Supplementary Material for *SwapAnything*: Enabling Arbitrary Object Swapping in Personalized Image Editing

Jing Gu[1]⋆ Nanxuan Zhao[2] Wei Xiong[2] Qing Liu[2] Zhifei Zhang[2] He Zhang[2]
Jianming Zhang[2] HyunJoon Jung[2] Yilin Wang[2]⋆⋆ Xin Eric Wang[1]⋆⋆

[1]University of California, Santa Cruz [2]Adobe
https://swap-anything.github.io/

A Variable Swapping Details

We use Stable Diffusion 2.1 as the pre-trained text-to-image diffusion model. DreamBooth [8] is used to convert the concept into textual space. The learning rate for this process is set at 1e-6, and we use the Adawm optimizer for 800 steps. The U-net and the text encoder are fine-tuned during this process, typically taking about 2 minutes on a machine equipped with 8 A100 GPUs. The target prompt is essentially the source prompt with a swap in object tokens to introduce a new concept.

For object mask, we first detect the object with Grounding DINO [7] and then extract the mask using Segment Anything [4]. For the targeting variable swapping, we do 30 for latent image feature z, 20 steps for cross-attention map, 25 for the self-attention map, and 10 for the self-attention output, we conduct swapping in all U-Net layer.

For area mask smooth, we first enlarge the masked areas using a dilation operation with an elliptical kernel, which can be adjusted in size. After dilation, the mask edges are smoothed using a Gaussian blur, creating a gradient effect at the boundaries. For the smooth over diffusion step, we linearly increase the mask rate from 0 to 1 during the first 30 steps. For a better understanding, we mark the masked area using a circle in most figures in this paper.

B Adaptation Details

Style Adaptation. This operation adjusts the mean and variance of content image features to match those of the style features, facilitating the transfer of artistic styles onto content images. The AdaIN technique is renowned for its efficiency and flexibility, making it a go-to choice for real-time style transfer and artistic image manipulation. Building on this, we introduce Masked-AdaIN. Unlike traditional AdaIN which applies style alignment across the entire image, Masked-AdaIN focuses this alignment only on a specific target area. In

⋆ This work was partly performed when the first author interned at Adobe.

⋆⋆ Equal advising.

this approach, mean and variance calculations are exclusively performed on the designated masked area, allowing for more precise and localized style transfers.
Scale Adaptation. We adapt the scale of the object in latent space to the shape of the mask. The object shape is indicated in the cross-attention map at each diffusion step [2, 3]. $Shape(\mathbf{M_{src}})(k)$ means the attention map for object text token k, which is obtained through binary-like transformation to the attention map. We apply a threshold of 0.4 after using sigmoid to normalize the attention value between 0 and 1.
Content Adaptation. In the Linear Boundary Interpolation process, the structuring element K is a predefined shape used in the dilation process to create the dilated image. The structuring element K slides over the binary mask $\mathbf{M_{src}}$ and at each position. If at least one pixel under K is 1, the pixel in the output image under the center of K is set to 1. This operation typically results in the enlargement of the regions with 1s in the binary mask, effectively smoothing the boundary and filling small holes and gaps. The subsequent convolution with a Gaussian kernel G further smooths the mask by averaging values in the vicinity of each point, thereby creating a gradient effect. The combination of dilation and Gaussian smoothing prepares the mask S' for linear boundary interpolation, where the sharp transitions are made gradual, and the final soft mask S' is obtained by selectively setting pixels to 1 based on the original mask and the smoothed values. In Gradual Boundary Transition, we set the transition step parameter as 30 to anneal $\mathbf{M_{src}}$ from 0 to the set value.

C Dataset

We conducted experiments on both human and non-human objects. For human swapping, we collect 50 faces as concepts. We also collected 500 images containing 1 or more people as the source images. For non-human object, we include DreamEdit [6] dataset and more concepts and its corresponding source images. In total, we aggregated 1,000 images.

D Object Insertion

SwapAnything is a general framework and is also capable of object insertion. With the same process as single-object swapping, we could insert and adapt a concept into background pixels, while preserving the composition and style of the source image. In Fig. 1, we insert a puppy and a butterfly into The Starry Night from Vincent van Gogh.

