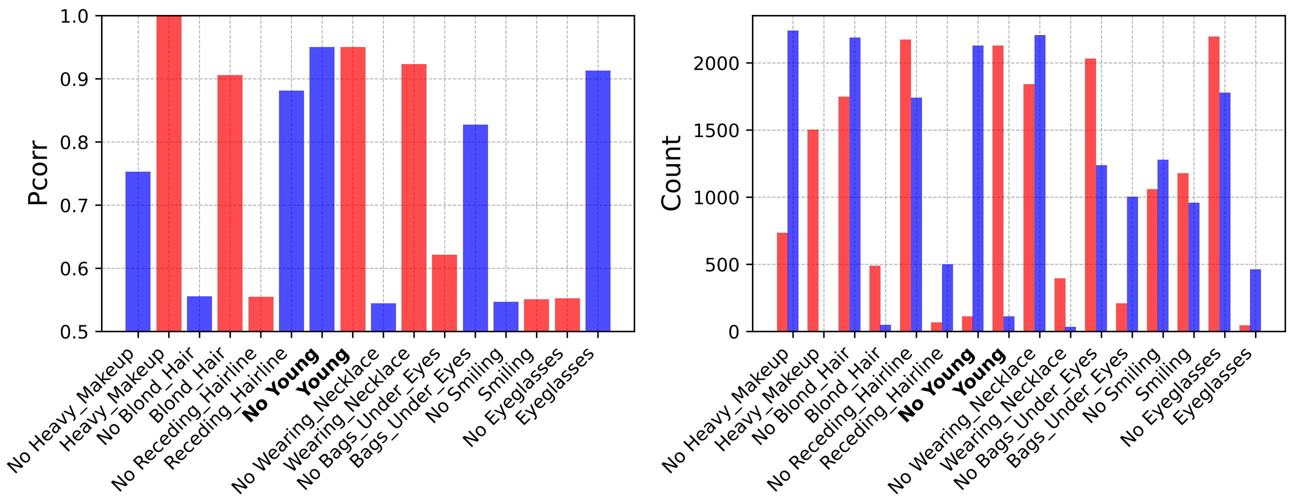


Figure 3. Task 3

***Task* 4 - Train Split**



***Task* 4 - Test Split**

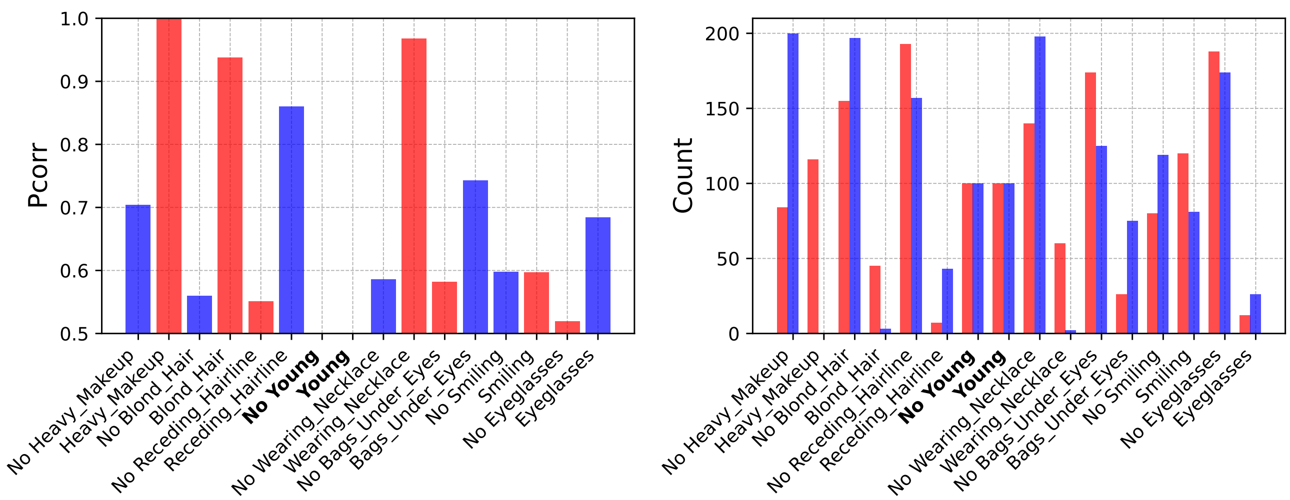
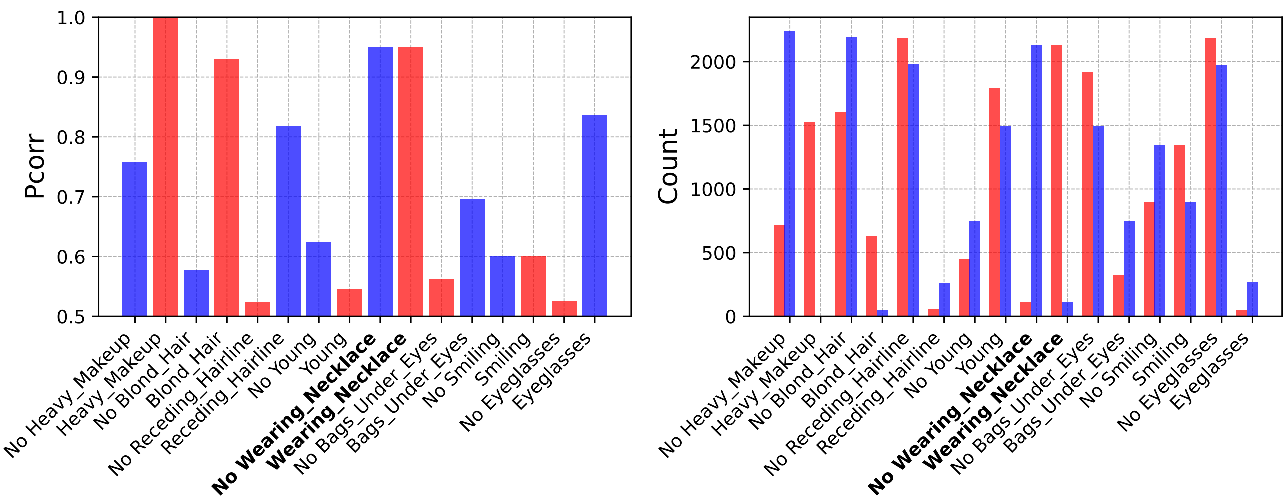


