Perceptual Artifacts Localization for Image Synthesis Tasks Supplementary Materials

A. Details on Data Labeling

We have discussed about the overall data labeling and statistics in the main paper. Here, we added more details regarding the data labeling.

Labeling Criterion Labeling perceptual artifacts is a highly subjective task, and therefore different workers may have varying opinions on which regions should be considered as 'artifacts'. We instruct the workers to keep a specific criterion in mind while labeling, which is to imagine that we have a perfect artifact-fixing model that can correct any marked region. Hence, if a worker believes that any region in the image can be enhanced or refined, they should mark those regions accordingly.

User Interface We use a labeling interface similar to the one used in [15], where we duplicate the generated image and stitch the copies side-by-side. During labeling, workers can identify perceptual artifacts on the right side of the image and refer to the 'unmarked' image on the left side as a reference.

Visualization of Labels We show more visualization of the perceptual artifacts labels in Fig 2.

B. Implementation Details

Training Details of PAL Models We implement our PAL model using the Swin-L as the backbone, UperNet as the head with a loss weight of 1.0, and a FCN auxiliary head with a loss weight of 0.4. During training, we use random crop with max cutout ratio of 0.75, and random flip with a probability of 0.5. Our code implementation is based on MMSegmentation [1].

Code Resources for Data Generation We use the official github repos to generate the images for each synthesis tasks. The StyleGAN 2 images from domain ffhq, afhq-dog, afhqcat, and afhqwild are generated using the official NVIDIA StyleGAN repo [2]. For StyleGAN 2 human images, we use the StyleHuman repo [3]. For unconditional generation with Latent-Diffusion Model (LDM), we use the code from [4]. We generate Anyres GAN images using [5], and super resolution with Real-ESRGAN [6]. For Edge-to-Image and Mask-to-Image, we use the same diffusion-based model PITI [7] but different checkpoints. For DALL-E 2 text-to-image synthesis and inpainting, we use the OpenAI API [8]. For Stable Diffusion, we use the v1.4 checkpoint from [9] for text-to-image synthesis, and v1.5 checkpoint from [10] for inpainting. We use official latent composition repo [11] for image composition synthesis. Finally, we directly use the synthesized images from repo [12] for virtual try-on task and repo [13] for portrait shadow removal. For other inpainting models used in artifacts fixing pipeline, we use official LaMa github repo [14], and official CoMod-GAN github repo [15].

Prompt for Text-based Inpainting Text-based diffusion inpainting requires additional text prompt besides the image and mask inputs. In this section, we discuss how we decide the fixed text prompts for each type of generated images. Generally, we use "a person's face" as the text prompt for all facial images generated by StyleGAN2 [5] and LDM [8], and use "a person" for all human images in StyleGAN2 human and virtual try-on task. For LDM LSUN bedroom images, we just use "bedroom" as the text prompt. For the rest of in-the-wild images, we use "photograph of a beautiful empty scene, highest quality settings" as the fixed text prompt, which is the default option used in Stable Diffusion inpainting.

Selecting Multimodal Outputs As text-based inpainting models, i.e. DALL-E 2 [7], have multimodal outputs, we select the final output image based on the Perceptual Artifacts Ratio (PAR), which has some correlation with hu-

[1] mmsegmentation: https://github.com/open-mmlab/mmsegmentation
[2] stylegan: https://github.com/NVlabs/stylegan3
[3] StyleGAN-Human: https://github.com/stylegan-human/StyleGAN-Human
[4] latent-diffusion: https://github.com/CompVis/latent-diffusion
[5] anyres-gan: https://github.com/chail/anyres-gan
[6] Real-ESRGAN: https://github.com/xinntao/Real-ESRGAN
[7] PITI: https://github.com/PITI-Synthesis/PITI
[8] dalle-api: https://openai.com/api/
[9] stable-diffusion-v1.4: https://github.com/CompVis/stable-diffusion
[10] stable-diffusion-v1.5: https://github.com/runwayml/stable-diffusion
[11] latent-composition: https://github.com/chail/latent-composition
[12] c-vton: https://github.com/benquick123/C-VTON
[13] portrait-shadow-manipulation: https://github.com/google/portrait-shadow-manipulation
[14] lama: https://github.com/saic-mdal/lama
[15] co-mod-gan: https://github.com/zsyzzsoft/co-mod-gan

man judgement as described in section 5 of the main paper. Specifically, suppose we have N multimodal outputs, we denote the candidate images as I_i, where $i = 1, ..., N$. We compute the PAR scores for each image, which is denoted as $PAR(I_i)$. The finally selected output image is determined by $\text{argmin}_i\, PAR(I_i)$.

C. Statistical Analysis of User Study

We conduct user studies to evaluate whether the artifacts fixed images are better, same, or worse the original generated images. We perform statistical hypothesis testing using a null hypothesis that the mean of preferences is zero, where the preference is -1 if the original image was preferred, 0 if no preference, and +1 if the artifacts-fixed image was preferred. We use a one sample permutation t test with 10^6 permutations. If we combine all user votes into a single list, the null hypothesis is rejected with $p = 0$. If we run a test per task, using Holm-Bonferroni correction and a familywise error rate of 0.05, we find the null hypothesis is rejected for every task except super-res, text-to-image, and shadow removal. This indicates that for 6 out 10 tasks and for the combination of all user votes across tasks, there is a significant preference, which per our data is the artifacts fixed image.

D. More Qualitative Results

In this section, we show more visualization results.

D.1. PAL and Artifacts Fixed Results

We show more qualitative results of perceptual artifacts segmentation and artifacts fixed results for ten synthesis tasks. These visual results are shown in Fig 3 - 12. In each example, first image is the generated image with perceptual artifacts localization (PAL), which is indicated by the pink mask. The second image is the original generated image, and the third is the corresponding artifacts fixed image using the predicted PAL. We put the original and artifacts refined images side-by-side for more direct visual comparison.