E More Qualitative Results

Here we first show the comparison with baselines in their original setting on single-object swapping. On other more challenging tasks, we also show the results of Photoswap since it is the state-of-the-art method of subject swapping.

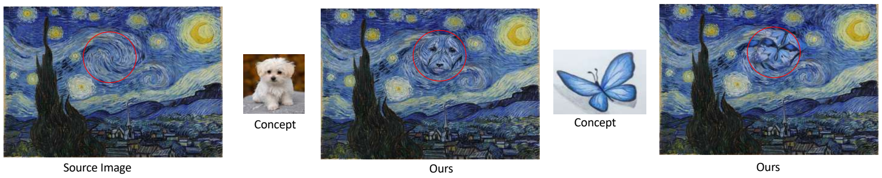


Fig. 1: Results on object insertion. *SwapAnything* can insert and adapt an object into a certain location of an image.

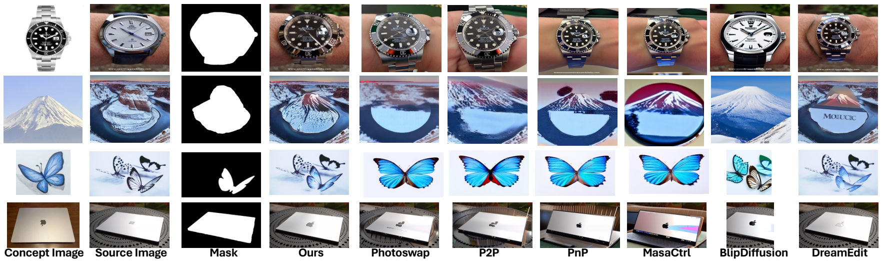


Fig. 2: Comparison on single-object swapping with baselines in their original components. Please zoom in for a clear visual result.

For the implementation details, we use external mask to help the inpainting process in DreamEdit. *SwapAnything* is also compared with BlipDiffusion [5]. Photoswap, P2P, PnP, and MasaCtrl, DreamEdit were equipped with the same DreamBooth model to grasp the new concept. Note that this would also indirectly include comparison with CustomEdit [1], since it also achieved personalized object swapping via equiping P2P with concept learning. CopyPaste involves directly transplanting the concept object in the concept image into the source object's position.

Single-object Swapping. Fig. 2 shows comparisons between *SwapAnything* and baselines. *SwapAnything* consistently outperforms other models in terms of background preservation, identity swapping, and overall quality. Note that there is also a huge performance gap between some baselines and their counterpart in Fig. 4 in the main paper, which further validates the efficacy of targeted variable swapping and location adaptation, which was applied to Photoswap, P2P, PnP, and MasaCtrl in Fig. 4 in the main paper.

Partial Object Swapping. As in Fig. 3, our method precisely swaps the cat head with a raccoon head harmoniously without influencing other pixels. Meanwhile, Photoswap swaps the whole body and modified the context pixels. When our proposed masked variable swapping is added, Photoswap achieves a better background preservation performance.

Cross-domain Swapping. *SwapAnything* is capable of swapping between styles and textual. In Fig. 3, a bear is adapted into a logo while keeping the gesture of the source object horse. Meanwhile, Photoswap fails to complete the challenging task. Also, when masked variable swapping is added, Photoswap achieves a better adaptation performance.

