Diffusion Handles
Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D

Supplementary Material

S1. Overview

In the supplementary material, we provide a full quantitative comparison to all baselines on our `Benchmark` dataset in Section S2, give full qualitative comparison on both the `PhotoGen` and the `Benchmark` datasets in Sections S3 and S4, provide more information about the foreground selection diffusion sampler in Sections S6 and S7, and additional information about our user study in Section S8.

A webpage for our method that provides an overview, more detailed stop-motion results, and also hosts the full qualitative results is available in the supplementary material.

S2. Full Quantitative Comparison

We provide a full quantitative comparison on the `Benchmark` dataset, using the same metrics described in the main paper:
(A) For *Identity Preservation*, we use a cycle consistency metric that measures the difference between the original image and the edited image transformed back to the original object configuration using the inverse 3D transform. Denoting the target 3D edit as T and the image edit performed by our method or a baseline as $\mathcal{E}$ (lower is better):

$$E_{\text{id}}^{\text{L1}} = \|x_0 - \mathcal{E}(T^{-1}, \mathcal{E}(T, x_0))\|_1, \text{ and} \quad (10)$$

$$E_{\text{id}}^{\text{LPIPS}} = \text{LPIPS}\big(x_0,\ \mathcal{E}(T^{-1}, \mathcal{E}(T, x_0))\big), \quad (11)$$

where we use either $L1$ or LPIPS [60] to measure the image difference.
(B) For *Edit Adherence*, we measure the Intersection over Union (IoU) between the mask M_e of the edited foreground object and the corresponding ground truth mask M_e^{gt} (higher is better):

$$S_{\text{edit}} = \text{IoU}(M_e, M_e^{\text{gt}}). \quad (12)$$

The mask M_e is using the same foreground segmentation approach described in Section 5.1 (Step 2) that is based on SAM [24], starting from a Grounding DINO [28] bounding box. The ground truth mask M_e^{gt} is obtained by applying the target 3D edit to the synthetic scenes our `Benchmark` dataset was created from.

Results are shown in Table S1. For additional insight, compare to the qualitative results given in Section S4. We can see that Zero123 performs worst on identity preservation, as it introduces significant distortions during object edits that accumulate in the edit cycle. ObjectStitch suffers less from distortions, but has a lower degree of identity preservation in each edit. 3DIT performs second best, although looking at the qualitative results, we can see that this good quantitative performance is deceptive: 3DIT often fails to change the input image at all, resulting in good identity preservation, but bad edit adherence. For fairness, we ignore 3DIT results that do not change the foreground object at all when computing identity preservation. All baselines have relatively low edit adherence. ObjectStitch does not provide 3D controls, while Zero123 and 3DIT lack accuracy in their 3D controls. Our 3D-aware guidance provides both more accurate control and better identity preservation.

Table S1. **Quantitative comparison on the `Benchmark` dataset.** We compare *identity preservation*, based on the cycle consistency of performing the edit, followed by its inverse; and *edit adherence*, as measured by the IoU between image region covered by the edited foreground object and the corresponding ground truth image region.

	Identity Preservation		Edit Adherence
	$E_{\text{id}}^{\text{L1}}(\times 10)\downarrow$	$E_{\text{id}}^{\text{LPIPS}}\downarrow$	$S_{\text{edit}}\uparrow$
Obj.Stitch [52]	0.89	<u>0.25</u>	0.37
Zero123 [27]	1.05	0.31	<u>0.52</u>
3DIT [31]	<u>0.74</u>	0.27	0.15
Ours	**0.71**	**0.19**	**0.85**

S3. Qual. Comparison on `PhotoGen` dataset

The full qualitative comparison on all samples from the `PhotoGen` dataset is linked from the webpage.

S4. Qual. Comparison on `Benchmark` dataset

The full qualitative comparison on all samples from the `Benchmark` dataset is is linked from the webpage.

S5. Object Insertion

Our approach also supports insertion without significant changes, by transplanting features and depth from another image. Initial results in Figure S2 look promising; background preservation can likely be improved by adjusting parameters.

S6. Foreground Selection Details

The choice of foreground segmentation method is orthogonal to our approach. Any segementation method that returns a mask of the foreground object can be used. In our experiments, we opt for an open-set segmentation using SAM [24], starting from a bounding box found with Grounding DINO [28]. This gives the user the option to select the