Figure 4. Task 4

Task **5 - Train Split**



Task **5 - Test Split**

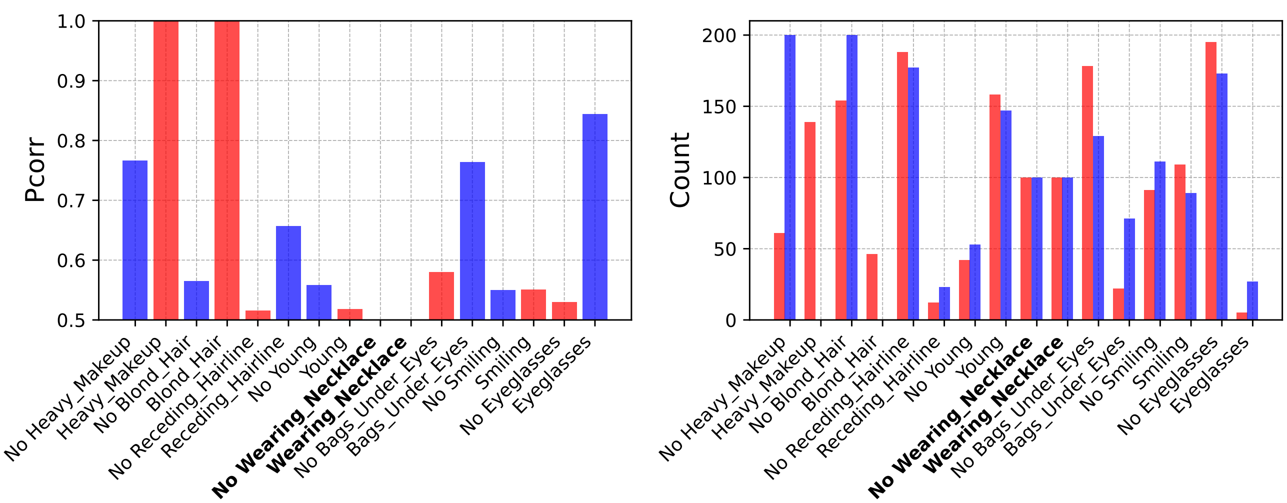
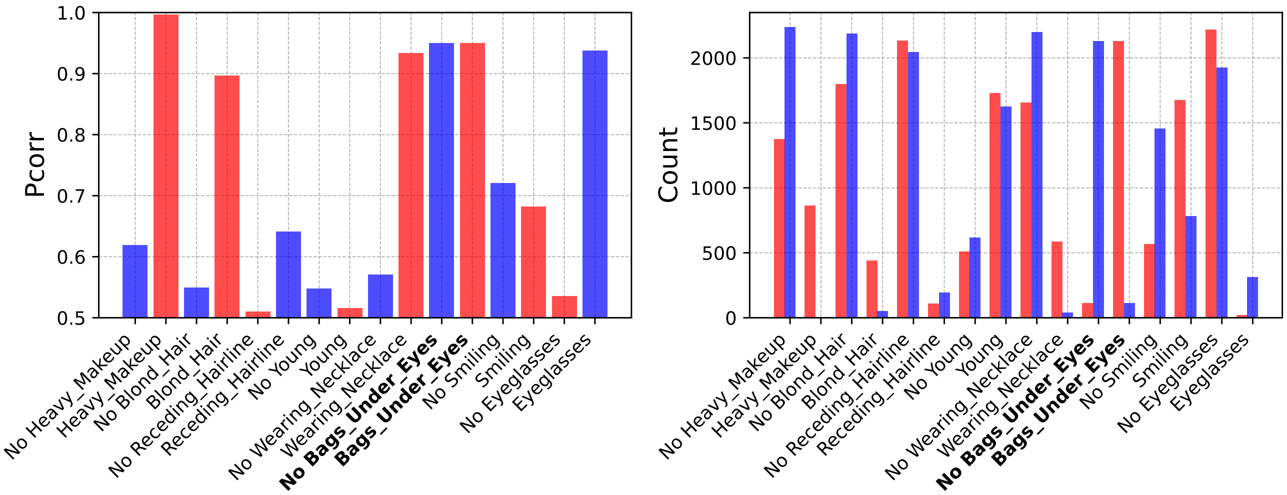


Figure 5. Task 5

***Task* 6 - Train Split**



***Task* 6 - Test Split**

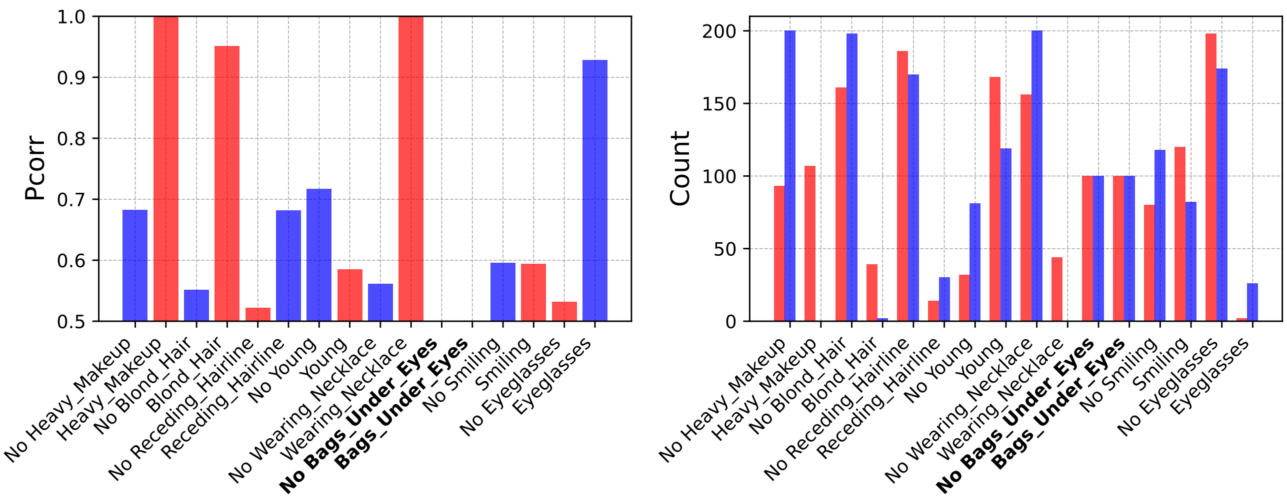
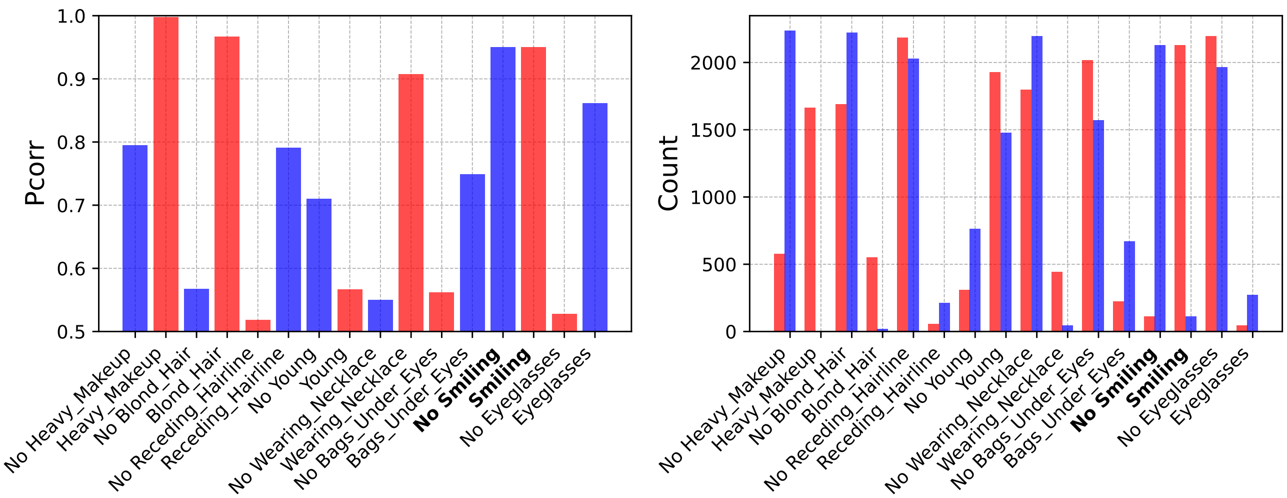