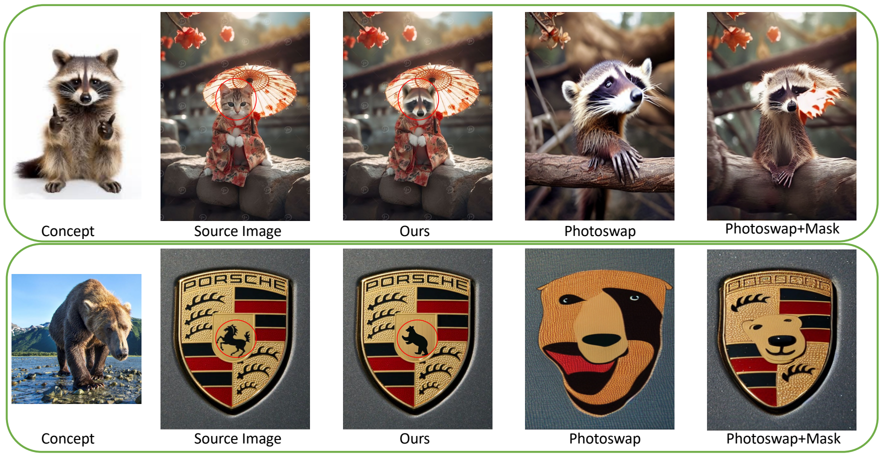


Fig. 3: Comparison with Photoswap on partial object swapping and cross-domain swapping. The upper part shows *SwapAnything* could localize the swapping area while Photoswap inevitably modified the background. In the lower part, *SwapAnything* adapts a bear into the style of a logo, while Photoswap failed on this cross-domain swapping task.

Multi-object Swapping. Multi-object swapping is a big step after single-object swapping. First, previous methods usually have a background modification such that continuous editing would accumulate unwanted distortion, which leads to a totally different image and fails the task of swapping. The second issue is that previous methods are usually designed for main subject swapping and do not pay attention to other objects. In this case, the objects in the following swapping steps could disappear in the previous swapping process.

F More Qualitative Results

Table 1: Automatic evaluation results. *SwapAnything* outperforms all other methods across all metrics.

	Ours	Photoswap	P2P	PnP	MasaCtrl	BlipDiffusion	DreamEdit	CopyPaste
$DINO_{fore}$	**0.61**	0.55	0.47	0.49	0.29	0.44	0.52	0.56
$CLIP_{fore}$	**0.79**	0.53	0.71	0.73	0.46	0.54	0.61	0.75
$DINO_{back}$	**0.79**	0.75	0.68	0.64	0.71	0.71	0.76	0.77
$CLIP_{back}$	**0.89**	0.86	0.75	0.70	0.67	0.76	0.82	0.79

We also conducted automatic evaluation. Following [2,6,8], we employ both DINO and CLIP-I as tools to evaluate the quality of the images generated. These two metrics serve as complementary indicators to the results obtained from human evaluations. As in Tab. 1, *SwapAnything* outperforms all other baselines in terms of both subject identity swapping and background preservation, which is consistent with the results of human evaluation.

Tab. 2 shows results on human evaluation on both human and non-human images on top baselines. PS means Photoswap [2]; P2P means Prompt-to-Prompt [3]; PnP means Plug-and-Play [9]; DE means DreamEdit [6]. We also conduct comparisons with another baseline PbE, Paint-by-Example [10].

Table 2: User study results. The 2nd to 5th rows and 6th to 9th rows show results on human objects and non-human objects.

	Ours	PS	Tie	Ours	P2P	Tie	Ours	PnP	Tie	Ours	DE	Tie	Ours	PbE	Tie
SS	**59.0**	10.0	31.0	**52.7**	18.2	24.1	**58.8**	29.2	12.0	**53.4**	16.5	30.1	**62.1**	12.0	28.0
SG	**44.0**	33.7	22.3	**54.5**	29.1	16.4	**61.6**	33.3	5.1	**42.4**	17.2	40.4	**73.1**	15.8	12.0
BP	**45.4**	32.2	22.4	**49.9**	26.9	23.2	**49.7**	22.0	28.3	**43.8**	18.0	38.2	**42.1**	31.3	27.0
OQ	**47.3**	24.3	28.4	**58.4**	23.3	18.3	**51.6**	31.1	17.3	**47.5**	26.5	26.0	**52.1**	21.9	30.0
SS	**45.6**	15.0	39.4	**52.9**	17.6	29.5	**45.8**	30.6	23.6	**56.8**	23.7	19.5	**58.1**	12.0	27.8
SG	**45.0**	34.3	20.7	**44.5**	34.9	20.6	**49.4**	32.7	17.9	**46.2**	20.4	33.4	**69.3**	15.0	14.8
BP	**37.6**	31.8	30.6	**39.1**	30.9	30.0	**50.7**	19.8	29.5	**44.2**	19.8	36.0	**40.5**	31.5	27.6
OQ	**51.3**	31.3	17.4	**51.6**	31.1	17.3	**47.5**	26.5	26.0	**48.1**	21.2	30.7	**48.1**	18.3	29.6

G Failure Cases

We highlight one common failure scenario encountered in our experiments. The challenge arises when dealing with subjects that exhibit a high degree of variability or freedom of movement. In such cases, as shown in Fig. 4, accurately replicating the concept subject becomes difficult. To address this, we are considering the implementation of explicit alignment, which we aim to explore in our future work.

