PS-Diffusion: Photorealistic Subject-Driven Image Editing with Disentangled Control and Attention

Weicheng Wang[1], Guoli Jia[3], Zhongqi Zhang[1], Liang Lin[2,4], Jufeng Yang[1,2,5*]

[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University, Tianjin, China.

[2] Pengcheng Laboratory, Shenzhen, China.

[3] Electrical Engineering Department, Tsinghua University, Beijing, China.

[4] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China.

[5] Nankai International Advanced Research Institute (SHENZHEN·FUTIAN), Shenzhen, China.

2120230639@mail.nankai.edu.cn, exped1230@gmail.com

15692233416@163.com, linliang@ieee.org, yangjufeng@nankai.edu.cn

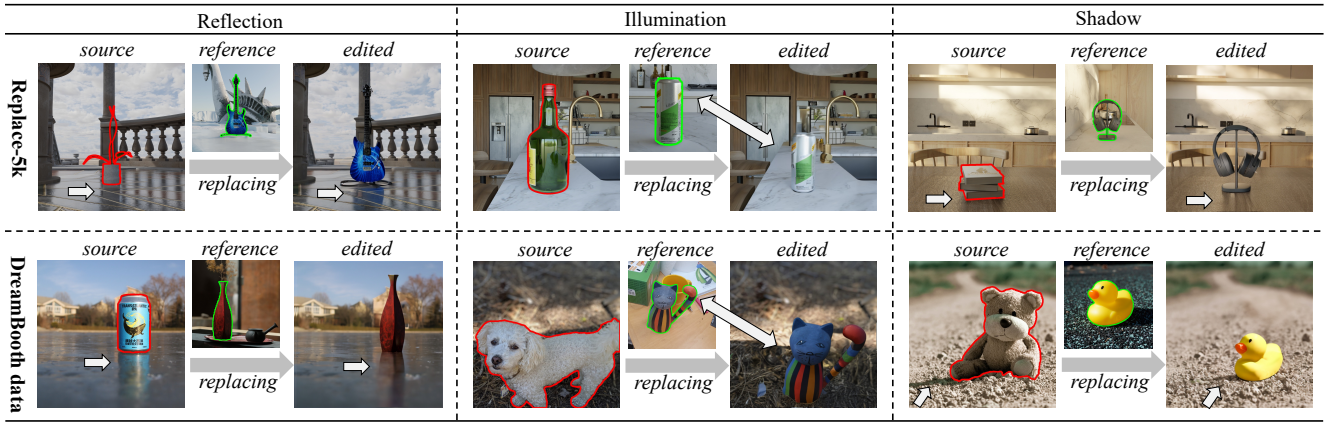

Figure 1. We aim to edit the object (red boundary) in the source with the object (green boundary) in the reference. PS-Diffusion simulates consistent contextual interactions of objects on scenes, such as reflection, illumination, and shadow, achieving **photorealistic** image editing and keeping the target appearance unchanged. Results from our proposed Replace-5K and real-world data in DreamBooth are shown.

Abstract

Diffusion models pre-trained on large-scale paired image-text data achieve significant success in image editing. To convey more fine-grained visual details, subject-driven editing integrates subjects in user-provided reference images into existing scenes. However, it is challenging to obtain photorealistic results, which simulate contextual interactions, such as reflections, illumination, and shadows, induced by merging the target object into the source image. To address this issue, we propose PS-Diffusion, which ensures realistic and consistent object-scene blending while maintaining the invariance of subject appearance during editing. To be specific, we first divide the contextual interactions into those occurring in the foreground and the background areas. The effect of the former is estimated through intrinsic image decomposition, and the region of the latter is predicted in an additional background effect control branch. Moreover, we propose an effect attention module to disentangle the learning processes of interaction and subject, alleviating confusion between them. Additionally, we introduce a synthesized dataset, Replace-5K, consisting of 5,000 image pairs with invariant subject and contextual interactions via 3D rendering. Extensive quantitative and qualitative experiments on our dataset and two real-world datasets demonstrate that our method achieves state-of-the-art performance. ***The code is available in the https://github.com/wei-cheng777/PS-Diffusion.***

* Corresponding author.

