DeCLIP: Decoding CLIP representations for deepfake localization
Supplementary material

1. Additional detailed results

In Figure 1 we provide additional detailed results on all train–test scenarios obtained using the PSCC method and DeCLIP on concatenated representations. Specifically, for DeCLIP we stack together the features from the 21st layer of CLIP ViT-L/14 and the features from 3rd layer of CLIP ResNet-50. The representations extracted from ResNet-50 are bilinearly upsampled from $14 \times 14 \times D$ to $16 \times 16 \times D$ to match the spatial resolution of the features extracted by ViT-L/14; here D denotes the feature dimension. The representations from both networks have the same dimension $D = 1024$. By concatenating the features along the last axis, we obtain a block of size $16 \times 16 \times 2048$, which then fed as input to the conv-20 decoder.

Compared to PSCC, DeCLIP shows better generalization capabilities (results in the out-of-domain setups, off-principal diagonal), especially when trained on LDM and P2. PSSC generally has better in-domain performance (principal diagonal), with the exception of the harder LDM–LDM case, where DeCLIP performs better (51.1% compared to 41.5% IoU)

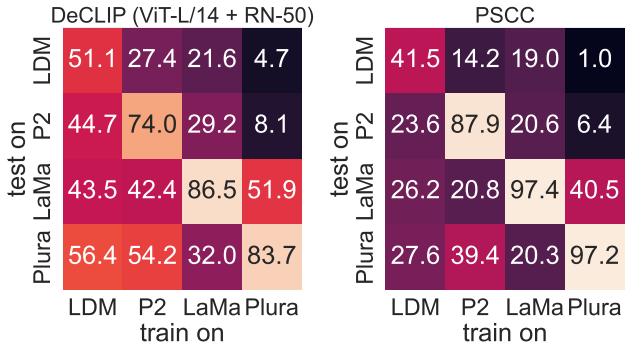

Figure 1. Detailed cross-generator performance (IoU) on the Dolos dataset (all 16 train–test combinations) for DeCLIP that used both ViT-L/14 and ResNet-50 representations and PSCC.

2. Additional qualitative results on Dolos

In Figures 3, 4, 5, 6 we show detailed visual results on Dolos dataset for all train–test scenarios for DeCLIP as well as four other methods trained and tested in the same way: Patch Forensics, CLIP-linear, PSCC and CAT-Net. The results show that although some train–test scenarios are considerably harder than the other, DeCLIP offers a plausible manipulation mask in the majority of cases. We showcase different types of masks, from the very small ones that cover only eyes to larger ones that correspond to face and hair. Patch Forensics and PSCC usually work well in domain (with the exception of LDM–LDM scenario), but genrally struggle in the out-of-domain cases. CLIP-linear and CAT-Net struggle both in domain and out of domain, producing masks with arbitrary activations that follow the face characteristics.

3. Additional qualitative results on COCO-SD

In Figures 7 and 8 we provide additional results on COCO-SD dataset for DeCLIP, Patch Foreniscs, PSCC and CAT-Net. Notice that even in a diverse visual domain, with arbitrary-shaped inpainted regions, DeCLIP has a more stable and precise localization of the manipulated area. The dataset is particularly hard as the inpainted objects are often parts of a larger one (e.g. the tie, the drawing of the dog on a cup), represent a single entity among similar of the same type (the doughnut, the bowl). Even in these conditions, DeCLIP provides plausible maps of the inpainting.

4. Illustration of LDM images

In Figure 2 we show how fingerprint and fake content are distributed in different types of LDM images. Green color corresponds to real content, red color corresponds to fake content and red dots symbolize fingerprint.

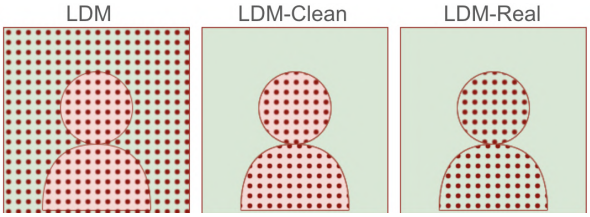

Figure 2. Schematic view of different types of inpaintings with LDM considered in Section 5, Table 4 in the main paper.