Figure 6. Task 6

Task 7 - Train Split



Task 7 - Test Split

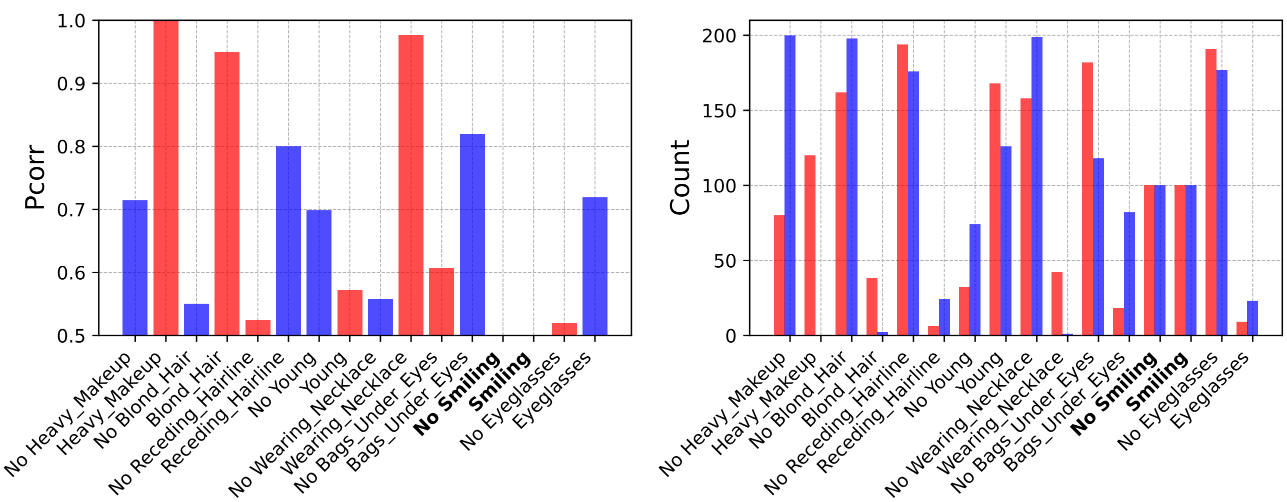
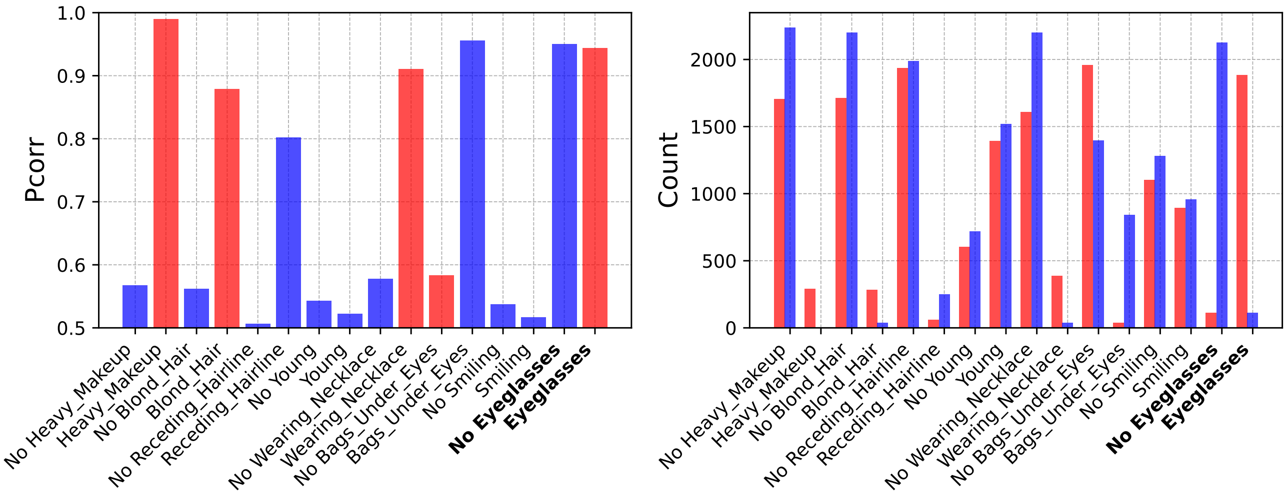


Figure 7. Task 7

Task 8 - Train Split



Task 8 - Test Split

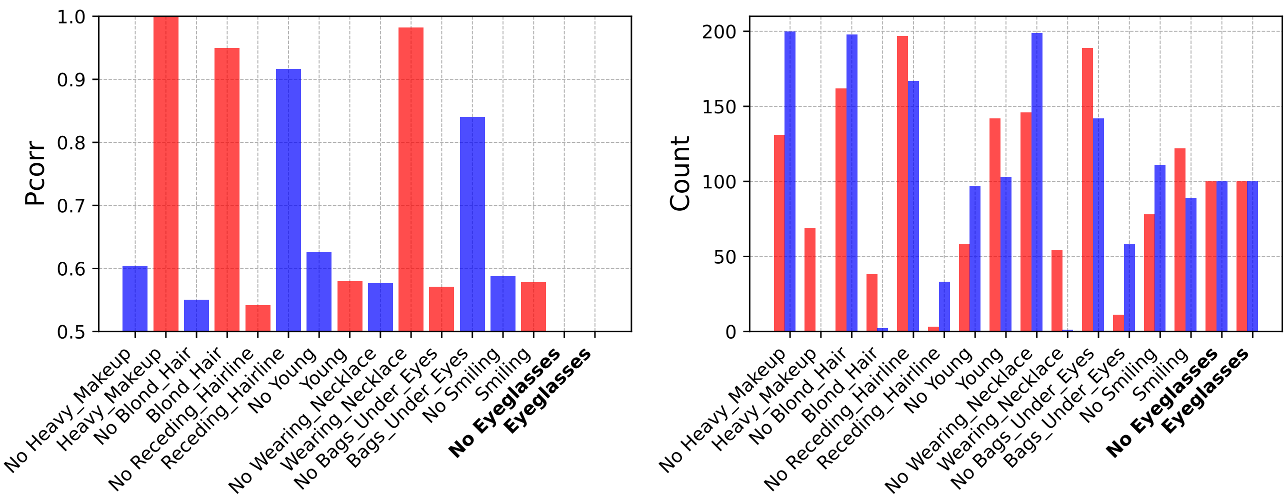
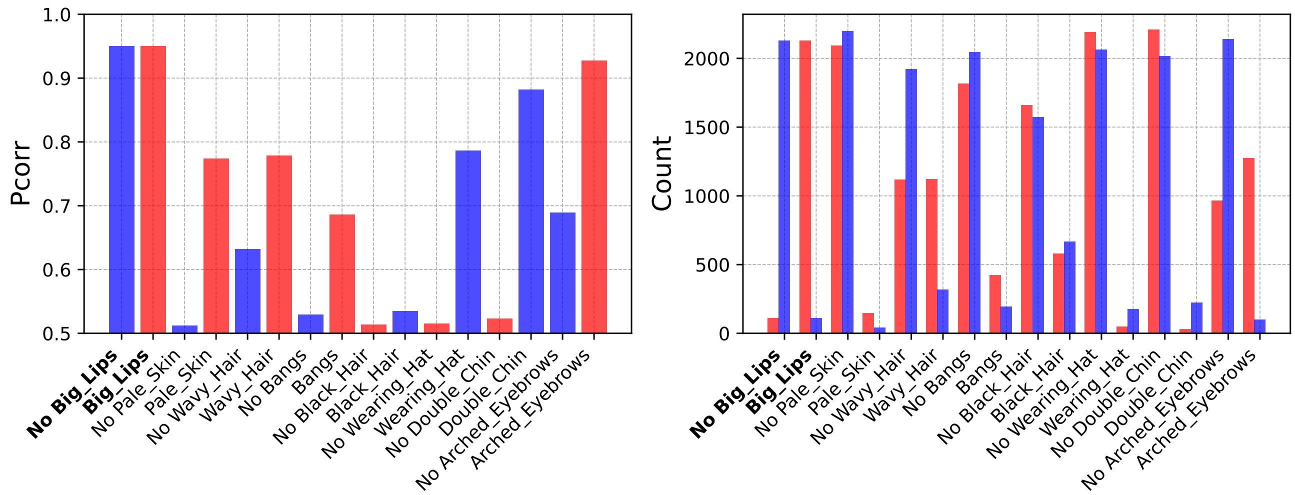