D.2. The Choices of Inpainting Models

In this paper, we mainly use CoMod-GAN [17], LaMa [10], and DALL-E 2 inpainting [7] in our artifacts fixing pipeline, as discussed in section 4 in the paper. In this section, we show ablation studies on how different inpainting models can be used to fix the perceptual artifacts in different cases. As shown in Figure 14, for face inpainting, CoMod-GAN trained on FFHQ [4] face dataset produce more realistic results than the CM-GAN [18], and has similar performance to DALL-E 2 inpainting. Since CoMod-GAN has faster inference speed than DALL-E 2 by a order of magnitude, we choose CoMod-GAN for general face inpainting cases. For other in-the-wild inpainting cases, as shown in Figure 13, we observe that GAN-based models LaMa and CM-GAN have reasonably good performance on the relatively easy cases, such as the first two rows. However, when the images are under perspective (3^{rd} row) or involve object completion (4^{th} and 5^{th} rows), diffusion-based models generally produce much better results. Within diffusion-based models, DALL-E 2 produce much more realistic details than Stable Diffusion inpainting [8] with v1.5 checkpoint. Therefore, we use LaMa for the easy background inpainting in tasks like Anyres GAN [1], and DALL-E 2 for the rest of tasks with complex scene or object completion.

D.3. Zoom-in Effect on Inpainting

In the main paper, we discuss that diffusion-based models, i.e. DALL-E 2 [7], systematically struggles to generate high-fidelity object details, such as faces and hands. Here, we show more qualitative results. Inspired by this insight, we further propose a 'zoom-in' inpainting pipeline that can fix the perceptual artifacts in the object detail level. As show in Figure 16, we can see that this zoom-in inpainting pipeline can significantly refine the object details and outperform naively inpainting using the full images and masks. More detailed comparisons on hands and faces are illustrated in Figure 15. In this work, we use the fixed text prompt for all the patches, but more tailored text prompt for the individual cropped patch should theoretically improve the visual quality, which we leave as future work.

E. SDEdit for Perceptual Artifacts Fixing

Using inpainting methods to fix the perceptual artifacts might not be ideal for certain synthesis tasks, since it could change too much of the original generated image identity. We also explore an alternative approach SDEdit [6], which enables stroke-based editing using a diffusion model generative prior DDIM [9]. In the implementation, we convert the pixels in the perceptual artifacts region into stroke painting by RTV smooth algorithm [13], and then run SDEdit to re-generate pixels in the artifacts region. As shown in Figure 1, SDEdit preserves more image identity with respect to the original generation, but underperforms DALL-E 2 inpainting [7] in terms of realism. SDEdit also has a hyperparameter that controls the tradeoff between realism and faithfulness (identity preservation), and this can be adjusted for different tasks. In this work, we showcase the usage of SDEdit with DDIM trained on LSUN Church dataset. To apply this in the wild, we might either need to re-train DDIM in larger diverse dataset or integrate SDEdit with other diffusion-based models, i.e. Stable Diffusion [8], and we leave this as future work.

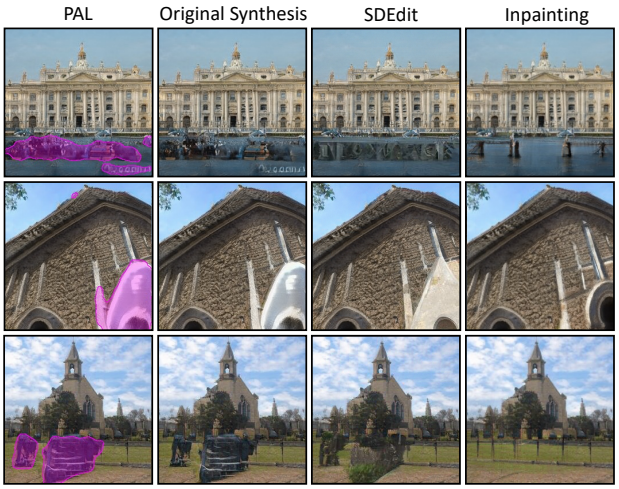

Figure 1. Qualitative comparison between SDEdit [6] and DALL-E 2 inpainting [7] for artifacts fixing. In general, we can see that SDEdit preserves more image identity (more similar to the original synthesis), while DALL-E 2 inpainting produces better realism.

References

[1] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis. *arXiv preprint arXiv:2204.07156*, 2022. 2, 7

[2] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in gans. *arXiv preprint arXiv:2103.10426*, 2021. 8

[3] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-vton: Context-driven image-based virtual try-on network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3144–3153, 2022. 7

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 9

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 5

[6] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 3

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 3, 7, 9, 10

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 5, 9

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[10] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2, 9

[11] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 6

[12] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 6

[13] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 2

[14] Lingzhi Zhang, Connelly Barnes, Kevin Wampler, Sohrab Amirghodsi, Eli Shechtman, Zhe Lin, and Jianbo Shi. Inpainting at modern camera resolution by guided patchmatch with auto-curation. In *European Conference on Computer Vision*, pages 51–67. Springer, 2022. 9

[15] Lingzhi Zhang, Yuqian Zhou, Connelly Barnes, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for inpainting. *arXiv preprint arXiv:2208.03357*, 2022. 1

[16] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020. 8

[17] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 9

[18] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Image inpainting with cascaded modulation gan and object-aware training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 277–296. Springer Nature Switzerland Cham, 2022. 2, 9