H Human Evaluation Interface

Amazon Turker was presented with one reference image mainly containing the concept subject, one source image to be swapped, and two generated images from *SwapAnything* and a baseline.

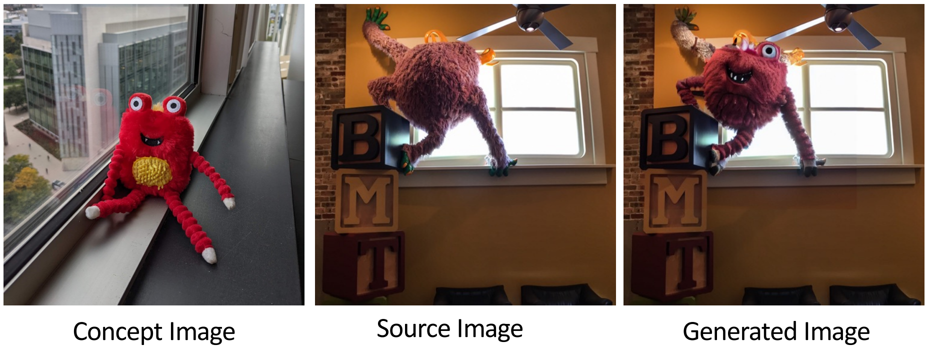
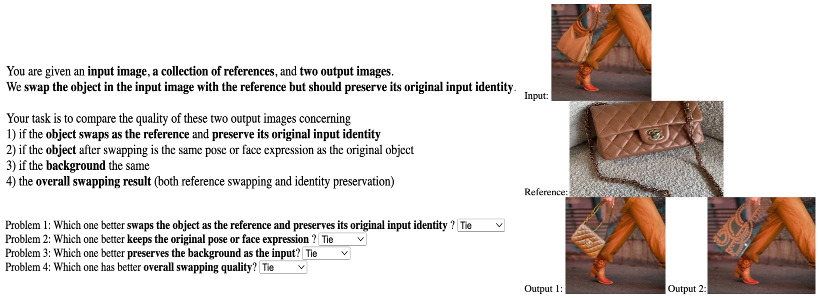


Fig. 4: Examples of failure cases. The model sometimes struggles to keep the details inside the mask area and could fail if the object has a high degree of freedom.



You are given an **input image**, **a collection of references**, and **two output images**.
We **swap the object in the input image with the reference but should preserve its original input identity**.

Your task is to compare the quality of these two output images concerning
1) if the **object swaps as the reference** and **preserve its original input identity**
2) if the **object** after swapping is the same pose or face expression as the original object
3) if the **background** the same
4) the **overall swapping result** (both reference swapping and identity preservation)

Problem 1: Which one better **swaps the object as the reference and preserves its original input identity** ? Tie
Problem 2: Which one better **keeps the original pose or face expression** ? Tie
Problem 3: Which one better **preserves the background as the input**? Tie
Problem 4: Which one has better **overall swapping quality**? Tie

Fig. 5: The illustration of the user study interface.

References

1. Choi, J., Choi, Y., Kim, Y., Kim, J., Yoon, S.: Custom-edit: Text-guided image editing with customized diffusion models. arXiv preprint arXiv:2305.15779 (2023)
2. Gu, J., Wang, Y., Zhao, N., Fu, T.J., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., Wang, X.E.: Photoswap: Personalized subject swapping in images (2023)
3. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2022)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
5. Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems (2023)
6. Li, T., Ku, M., Wei, C., Chen, W.: Dreamedit: Subject-driven image editing. arXiv preprint arXiv:2306.12624 (2023)
7. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
8. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In: CVPR (2023)
9. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
10. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)