Figure 8. Task 8

***Task* 1 - Train Split**



***Task* 1 - Test Split**

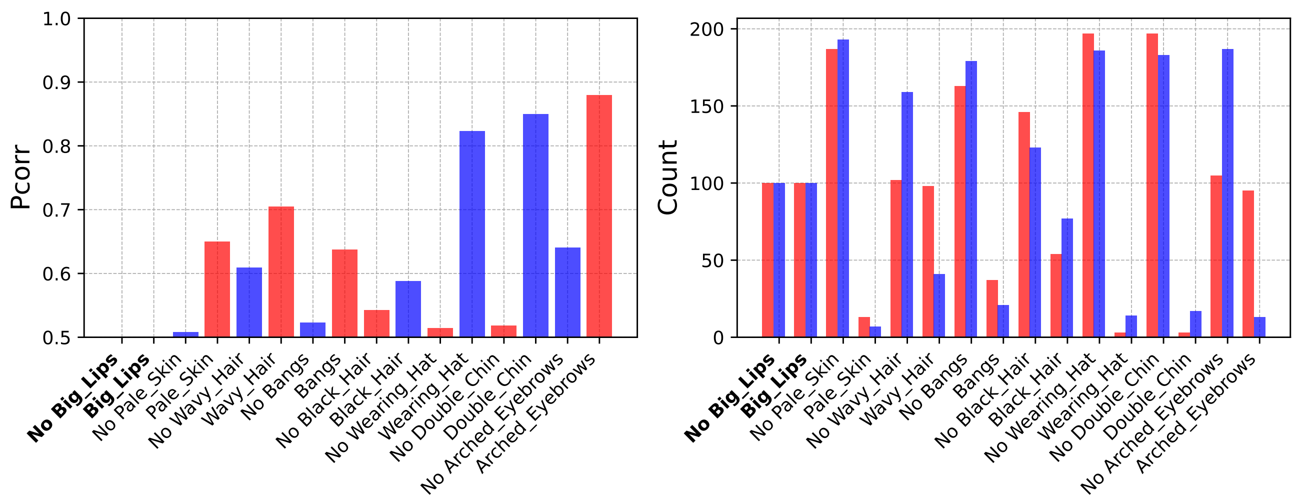
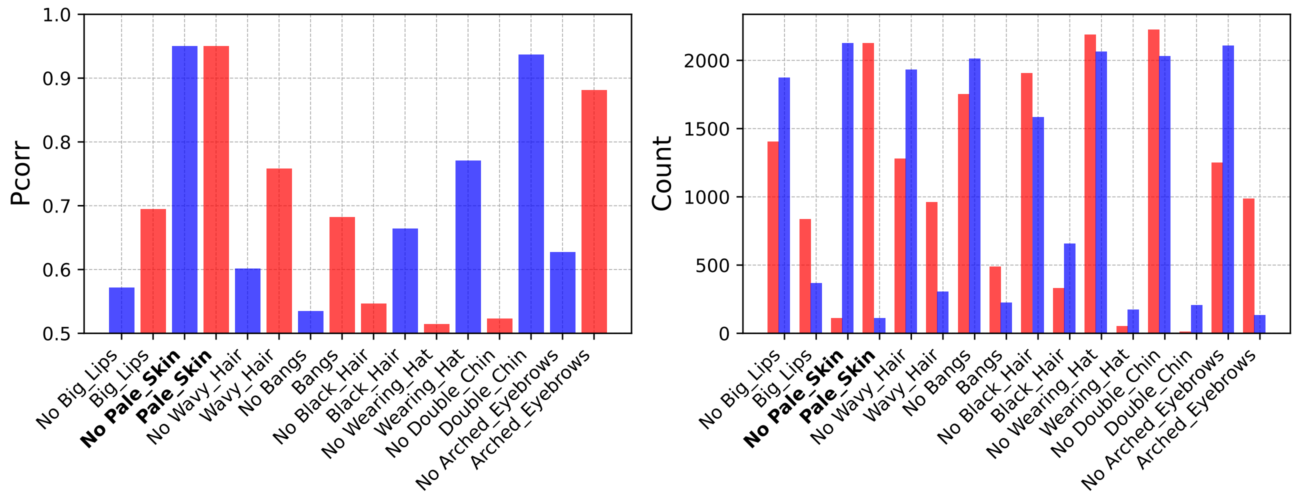