1. Introduction

Image editing aims to modify images following user-specified editing targets, and unrelated content remains consistent with the source images. Benefiting from the recent success of diffusion models, image editing has received unprecedented progress. As a highly controllable task, image editing based on diffusion models [18, 37, 66] is widely applied in real-world scenarios such as personalization [35, 50], try-on [29, 43, 73], and E-commerce [5, 10].

To convey fine-grained visual concepts and precisely articulate the user's desires, subject-driven image editing [35, 45, 62] (SDIE) integrates the subject in reference image into an existing scene. Built on the diffusion model, Paint-by-Example [62] proposes a subject-driven editing framework for inpainting masked regions guided by the CLIP embeddings of reference images. Anydoor [8] enhances identity consistency by injecting self-supervised representation and detailed maps. To directly edit image pixels, Copy-Paste [14] is another simple method to integrate the entirely unchanging subject into the source image by copy and paste. However, the current diffusion-based methods mainly focus on preserving semantic consistency. Indiscriminately enhancing consistency via copy overlooks the contextual interaction of the object on the scene, such as reflection, illumination, and shadow in Fig. 1. This poses a significant challenge to acquiring photorealistic results.

To achieve photorealistic SDIE, there are pioneering works initially explored to simulate contextual interactions of objects on scenes, as well as maintain high consistency of subject identity. Graphics-based methods [36, 52–54, 61] render contextual interactions based on 3D assets and physical parameters. To relax the requirement for highly-cost parameters, an alternative solution is utilizing two-stage generative editing methods, removing the source object by inpainting [27, 47, 63] and then inserting [19, 59] the targets. However, they still face the following challenges in achieving photorealism: 1) Removal stage: Most inpainting models [47, 74] struggle to eliminate the original contextual interactions of the source object outside the mask indicating the editing area. 2) Insertion stage: Harmonization-like methods [19, 58] primarily affect the object area within the mask. 3) Consistency across stages: Although some object insertion or removing methods [41, 57, 59] account for physical laws, the individual process overlooks the consistency of contextual interactions before and after editing.

Facing the above issues, we propose an end-to-end **P**hotorealistic **S**ubject-driven image editing approach, PS-Diffusion. To cover most contextual interactions of the object on the scene, we divide them into two aspects based on affecting regions: 1) Effects on the foreground within the mask, such as lighting. 2) Effects on the background outside the mask, such as shadow and reflection. Then, we use disentangled control signals to guide respective editing processes in the diffusion model. For the former, we condition the diffusion denoising process with illumination-dependent properties estimated by intrinsic image decomposition. These properties derived from the source images improve the consistency of contextual interactions during editing. For the latter, we introduce an additional background effect control branch to perceive effect regions outside user-defined masks by learning an effect map. Unlike object editing, modifications to scene effects typically build upon existing images without significantly altering the original pixels. Therefore, directly weighting the two branches causes confusion between the learning of interactions and subjects. To separate them, we introduce the Effect Attention Module (EAM) in denoising U-Net, which reorganizes the structural features of interaction regions and lighting conditions to improve the plausibility of effects. Finally, due to the scarcity of paired data, we curated a dataset comprising 5,000 data pairs, *i.e*. Replace-5K, by Blender.

Our contributions are four-fold: 1) We propose PS-Diffusion for photorealistic SDIE, ensuring consistent contextual interactions and object identity. Disentangled control allows the model to recognize interactions without being restricted by masks. 2) EAM enhances visual plausibility by separating interaction and subject learning. 3) We introduce an exquisite synthetic dataset, Replace-5K, with 5,000 image pairs that simulate the interactions between scenes and objects. 4) Extensive experiments on synthetic and real-world datasets demonstrate the effectiveness.

2. Related Work

2.1. Image Editing

Priors of generative models play a crucial role in effective image editing. In GAN-based methods [60, 72], the image is inverted into the latent code of the pre-trained GAN [3, 28]. Recently, text-to-image (T2I) diffusion models [46, 47, 49] enhance image editing controllability. Benefiting from pre-training on image-text pairs, it is efficient to guide editing through texts of the target [13, 20, 55] or instructions that express the editing purpose [4, 26]. However, texts struggle to control fine-grained visual details. Utilizing the inherent visual information in images to guide editing has been extensively explored [17, 50, 65]. To precisely locate the editing position, the mask [1, 9, 40, 51] is another visual cue to control the editing area. Moreover, combining multiple control signals [12, 23] can address more complex needs. In this paper, we use image guidance and mask localization to achieve precise image editing.