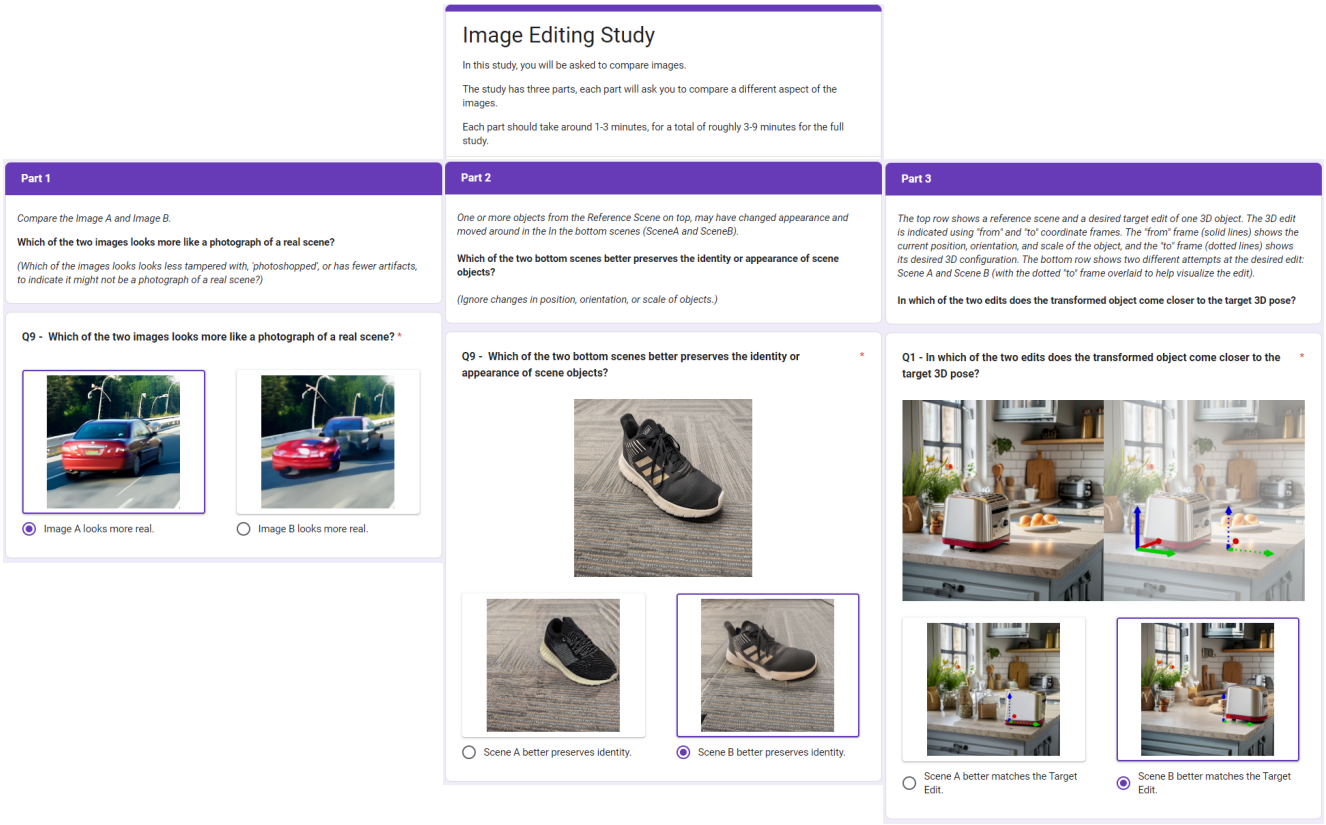

Figure S1. **User Study Screenshot.** We asked three types of questions in the user study to measure plausibility (left), identity preservation (middle), and edit adherence (right). In each question, the user chooses between two images.

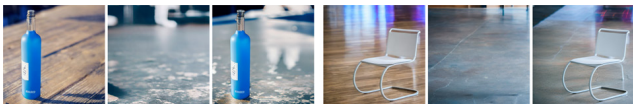

Figure S2. **Examples of object insertion.** We show examples of object insertion (showing foreground, background, result).

object of interest interactively, by describing the object with a text prompt and selecting one of the resulting candidate segments. Other approaches, like using SAM to identify the object based on a click or a bounding box could also be used.

S7. Diffusion Sampler Details

We use the DDIM sampler as described in Denoising Diffusion Implicit Models [51] in in all our experiments, using 50 denoising steps. When computing the edited image with our guidance energy $\mathcal{G}$, in each iteration we perform three steps of gradient descent $\nabla_{\tilde{x}(t)}\, \mathcal{G}(\tilde{x}(t); t, y, d)$ (see Eq. 3) on our energy to nudge the denoising trajectory in a direction that minimizes the guidance energy.

S8. User Study Details

Figure S1 shows screenshots with examples for each type of question in our user study. We split the study into three parts, corresponding to plausibility, identity preservation, and edit adherence. In each part we ask one specific type of question, and the user chooses between two images as response. We start by asking about image plausibility in the first part, as this does not require introducing the notion of a 3D edit. In the second part, we ask about identity preservation, using the input image as reference, and in the third part, we ask about edit adherence, using both the input image and a visualization of the target edit as reference.