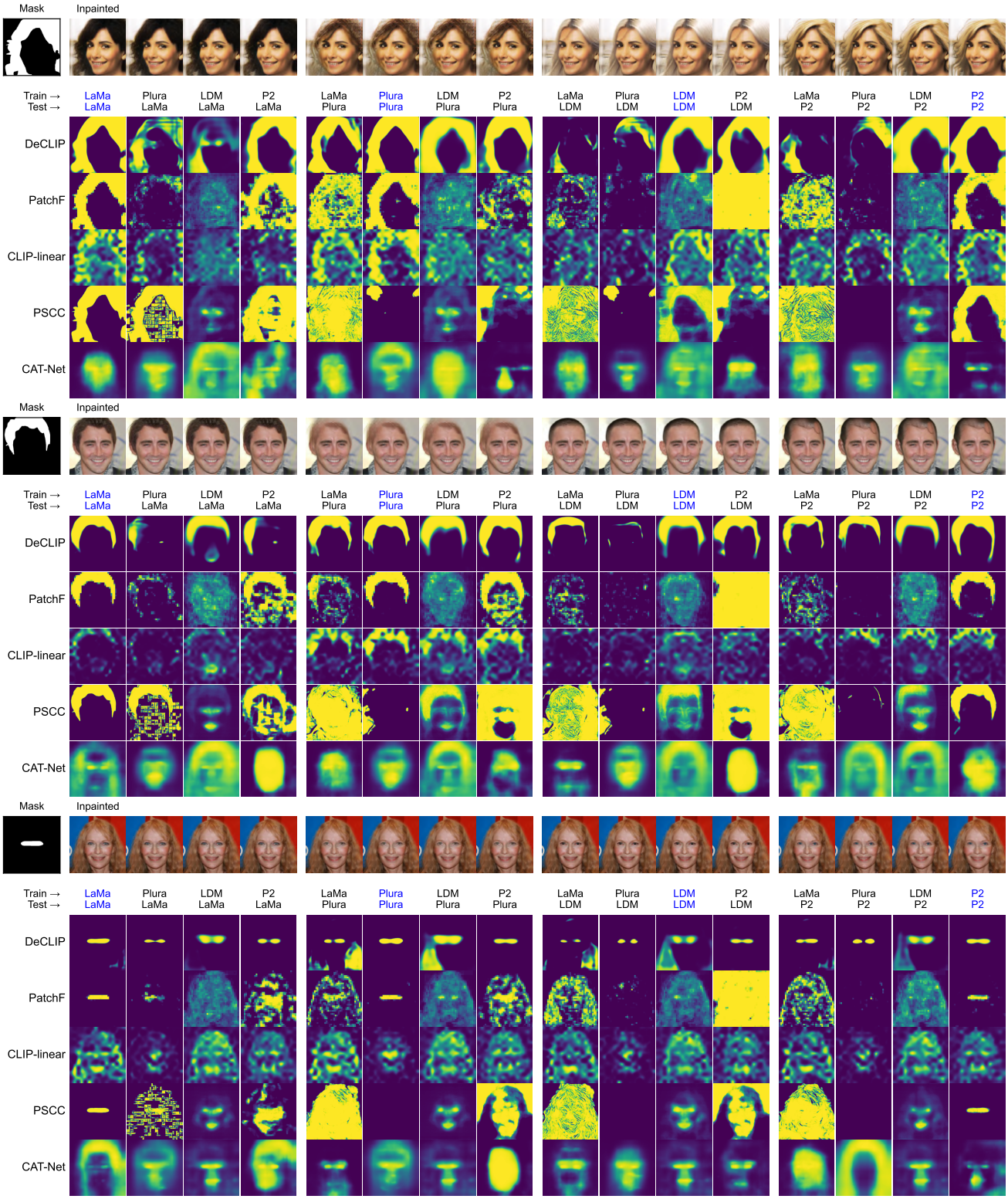

Figure 3. Sample predictions for DeCLIP (second row) and four other methods (Patch Forensics, CLIP-linear, PSCC, CAT-Net) on all 16 train–test combinations from the Dolos dataset. The in-domain combinations are highlighted in blue; the others are out-of-domain combinations. The black-and-white image in the top left corner shows the inpainting mask (white is the inpainted region) and the rest of the images in the first row are the inpainted images with one of the four test datasets (LaMa, Pluralistic, LDM, P2).