2.2. Subject-Driven Control in Diffusion Models

Subject-driven control aims to align the subject of outputs with user-provided references [24, 35, 68]. Subject-driven image generation [13, 33, 50] creates new scenes containing

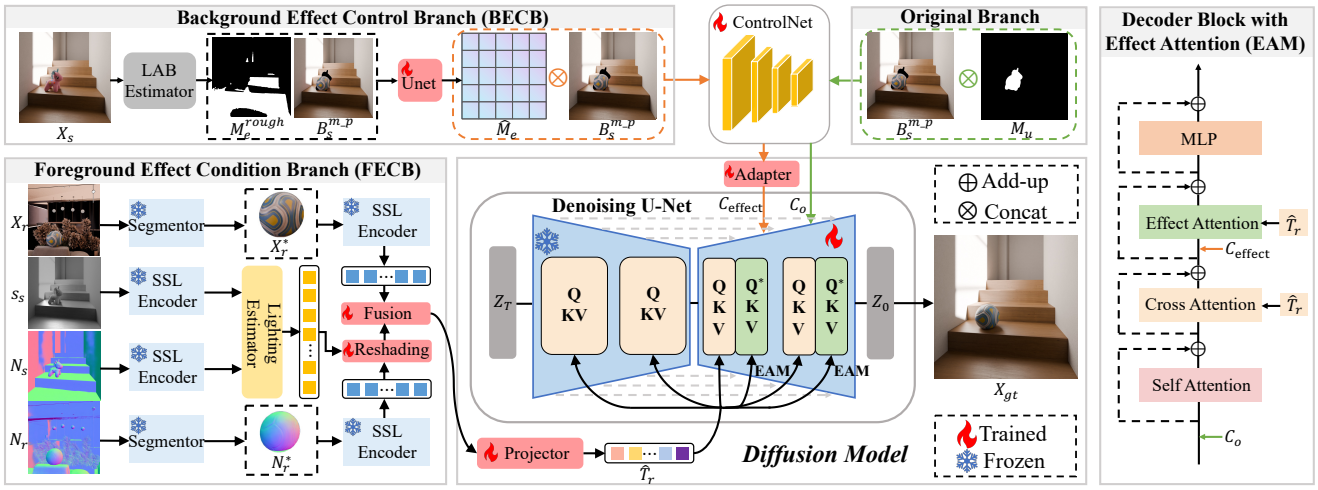

Figure 2. Pipeline of our method. We disentangle the control signals of contextual interactions into those within and outside M_u. EAM decouples the learning of the subject and the interaction. M_u is the union of the source object mask M_s and the target mask M_t. $B_s^{m\_p}$ is the background pasted with the object. S_s is the shading of X_s extracted by [6]. N_s and N_r^* are the normal of X_s and X_r^*.

user-provided subjects based on text. However, text alone only maintains the semantic consistency of scenes. Many researchers [35, 62] study subject-driven image editing to achieve finer control over background and foreground. Based on DreamBooth [50], DreamEdit [35] proposes an iterative strategy to inpaint the masked areas of the source image while protecting the background. Instead of fine-tuning each subject, Paint-by-Example [62] and ObjectStitch [56] train a unified model in a self-supervised manner by injecting image embeddings into the cross-attention layer. PhD [71] improves subject fidelity by inputting background images pasted with the subject into ControlNet [69]. AnyDoor [8] further improves identity preservation by extracting self-supervised embedding and high-frequency maps. Despite these advancements, most methods only focus on maintaining semantic similarity. Importantly, few works focus on the effects of interactions between scene and object during editing, which is crucial for achieving photorealism.