Figure 9. Task 1

Task 2 - Train Split



Task 2 - Test Split

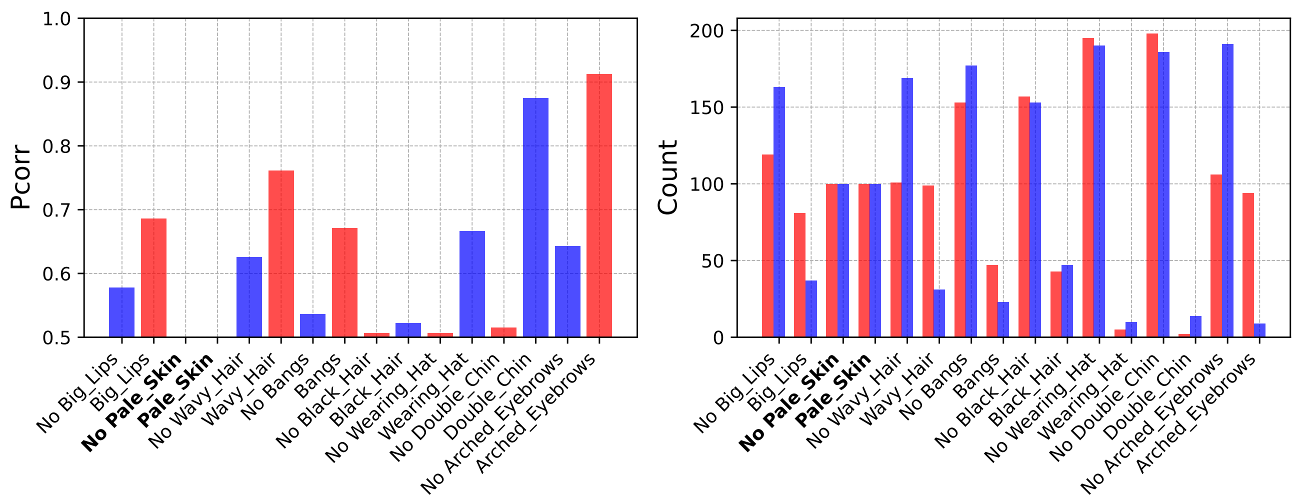


Figure 10. Task 2

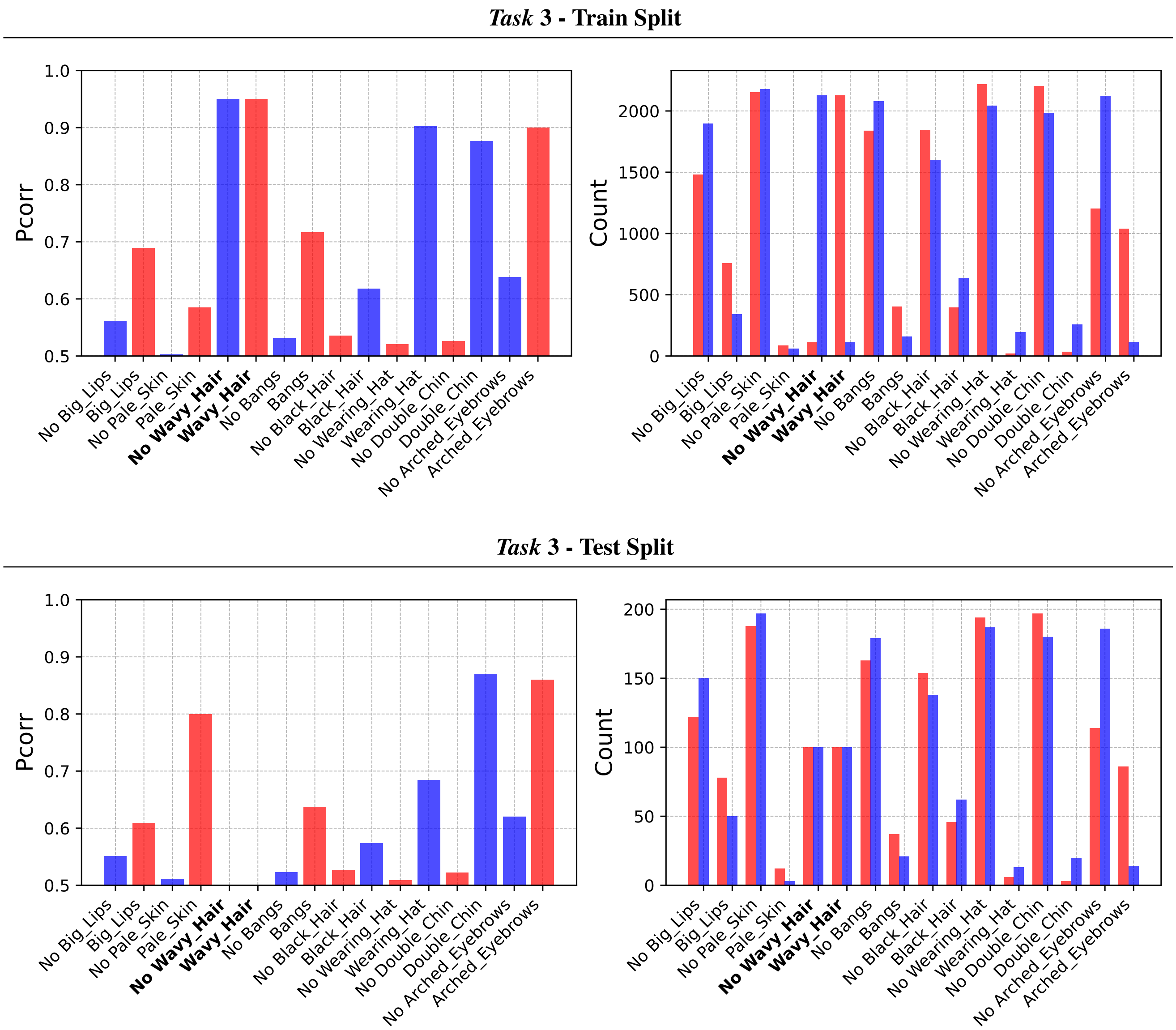
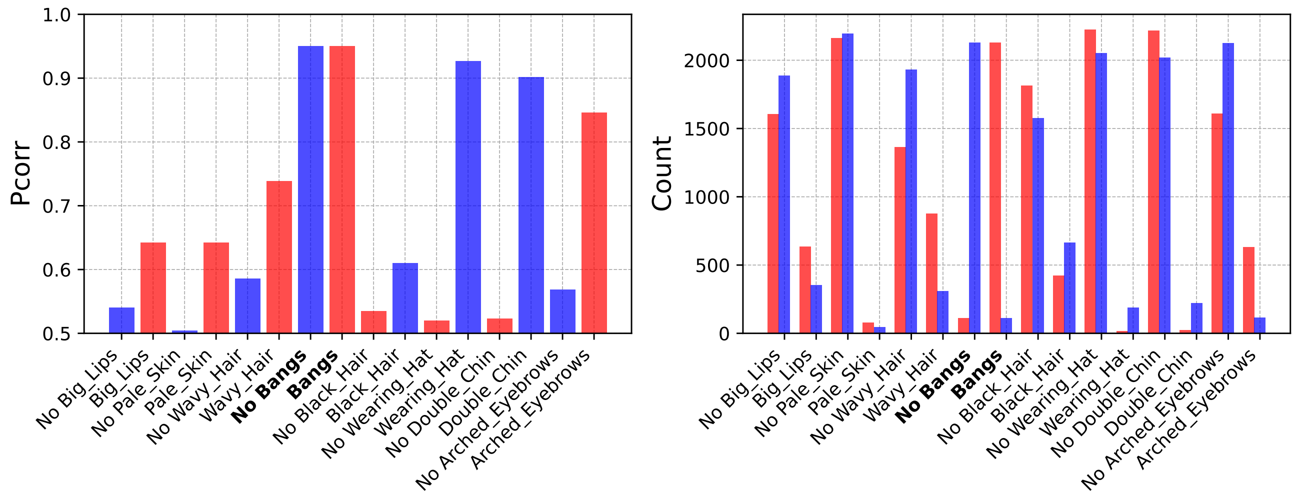