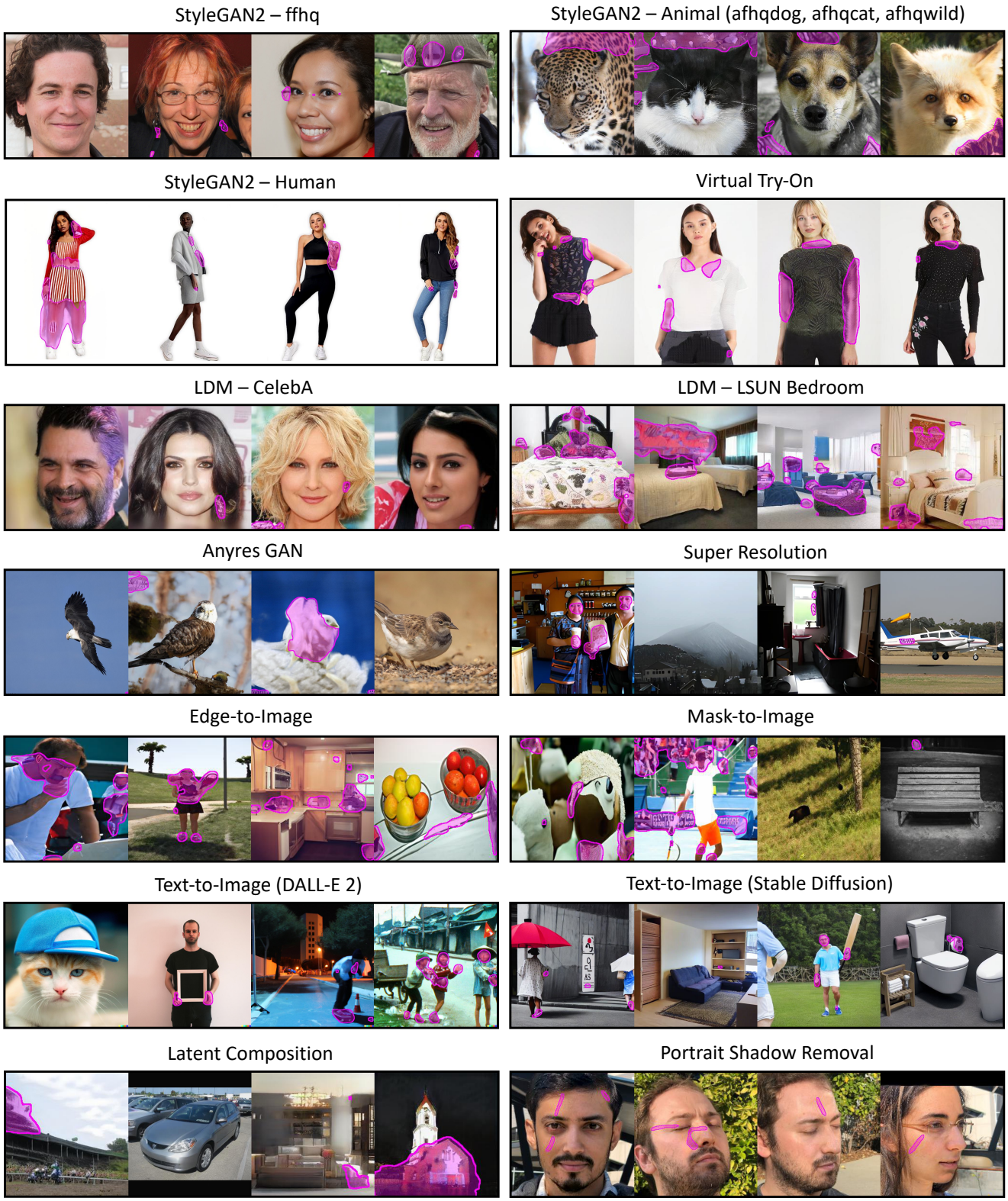

Figure 2. A sampled visualization of our labeled perceptual artifacts dataset in diverse synthesis tasks and domains. Note that if there is no mask in the image, it indicates that workers do not think there are any artifacts in the generated image.