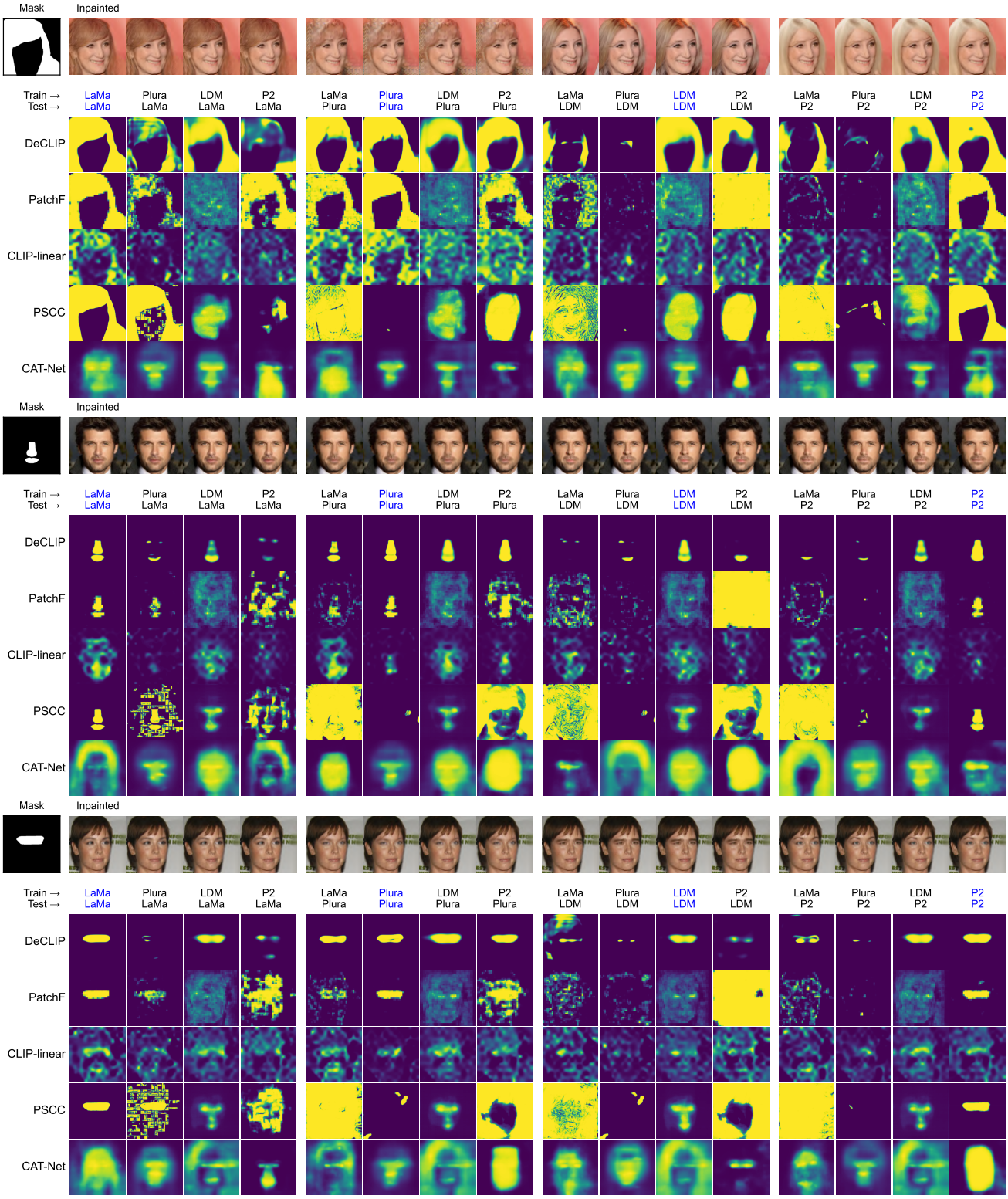

Figure 4. Sample predictions for DeCLIP (second row) and four other methods (Patch Forensics, CLIP-linear, PSCC, CAT-Net) on all 16 train–test combinations from the Dolos dataset. The in-domain combinations are highlighted in blue; the others are out-of-domain combinations. The black-and-white image in the top left corner shows the inpainting mask (white is the inpainted region) and the rest of the images in the first row are the inpainted images with one of the four test datasets (LaMa, Pluralistic, LDM, P2).