2.3. Physical Control in Diffusion Models

In image generation or editing, physical plausibility of contextual interactions is essential for realistic outcomes. Research on physical control can be classified into two main areas. On the one hand, some work targets specific physical properties, such as lighting [39] and shadows [34]. For instance, LightIt [31] incorporates lighting conditions and normal maps as additional controls in text-to-image diffusion models. SGDiffusion [38] proposes diffusion-based shadow generation methods. On the other hand, some incorporate general scene effects. ObjectDrop [59] proposes object removal and insertion complying with physical laws trained on a counterfactual dataset. Generating a dataset containing pairs with and without foreground, Tarrés *et al*. [57] allows the generation of shadows and reflections in object insertion. However, managing physical effects at an individual stage *e.g*. object removal or insertion, ignores the consistency of physical laws during editing. To ensure consistency supervision, we construct a physically consistent paired dataset by industrial rendering engines.

3. Method

3.1. Overview

We aim to edit the object in a source image with the subject from a reference image, guided by object masks as positioning information. Formally, an source image $X_s = \{F_s, B_s\} \in \mathbb{R}^{H\times W\times 3}$ consists of a foreground object F_s and a background B_s, while mask $M_s \in \mathbb{R}^{H\times W}$ indicates the location of the foreground object. During the editing, given a reference image $X_r = \{F_r, B_r\} \in \mathbb{R}^{H'\times W'\times 3}$ with mask $M_r \in \mathbb{R}^{H'\times W'}$ indicating the position of the reference object, F_s in the X_s is replaced by F_r. The target mask $M_t \in \mathbb{R}^{H^*\times W^*}$ is used to specify the location and size of F_r within X_s. Finally, result image $X_{gt} = \{F_r, B_s\}$ after editing will be obtained.

As shown in Fig. 2, our baseline consists of two main components: ControlNet [69] position control (original branch) and reference image condition. To control editing regions in the background, we first concatenate the masked source background B_s^m and mask M_u like [27, 62]. Here, M_u is the union of the source object mask M_s and the target mask M_t, $B_s^m = B_s \cdot (1 - M_u)$. To preserve the subject identity, we paste F_r in X_r onto the B_s^m as prior following [71], *i.e*. $B_s^{m\_p} = \text{Paste}(B_s^m, F_r)$. Then, the concatenated $B_s^{m\_p} \otimes M_u$ are input into ControlNet with coefficient α to get the feature C_o, where $\otimes$ is the concatenate opera-

tion. To guide the denoising process with the reference image, we replace the text embedding of Stable Diffusion [49] with the image embedding T_r extracted by DINOv2 [44]. Finally, a linear layer [8] is used to align the dimensions.

We incorporate the learning for contextual interactions of the object on the scene in SDIE by disentangled effect control and attention. The former decouples control signals of effects caused by interactions into those within and outside of M_u, and enhances respective performance by the condition of foreground effect and an additional background effect control branch in Stable Diffusion. EAM introduced in the latter improves the plausibility of effects by separating the learning of interactions from that of subjects.

3.2. Disentangled Effect Control

Foreground effect condition branch (FECB). FECB aims to estimate the effects on objects within M_u, such as lighting. To eliminate the influence of the background on the extraction of subject features, areas outside M_r are first filled with white pixels [8, 45], $X_r^* = \text{Filling}(X_r)$. However, the effect of B_r on the F_r remains. Given our goal to preserve the appearances of the subjects as much as possible, bias from B_r could lead to undesirable copy-paste artifacts. Given that lighting is the main factor, we condition the diffusion process with illumination-dependent effects using intrinsic image decomposition [6, 7]. To be specific, intrinsic image decomposition separates X_s and X_{gt} into the illumination-invariant properties A_s and A_{gt} (albedo), and the illumination-dependent effects S_s and S_{gt} (shading):

$$X_s = A_s \cdot S_s, X_{gt} = A_{gt} \cdot S_{gt} \tag{1}$$

To keep the consistency of illumination, we first estimate the lighting parameters l_s in the source image inspired by the parametric illumination model in [7]:

$$l_s = \text{Estimator}(T_{N_s}, T_{S_s}) \tag{2}$$

where T_{N_s} and T_{S_s} are embeddings extracted by DINOv2 from normal N_s and shading S_s of X_s. l_s is the high-dimensional lighting-related feature, not constrained by specific light models. Training on paired images rendered under diverse lighting and reflectance models enables PS-Diffusion to learn various lighting. Next, we obtain the illumination-dependent feature $T_{S_r^*}$ of X_r^* under l_s through reshading. Then, T_r in the baseline and $T_{S_r^*}$ are fused:

$$T_{S_r^*} = \text{Reshading}(T_{N_r^*}, l_s), \hat{T}_r = \text{Fusion}(T_r, T_{S_r^*}) \tag{3}$$