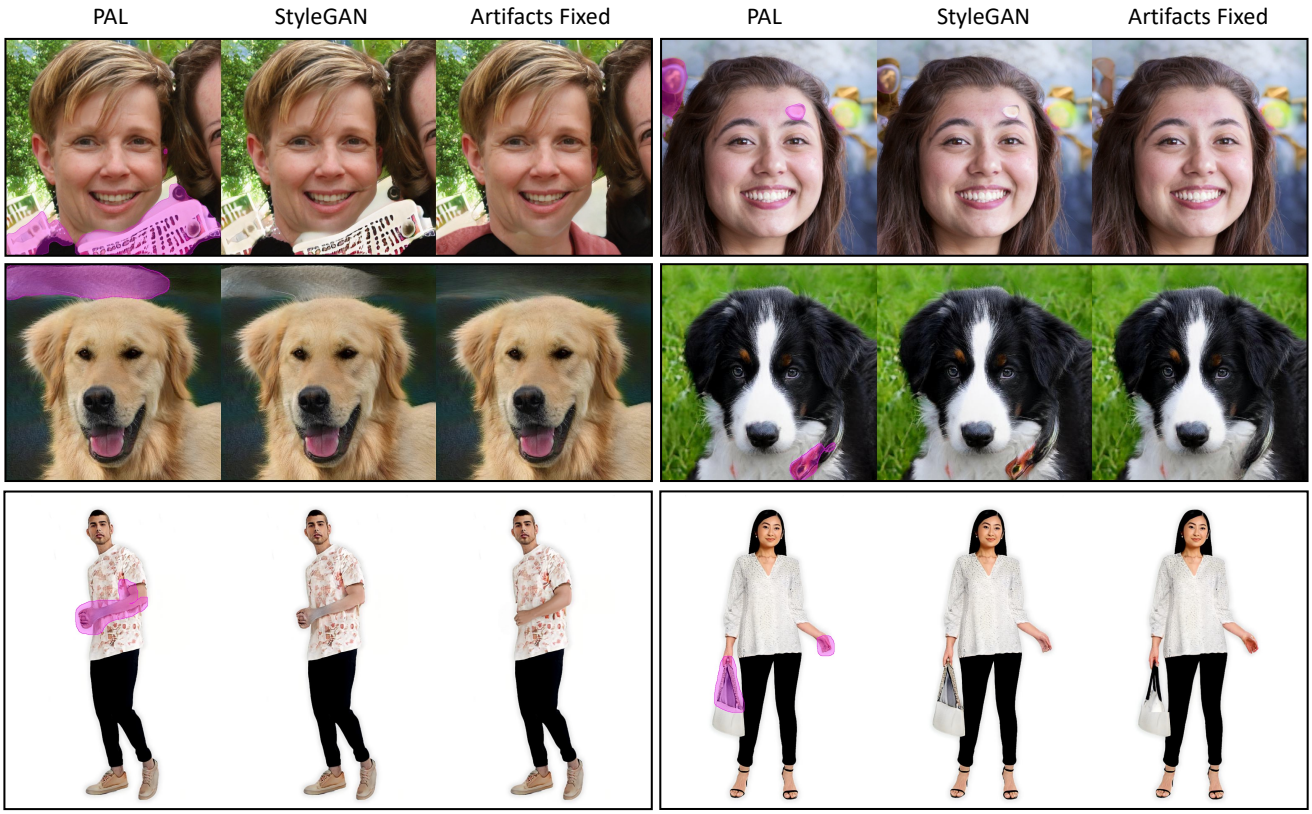

Figure 3. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for StyleGAN [5]. **Left**: original generated image with PAL prediction. **middle**: original generated image. **right**: artifacts fixed/refined generated image.

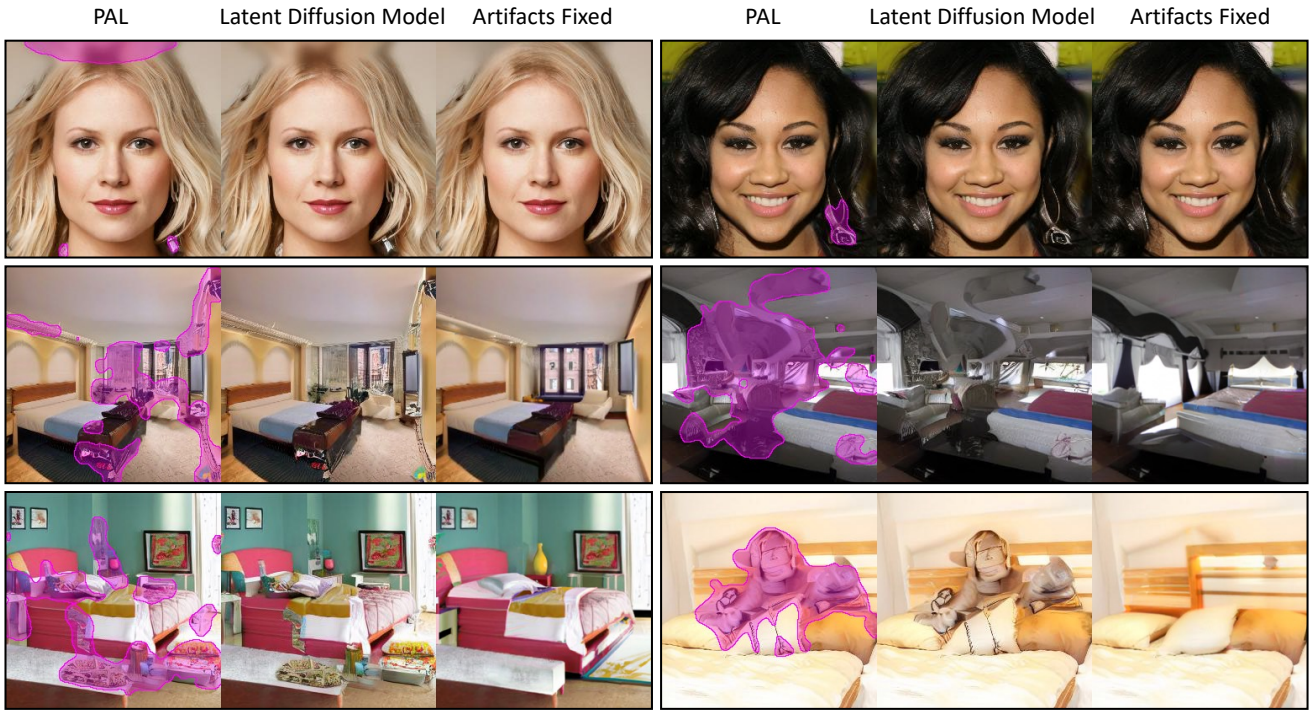

Figure 4. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Latent Diffusion Model [8]. **Left**: original generated image with PAL prediction. **middle**: original generated image. **right**: artifacts fixed/refined generated image.