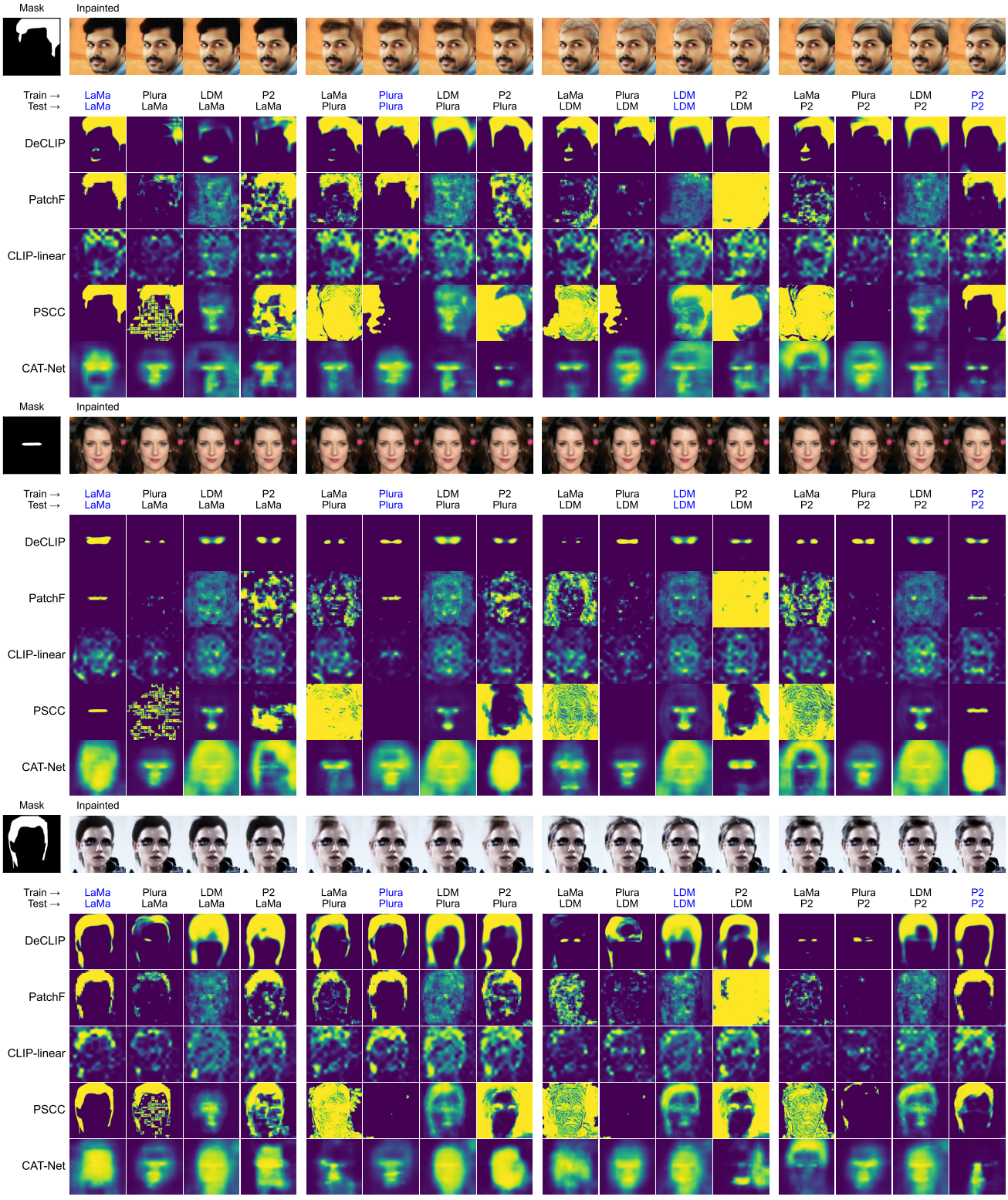

Figure 5. Sample predictions for DeCLIP (second row) and four other methods (Patch Forensics, CLIP-linear, PSCC, CAT-Net) on all 16 train–test combinations from the Dolos dataset. The in-domain combinations are highlighted in blue; the others are out-of-domain combinations. The black-and-white image in the top left corner shows the inpainting mask (white is the inpainted region) and the rest of the images in the first row are the inpainted images with one of the four test datasets (LaMa, Pluralistic, LDM, P2).