$T_{N_r^*}$ is embeddings from normal N_r^* of X_r^*. $\hat{T}_r$ is the final feature input into the projection layer. The Estimator, Reshading, and Fusion are two-layer MLPs. Moreover, we supervise $\hat{T}_r$ and $T_{S_r^*}$ with X_{gt} and S_{gt}:

$$\mathcal{L}_{sim} = 2 - \text{Cosine}(\hat{T}_r, T_{gt}) - \text{Cosine}(T_{S_r^*}, T_{S_{gt}}) \tag{4}$$

where T_{gt} and $T_{S_{gt}}$ are DINOv2 embedding of X_{gt} and S_{gt}. Benefitting from paired training data, FECB can autonomously learn additional attributes, such as roughness. By guiding the denoising process with $T_{S_r^*}$, our model not only accounts for effects of contextual interactions on the foreground but also potentially steers the editing of effects on the background through the cross-attention mechanism.

Background effect control branch (BECB). Although FECB transfers lighting, ControlNet's structural constraint in the original branch limits editing outside M_u. BECB aims to learn structural features for object-scene interactions projected onto the background outside M_u, such as shadows and reflections. Some works [57, 59] directly finetune on large-scale datasets to implicitly learn effects outside masks. We aim to explicitly learn the effects region in an additional branch. Specifically, given a background $B_s^{m\text{-}p}$ pasted with the reference object, we predict an effect map $\hat{M}_e$ through a function $\mathcal{F}$ to indicate where the effects of objects on the scene will appear in the background. Similarly, the concatenated $B_s^{m\text{-}p} \otimes \hat{M}_e$ is fed to ControlNet. The final feature C_{effect} is scaled with an additional coefficient β:

$$C_{\text{effect}} = \beta \cdot \text{ControlNet}(B_s^{m\text{-}p} \otimes \hat{M}_e) \tag{5}$$

Given that the original branch of ControlNet has learned the structural information related to the mask, we share their weights, avoiding additional training burden.

It is not trival to directly train $\mathcal{F}$ to predict $\hat{M}_e$ while keeping consistent physical laws during editing. To accelerate learning, we exploit the consistency prior of effects on scenes. Specifically, although objects differ, their effects on the same scene often exhibit strong consistency in certain attributes. For instance, if a scene is cast shadows of objects, the direction of these shadows typically remains consistent, regardless of changes to objects. To utilize this prior, we estimate a rough mask M_e^{rough} from X_s in LAB color space to further guide the $\mathcal{F}$, as most effects on the background involve changes in brightness [42]. Following [42], we calculate mean values μ_L, μ_A and μ_B of the pixels in L, A, and B planes of X_s. If $\mu_A + \mu_B \leq \tau$, we classify pixels in L plane as effect pixels when they meet $L(x,y) \leq \mu_L - \frac{\sigma_L}{3}$:

$$M_e^{\text{rough}} = \begin{cases} 1, & \text{if } L(x,y) \leq \mu_L - \dfrac{\sigma_L}{3} \ \wedge \ \mu_A + \mu_B \leq \tau \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where τ is a threshold, σ_L denotes the standard deviation of the L plane. Finally, we concatenate M_e^{rough} and $B_s^{m\text{-}p}$ and input them into $\mathcal{F}$, resulting in:

$$\hat{M}_e = \mathcal{F}(B_s^{m\text{-}p} \otimes M_e^{rough}) \tag{7}$$

Moreover, benefiting from our paired dataset, we can acquire the label M_e indicating the regions that need to be edited when the object changes. Given a pair of X_s and

X_{gt}, we first calculate the pixel-wise difference $M_e(i,j) = |X_{gt}(i,j) - X_s(i,j)|$. Then, we apply the threshold τ' to binarize and exclude the area of M_u. Under the supervision of M_e, we optimize $\mathcal{F}$ using a binary segmentation loss, specifically the BCEWithLogitsLoss:

$$\mathcal{L}_m = \text{BCEWithLogitsLoss}(\hat{M}_e, M_e) \tag{8}$$

In addition, $\hat{M}_e$ is a smoothed map rather than a binary mask, which supports the simultaneous training of $\mathcal{F}$ and Stable Diffusion. This effectively leverages the physical knowledge learned from large-scale pre-training [32, 67]. $\mathcal{F}$ in our method is a simple U-Net.