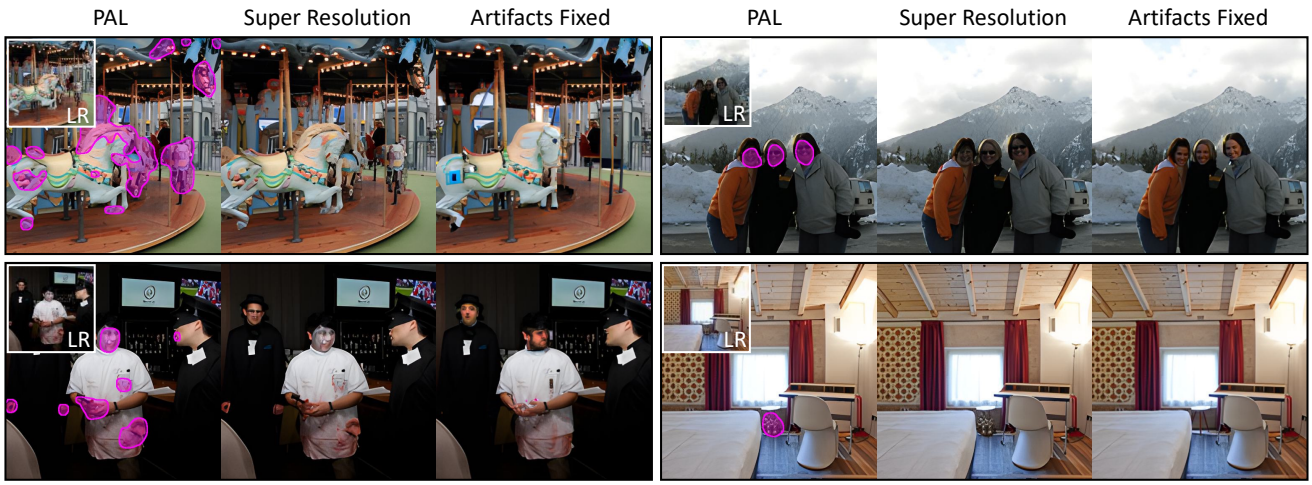

Figure 5. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for super resolution with Real-ESRGAN [12].

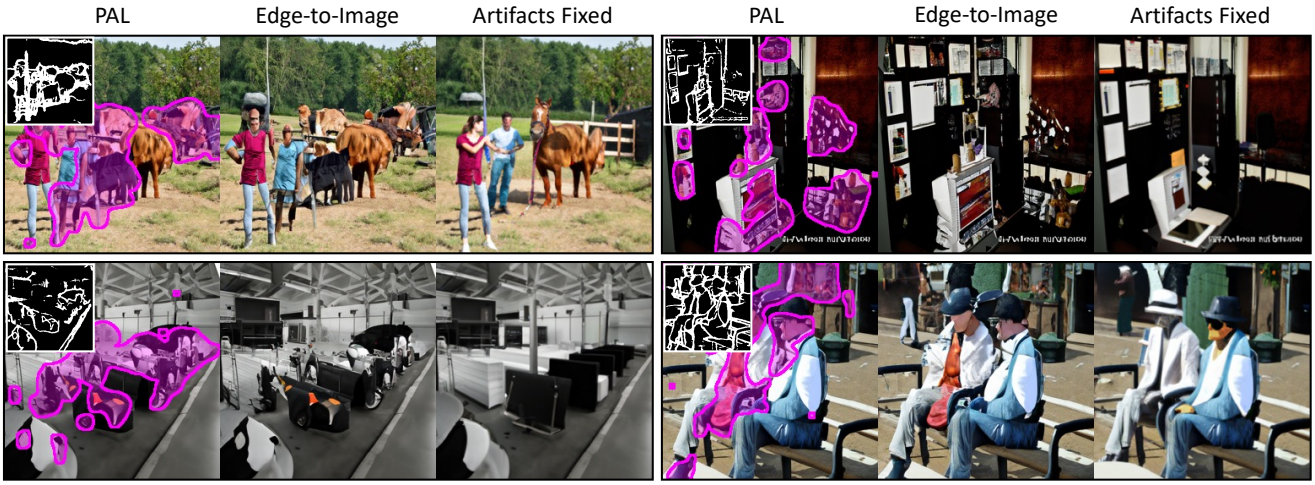

Figure 6. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Edge-to-Image translation with PITI [11].

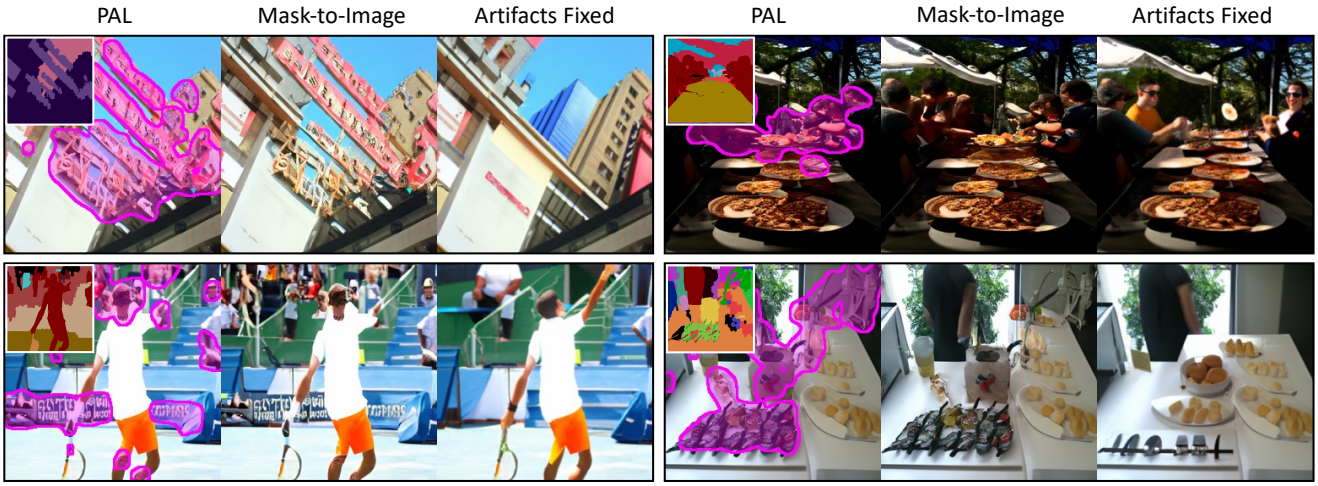

Figure 7. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Mask-to-Image translation with PITI [11].