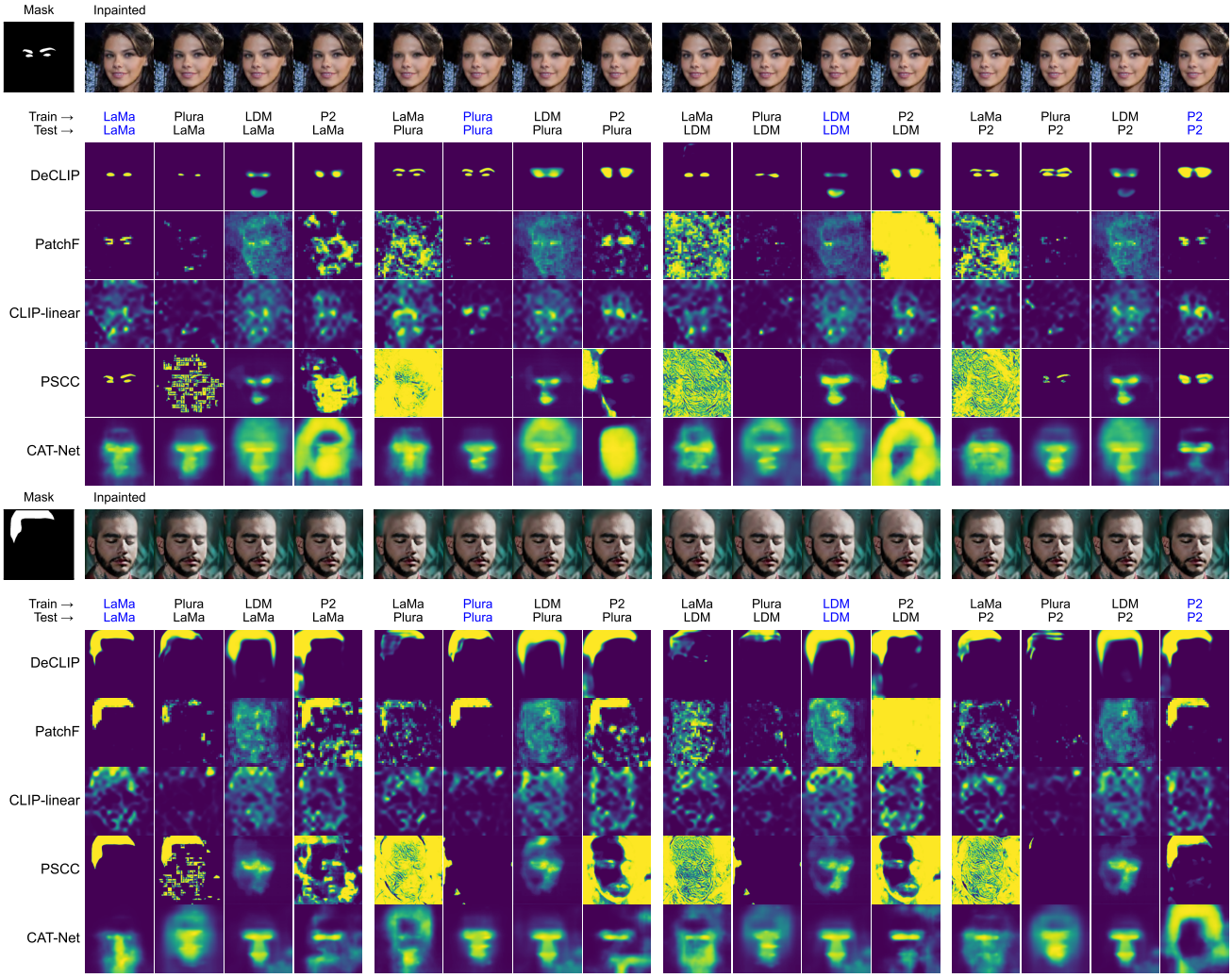

Figure 6. Sample predictions for DeCLIP (second row) and four other methods (Patch Forensics, CLIP-linear, PSCC, CAT-Net) on all 16 train–test combinations from the Dolos dataset. The in-domain combinations are highlighted in blue; the others are out-of-domain combinations. The black-and-white image in the top left corner shows the inpainting mask (white is the inpainted region) and the rest of the images in the first row are the inpainted images with one of the four test datasets (LaMa, Pluralistic, LDM, P2).