Figure 11. Task 3

Task 4 - Train Split



Task 4 - Test Split

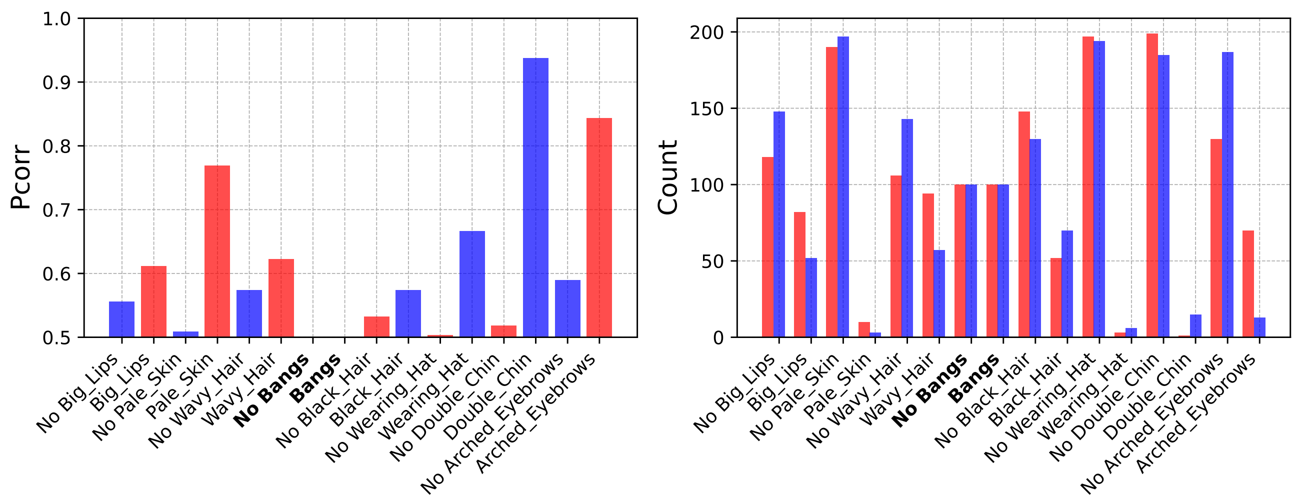
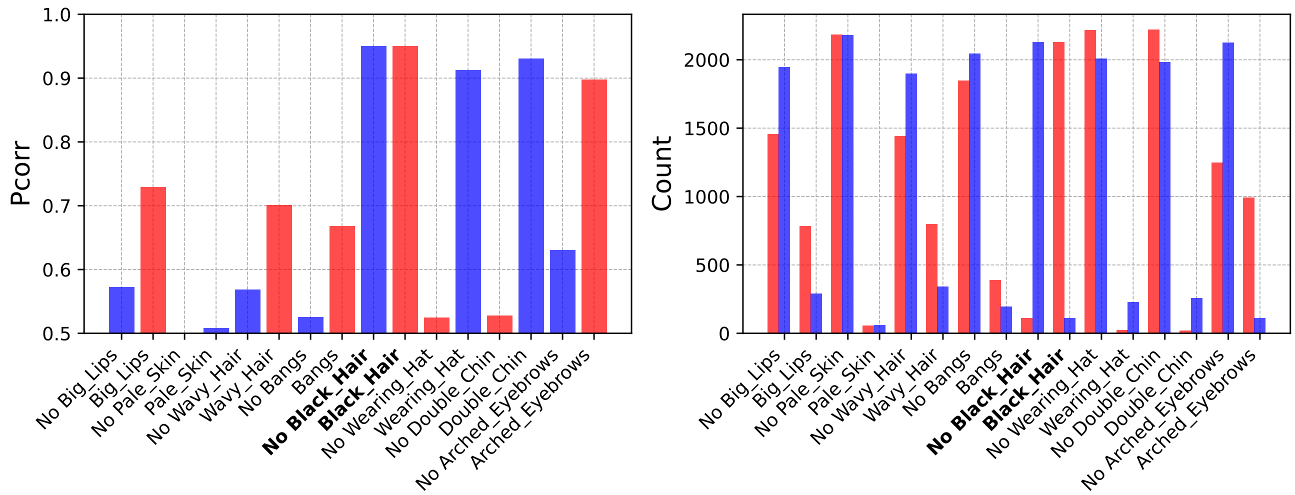


Figure 12. Task 4

Task 5 - Train Split



Task 5 - Test Split

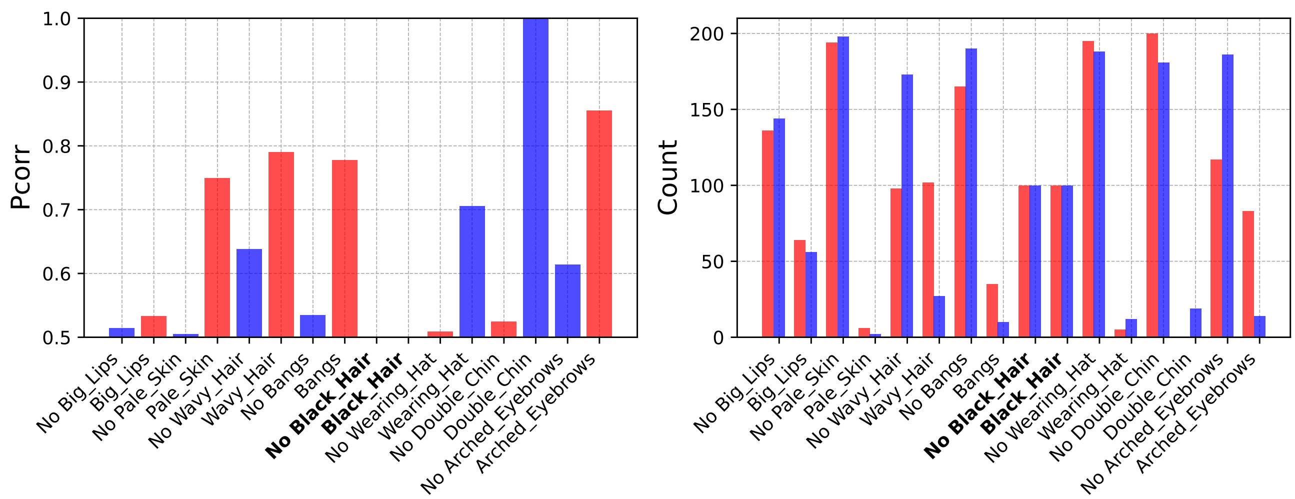
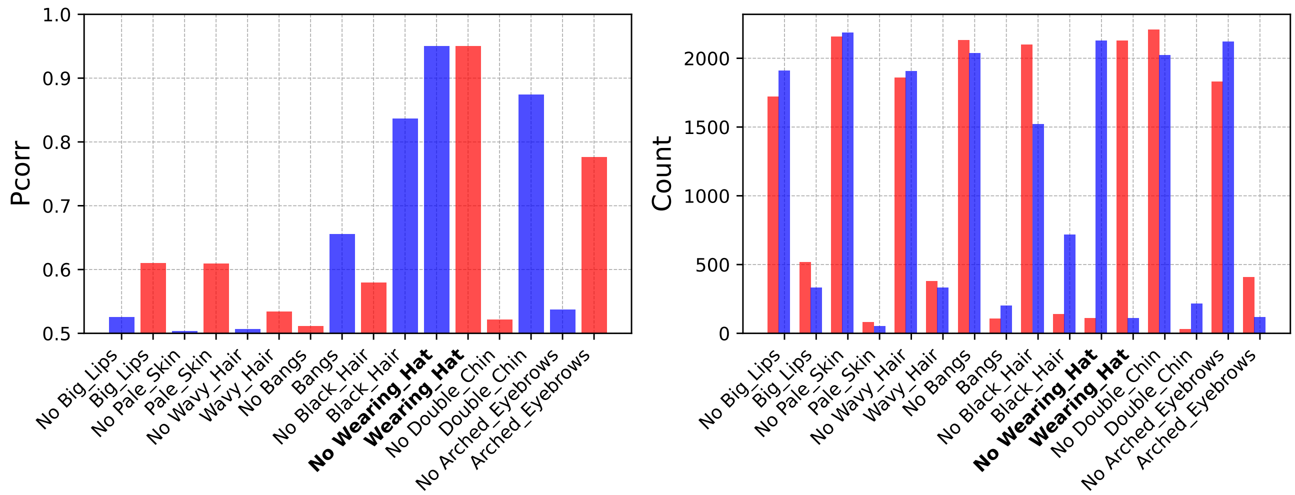