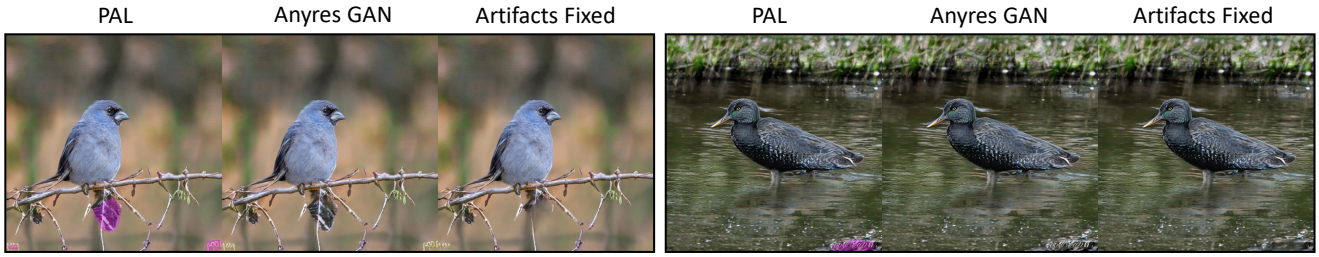

Figure 8. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Anyres GAN [1].

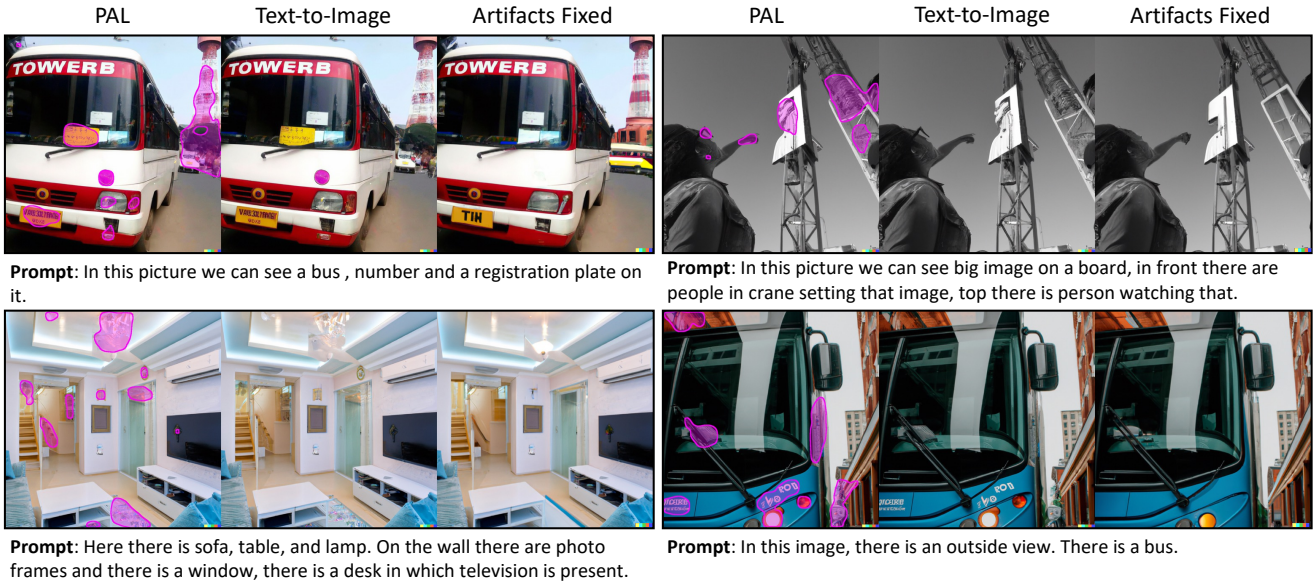

Figure 9. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Text-to-Image synthesis with DALL-E 2 [7].

Figure 10. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for virtual try-on with [3].

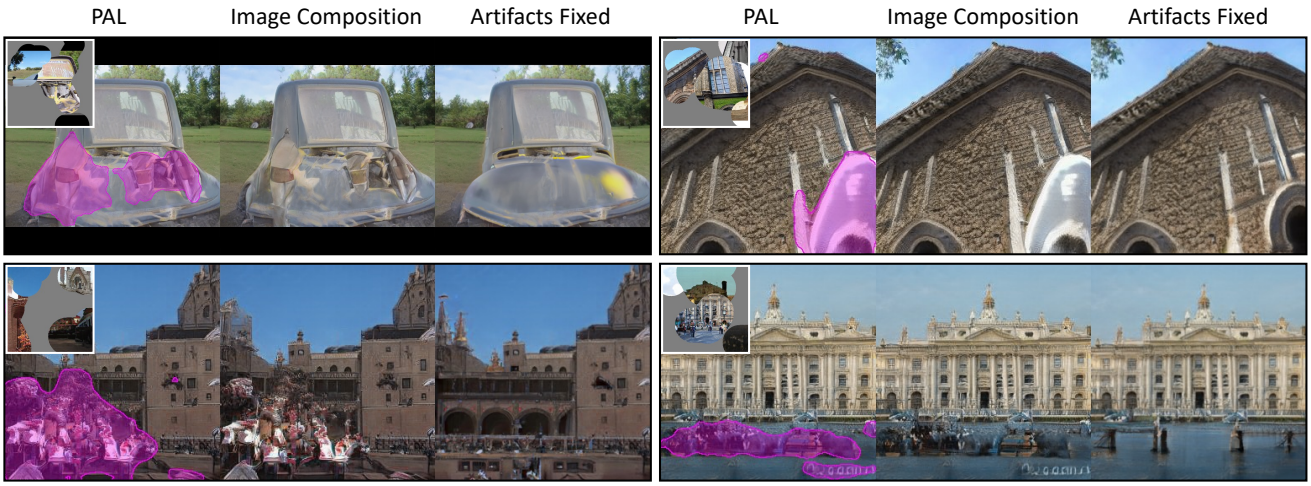

Figure 11. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for latent composition [2].