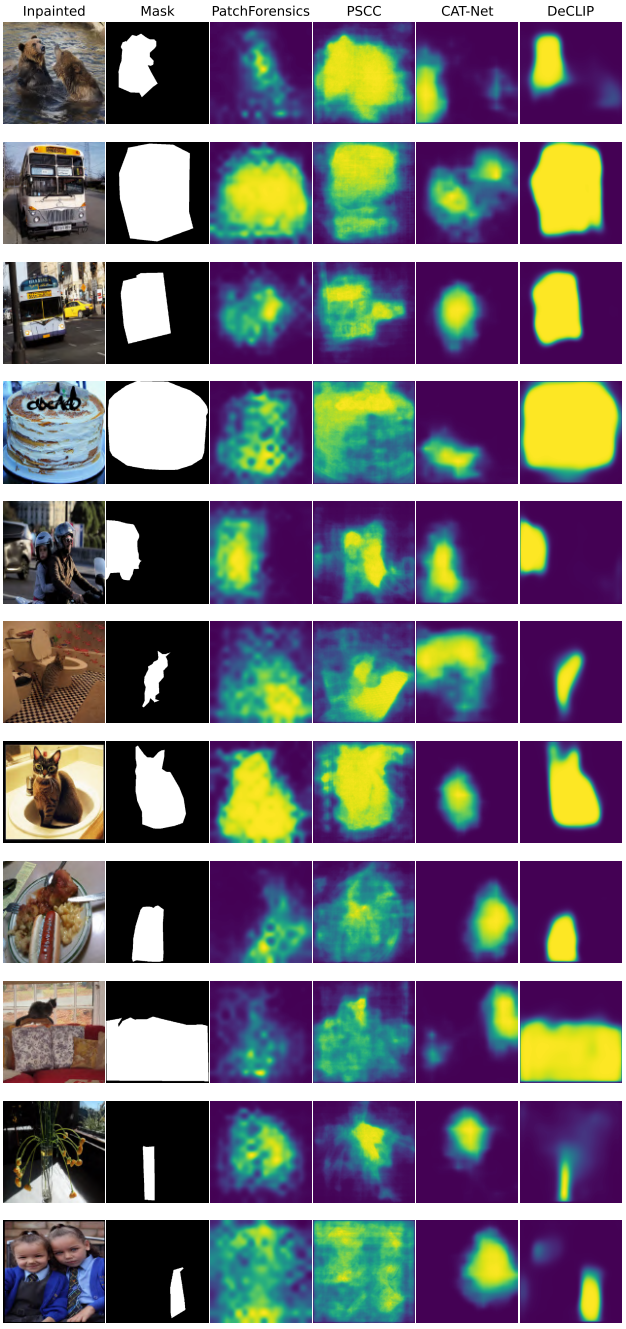

Figure 7. Manipulation localization results on COCO-SD, which has a more challenging set of masks and diverse content. DeCLIP offers a more precise localization of the manipulated area.

