A Results on Real Images

Figure 1: Face swapping on real images. Top row are real photos of Dr. Fei-Fei Li, and bottom row are real movie scenes. Photoswap precisely swaps human subjects while preserving the same subject pose and background context.

Figure 2: More results on human face swapping. Given a source human image (leftmost), Photoswap replaces the human identity with the reference person (bottom row) to generate the target image (top row).

This section showcases the practical effectiveness of our Photoswap method on real images in Figure[1] Figure[2] and Figure[3]. These results provide a visual testament to the successful execution...
of subject swapping in real-world instances. For the implementation of subject swapping on actual
images, we require an additional process that utilizes an image inversion method, specifically the
DDIM inversion, to transform the image into initial noise. This inversion method relies on
a reversed sequence of sampling to achieve the desired inversion. However, there exist inherent
challenges when this inversion process is applied in text-guided synthesis within a classifier-free
guidance setting. Notably, the inversion can potentially amplify the accumulated error, which could
ultimately lead to subpar reconstruction outcomes. To fortify the robustness of the DDIM inversion
and to mitigate this issue, we further optimize the null text embedding, as detailed in Mokady et al.
(3). The incorporation of this optimization technique bolsters the effectiveness and reliability of the
inversion process, consequently allowing for a more precise reconstruction.

B Results of Other Concept Learning Methods

Our mainly use Dreambooth as the concept learning method in the experiments, primarily due to its
superior capabilities in learning subject identities (5). However, our method is not strictly dependent
on any specific concept learning method. In fact, other concept learning methods could be effectively
employed to introduce the concept of the target subject.

To illustrate this, we present the results of Photoswap when applying Text Inversion (1). We train the
model using 8 A100 GPUs with a batch size of 4, a learning rate of 5e-4, and set the training steps to
1000. Results in Figure 4 indicate that Text Inversion also proves to be an effective concept learning
method, as it successfully captures key features of the target object. Nevertheless, we observe that
Text Inversion performance is notably underwhelming when applied to human faces. We postulate
that this is because Text Inversion focuses on learning a new embedding for the novel concept, rather
than finetuning the entire model. Consequently, the capacity to express the new concept becomes
inherently limited, resulting in its less than optimal performance in certain areas.
C Attention Swapping Step Analysis

In this section, we visualize the effect of the influence of swapping steps of different components. As discussed in the main paper, self-attention output $\phi$, and self-attention map $M$, derived from the self-attention layer, encompasses comprehensive content information from the source image, without relying on direct computation with textual features. Previous works such as Hertz et al. [2] did not explore the usage of $\phi$ and $M$ in the object-level image editing process.

Figure 5: Results at different swapping steps. With consistent steps, swapping the self-attention output provides superior control over the layout, including the subject’s gestures and the background details. However, excessive swapping could affect the subject’s identity, as the new concept introduced through the text prompt might be overshadowed by the swapping of the attention output or attention map. This effect is more clear when swapping the self-attention output $\lambda_\phi$. Furthermore, we observed that replacing the attention map for an extensive number of steps can result in an image with significant noise, possibly due to a compatibility issue between the attention map and the $v$ vector.

Figure [5] provides a visual representation of the effect of incrementally increasing the swapping step for one $\lambda$ hyperparameter while maintaining the other two at zero. Although all of them can be utilized for subject swapping, they demonstrate varying levels of layout control. At the same swapping step, the self-attention output $\phi$ offers more robust layout control, facilitating better alignment of gestures and preservation of background context. In contrast, the self-attention map $M$ and cross-attention map $A$ demonstrate similar capabilities in controlling the layout.

However, extensive swapping can affect the subject’s identity, as the novel concept introduced via the text prompt might be eclipsed by the swapping of the attention output or attention map. This effect becomes particularly evident when swapping the self-attention output $\lambda_\phi$. This analysis further informs the determination of the default $\lambda_\phi$, $\lambda_M$, and $\lambda_A$ values. While the cross-attention map $A$ facilitates more fine-grained generation control, given its incorporation of information from textual tokens, we discovered that $\phi$ offers stronger holistic generation control, bolstering the overall output’s quality and integrity.

D Ethics Exploration

Like many AI technologies, text-to-image diffusion models can potentially exhibit biases reflective of those inherent in the training data [4,6]. Given that these models are trained on vast text and image datasets, they might inadvertently learn and perpetuate biases, such as stereotypes and prejudices, found within this data. For instance, should the training data contain skewed representations or descriptions of specific demographic groups, the model may produce biased images in response to related prompts.

However, Photoswap has been designed to mitigate bias within the generation process of a text-to-image diffusion model. It achieves this by directly substituting the depicted subject with the
Figure 6: Human face swapping across different races. As you can see, the skin colors are also successfully transferred when swapping a white person with a black person, or vice versa.

intended target. In Figure 6 we present our evaluation of face swapping across various skin tones. It is crucial to observe that when there is a significant disparity between the source and reference images, the swapping results tend to homogenize the skin color. As a result, we advocate for the use of Photoswap on subjects of similar racial backgrounds to achieve more satisfactory and authentic outcomes. Despite these potential disparities, the model ensures the preservation of most of the target subject’s specific facial features, reinforcing the credibility and accuracy of the final image.

E Self-Attention Map Visualization

In Figure 7 we show more visualization on self-attention map for real images. Here we show four more examples of real images and synthetic images. The visualization results are consistent with those in the main paper.

Figure 7: Self-attention visualization results. The top two rows are synthetic images and the bottom two rows are real images. There is a high correlation between self-attention maps and the images.
Failure Cases

Here we highlight two common failure cases. First, the model struggles to accurately reproduce hands. When the subject includes hands and fingers, the swapping results often fail to precisely mirror the original hand gestures or the number of fingers. This issue could be an inherited challenge from Stable Diffusion. Moreover, Photoswap can encounter difficulties when the image comprises complex information. As illustrated in the lower row of Figure 8, Photoswap fails to reconstruct the complicated formula on a whiteboard. Therefore, while Photoswap exhibits strong performance across various scenarios, it’s crucial to acknowledge these limitations when considering its application in real-world scenarios involving intricate hand gestures or complex abstract information. Future work will focus on addressing these issues to enhance the overall performance and versatility of the model.

Figure 8: Failure cases.

References