Figure 13. Task 5

Task 6 - Train Split



Task 6 - Test Split

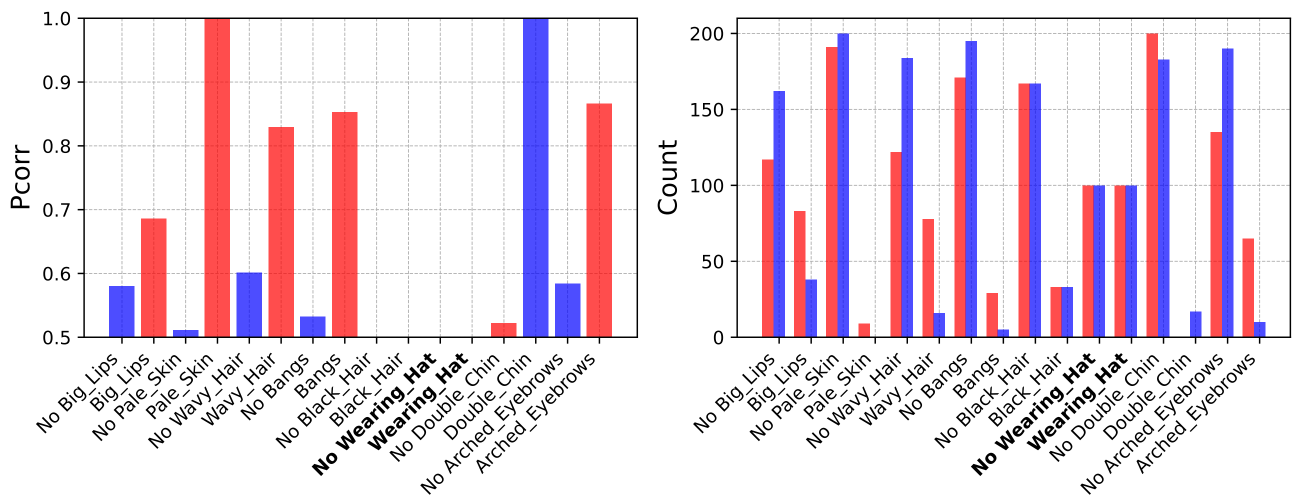
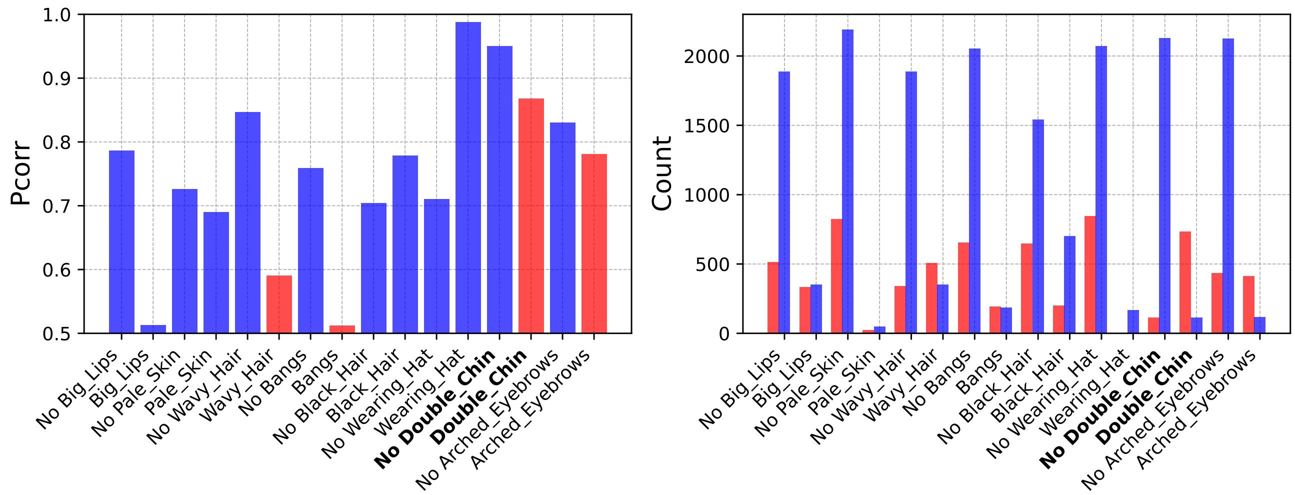


Figure 14. Task 6

Task 7 - Train Split



Task 7 - Test Split

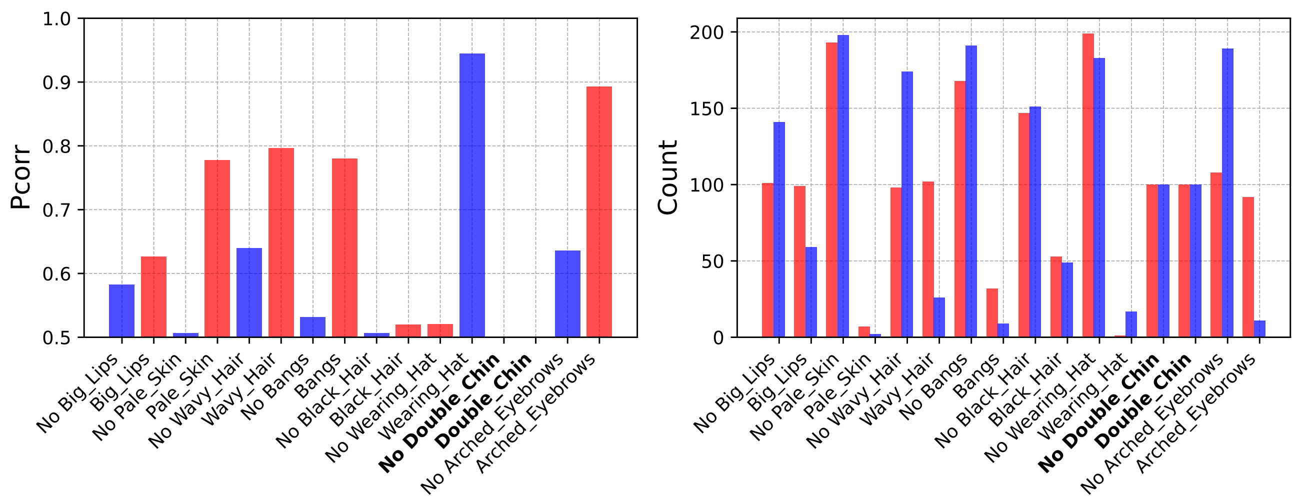
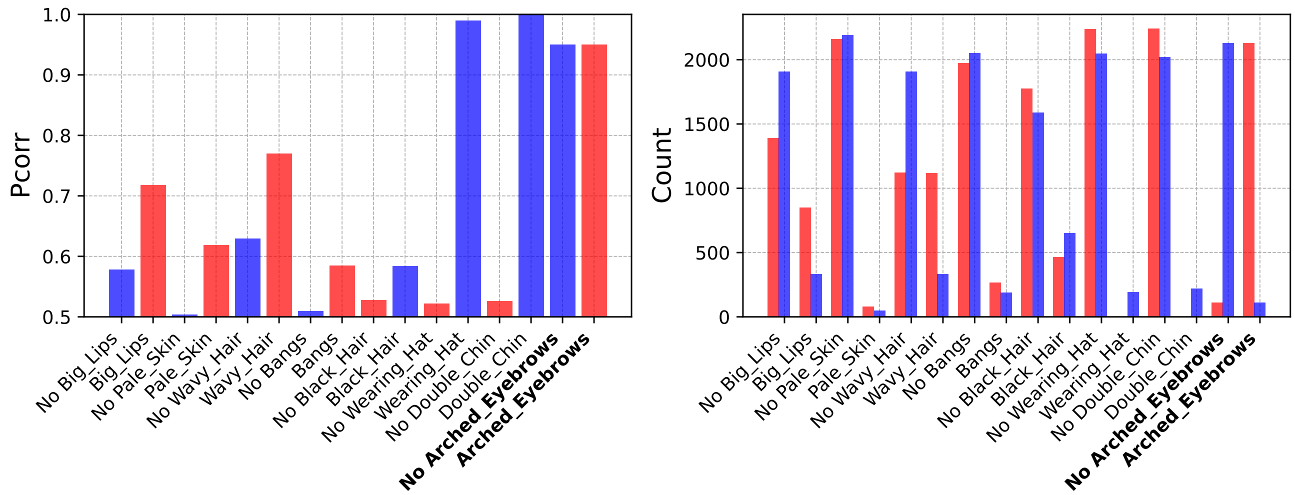