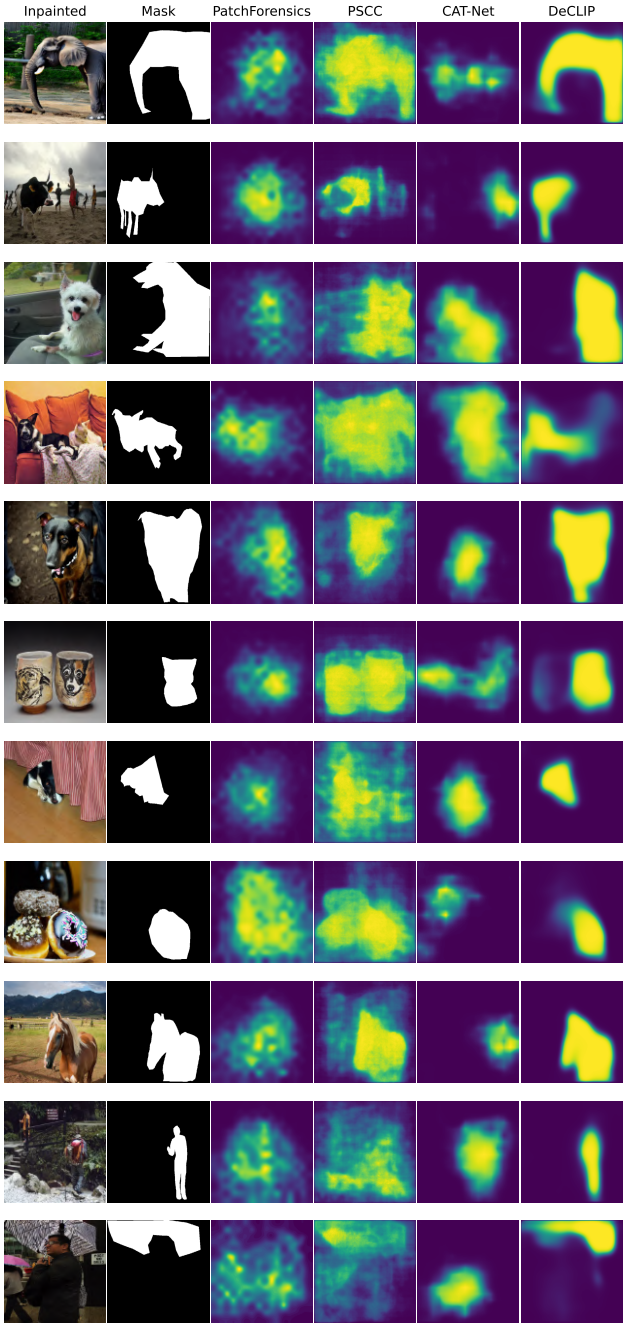

Figure 8. Manipulation localization results on COCO-SD, which has a more challenging set of masks and diverse content. DeCLIP offers a more precise localization of the manipulated area.