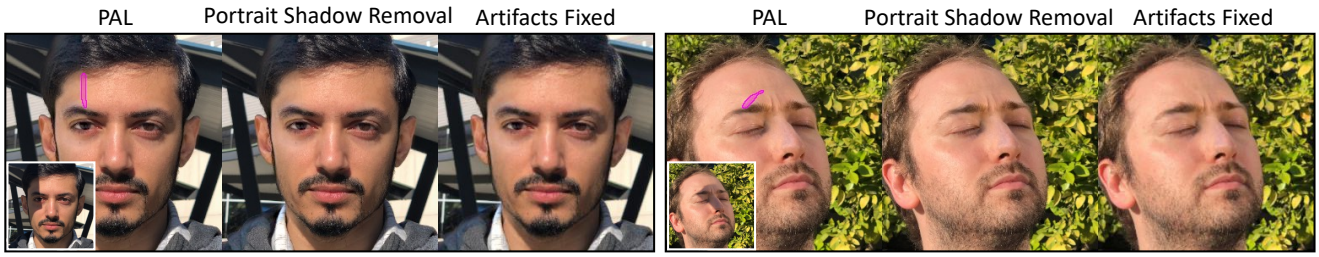

Figure 12. More qualitative results for perceptual artifacts localization (PAL) prediction and the artifacts fixed images for Portrait Shadow Removal [16]. Please *zoom in* to see the detailed comparisons.

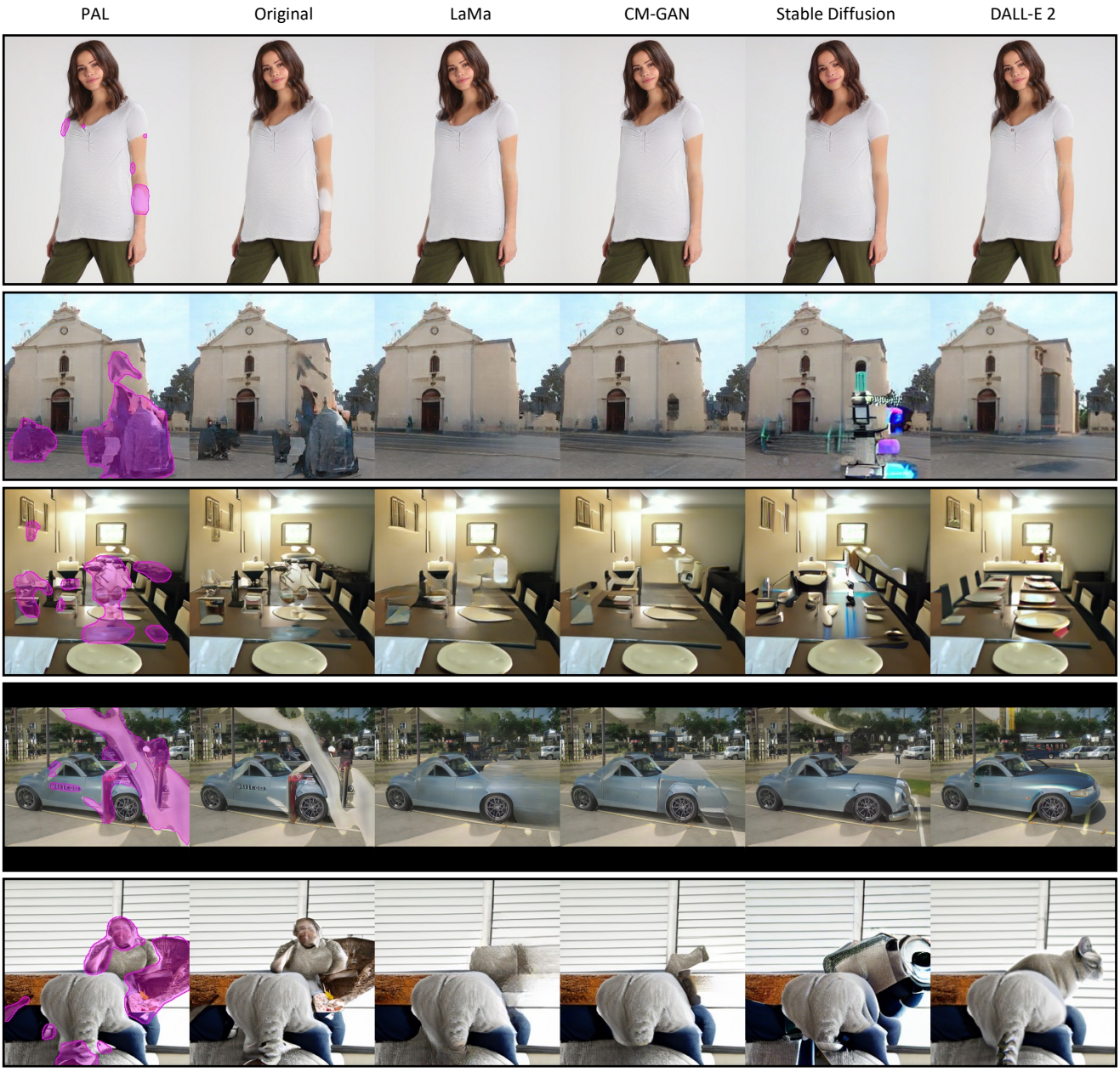

Figure 13. An ablation study on how four different state-of-the-arts inpainting models, including LaMa [10], CM-GAM [14], Stable Duffion [8], and DALL-E 2 [7], could fix the perceptual artifacts in types of generated images using our PAL prediction as the inpainting masks.