Figure 15. Task 7

Task 8 - Train Split



Task 8 - Test Split

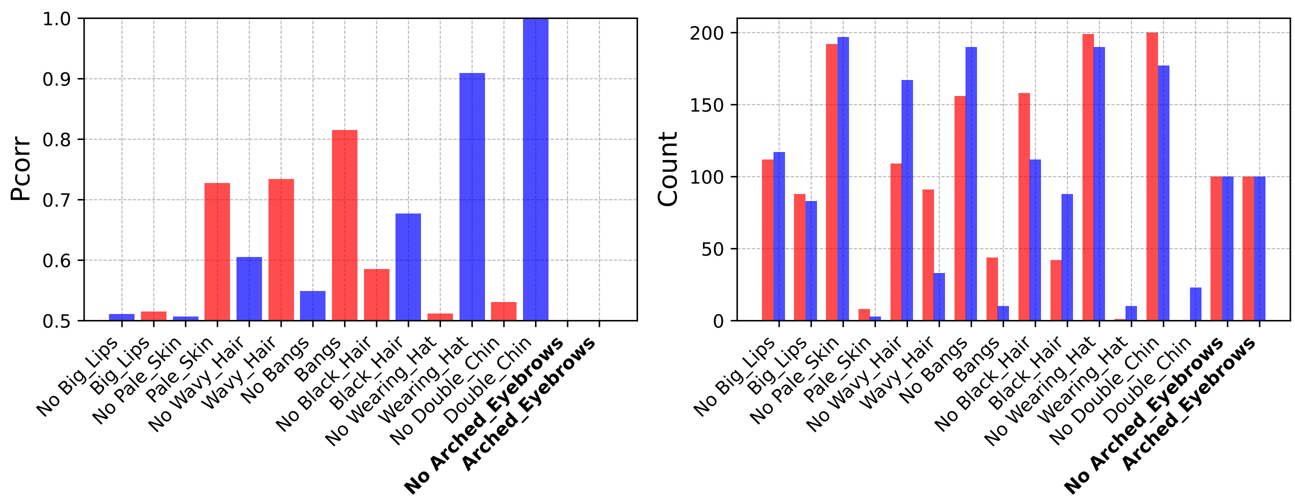
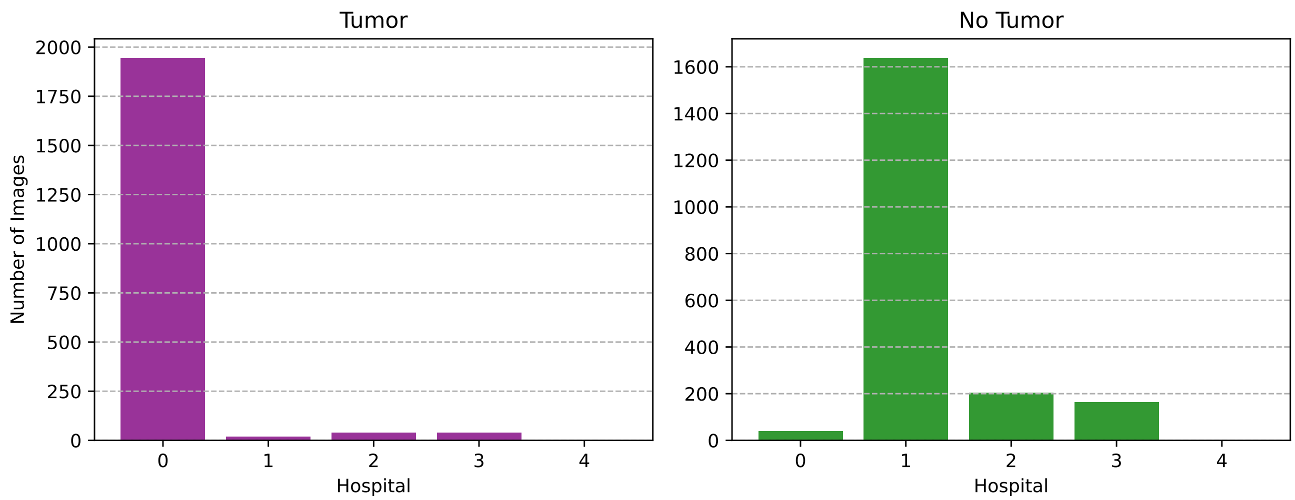


Figure 16. Task 8

***Task*{1,2,3,4} - Train Split**



Test Split

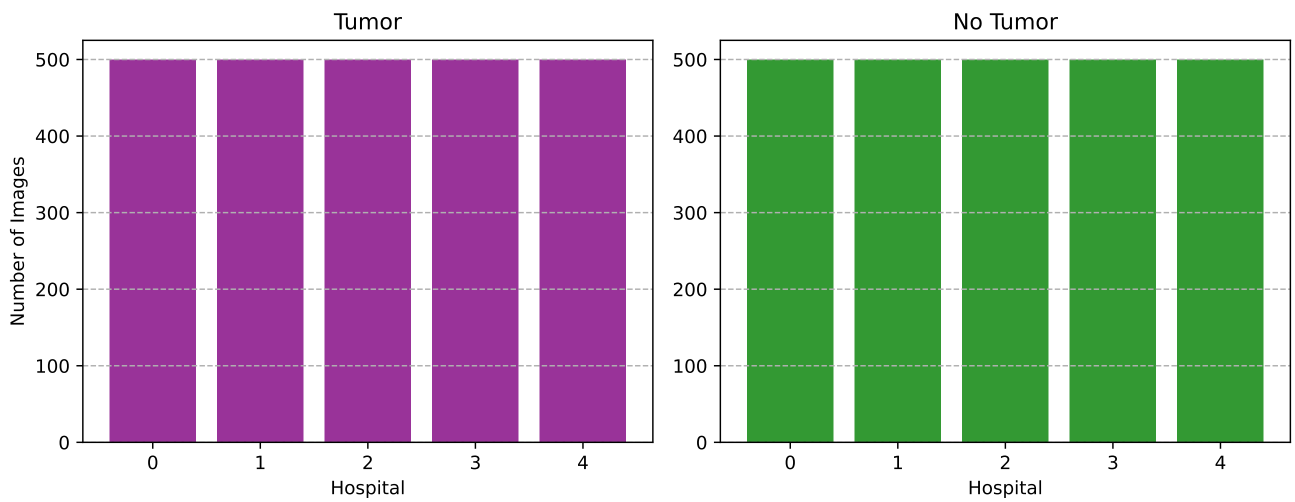


Figure 17. Task 1

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

[2] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[6] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.