3.3. Effect Attention Empowered Denoising U-Net

Unlike object editing tasks such as inserting or restoring, editing of effects augments existing images without significantly altering the original pixels [16]. For example, when shadows or reflections are projected onto the background, the underlying scene remains discernible to the viewer. On the one hand, since ControlNet controls the editing position, directly weighting the features from the two branches confuses the learning for editing objects and effects, leading to disruptions in the background or foreground. On the other hand, there is an inherent gap between the structural features of BECB and the lighting features of FECB. Therefore, we propose an Effect Attention Module (EAM) to separate them and reorganize disentangled effect control.

Given the effects are built upon the existing background or foreground, EAM introduces an extra attention layer after the cross-attention layer in the decoder of U-Net. The output F_{cross} from the cross-attention incorporates reference image features from condition and source background features from the original ControlNet branch. For effect control outside mask, the query Q_{EAM} for the EAM is formed by combining F_{cross} with C_{effect}, indicating effect regions:

$$Q_{\text{EAM}} = F_{\text{cross}} + (W \cdot C_{\text{effect}} + b) \tag{9}$$

W and b are the Adapter's weight matrix and bias vector. Adapter aligns dimensions and adjusts C_{effect} to effects editing task. The key K_{EAM} and value V_{EAM} are derived from $\hat{T}_r$, which captures the effects within mask. The attention weights for the EAM are then calculated as follows:

$$M = \text{Softmax}(\frac{Q_{\text{EAM}} K_{\text{EAM}}^T}{\sqrt{d}}), K_{\text{EAM}} = V_{\text{EAM}} = \hat{T}_r \tag{10}$$

where d is the dimensionality of the key and query. The output F_{EAM} of EAM is given by $F_{\text{EAM}} = M \cdot V_{\text{EAM}}$. Finally, EAM modulated F_{cross} in the form of a residual connection with a coefficient ω: $F_{\text{cross}} + \omega \cdot F_{\text{EAM}}$.

4. Benchmark

Many works employ existing datasets to construct self-supervised training, such as by data augmentation [62] or video data [8]. However, self-supervised is insufficient for learning physical laws [59]. We leverage the powerful 3D computer graphics suite Blender [21] to produce the necessary paired training data. With industrial renderer Cycles, Blender can capture physical effects such as lighting, shadows, and reflections [15]. Specifically, we construct each pair of data by two triplets: image triplet (X_r, X_s, X_{gt}) and corresponding object mask triplet (M_r, M_s, M_{gt}).

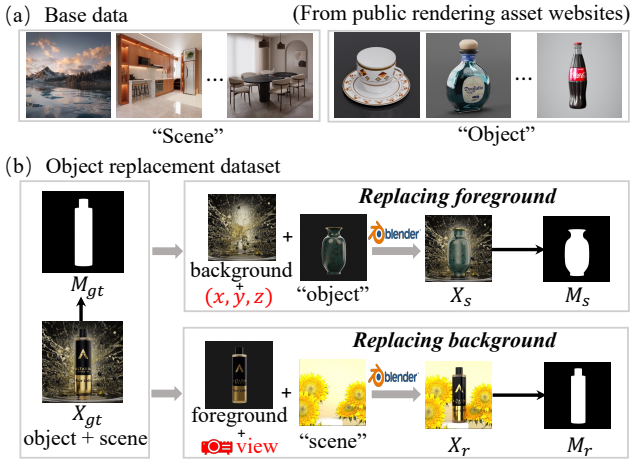

Figure 3. Dataset Preparation. (a) we collect the base data: **Scene** and **Object**. (b) We create the paired data with Blender.

Base data collection. As shown in Fig. 3 (a), we collect base data crafted by professional modelers from public rendering asset websites. For **Object** assets, we use classes in Open Images V7 [2] as keywords to enhance diversity while excluding categories unsuitable as foreground. The **Scene** assets are sourced from six categories: indoor, outdoor, e-commerce, technology, ancient buildings, and abstract.