Figure 14. An ablation study on how inpainting models work on face artifacts removal. Note that CM-GAN [18] and DALL-E 2 [7] are not tailored for face inpainting, while CoMod-GAN [17] is trained on the FFHQ [4] dataset for face inpainting specifically.

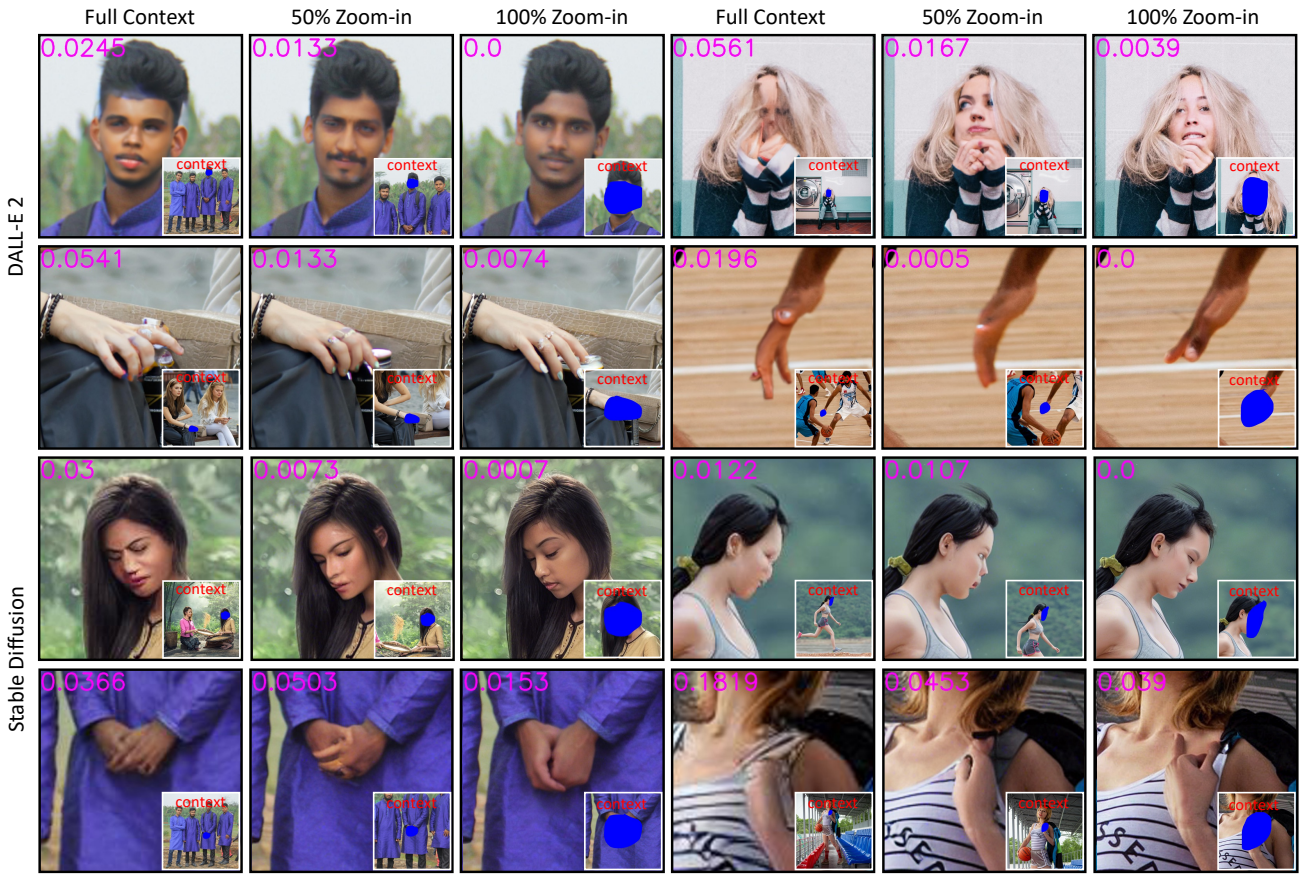

Figure 15. More qualitative results showing that DALL-E 2 inpainting [7] and Stable Diffusion [7] tend to generate less perceptual artifacts when zooming in around the object region, such as faces and hands. We show that our PAR scores, which are placed at the top left corner of the images, can be used to quantify this observation and confirm our insight.

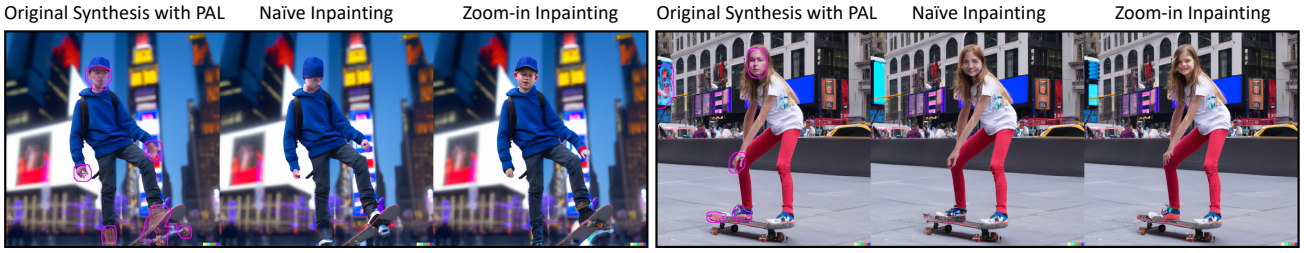

Figure 16. Qualitative comparison between naive inpainting and zoom-in inpainting for fixing perceptual artifacts in text-to-image outputs. In the above examples, we use DALL-E 2 [7] for both text-to-image generation and inpainting. Naive inpainting could fix certain artifacts compared to the original synthesis, but still struggles to generate high-fidelity object details. In contrast, zoom-in inpainting pipeline produces much more realistic object details.