Triplet data creation. As depicted in Fig. 3 (b), starting from X_{gt} composed of object and scene, we replace the foreground at the same coordinate positions with random objects from **Object** set, resulting in X_s. For the object in X_{gt}, we replace the background from **Scene** set while maintaining the same camera viewpoint, making X_r. Images with inharmonious foreground and background combinations are filtered out. Additionally, we render an image with only the object, making the scene transparent. By identifying non-transparent areas, we obtain object masks. Finally, we obtain 5,000 pairs of data, which constitute Replace-5K.

5. Experiments

5.1. Implementation Details

Following [8], Stable Diffusion V2.1 [49] is the base architecture. Intrinsic image decomposition and normal estimation use the off-the-shelf method [6] and [11]. α and β are set to 1.0 and 0.1. We set the initial learning rate as $1e^{-5}$. We train the model at a batch size of 8 for 40 epochs.

Table 1. Quantitative comparisons on Replace-5K. The standard reconstruction metrics and perceptual similarity metrics are applied. The consistency of subjects is evaluated via CLIP-I and DINO-I. PBE*, ObjectStitch*, and AnyDoor* are fine-tuned on our Replace-5K.

Method	PSNR↑	CLIP↑	DINO↑	LPIPS↓	FID↓	CLIP-I↑	DINO-I↑
IP-Adapter [65]	18.473	0.892	0.755	0.247	7.188	0.788	0.553
PBE* [62]	17.036	0.730	0.788	0.298	15.803	0.757	0.526
ObjectStitch* [56]	18.032	0.887	0.762	0.295	8.708	0.773	0.554
Anydoor* [8]	19.473	0.923	0.872	0.222	7.132	0.804	0.638
Remove [74]-Insert [19]	21.490	0.934	0.892	0.217	6.241	0.814	0.682
Ours	**22.583**	**0.945**	**0.921**	**0.174**	**5.109**	**0.827**	**0.688**

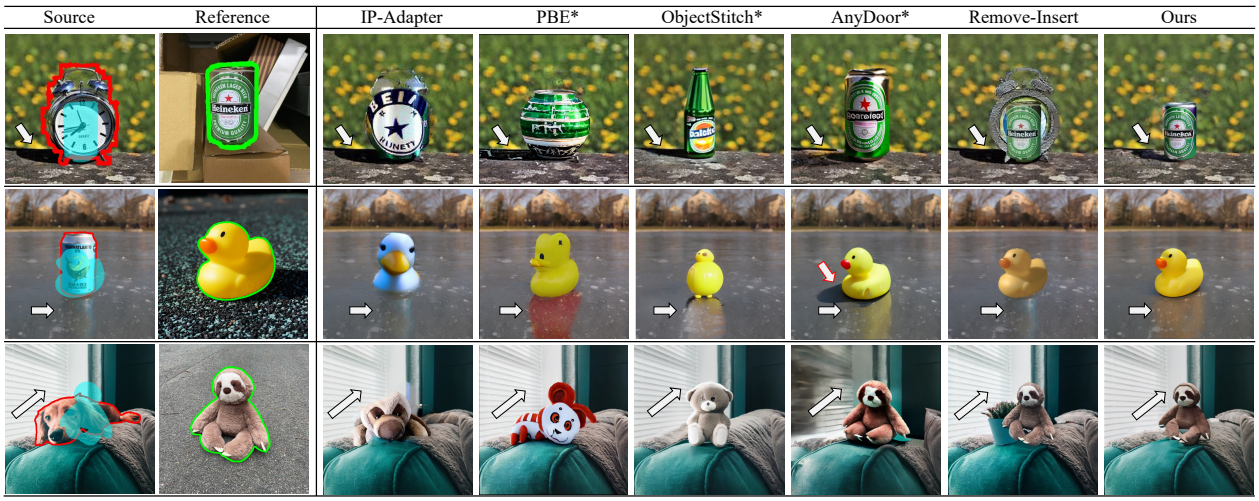

Figure 4. Qualitative comparison on data of DreamBooth and DreamEdit. We achieve consistent effect editing (shadows, reflections, lighting), precise control for the position and size of targets. The blue masks in source images indicate the target position and size.

5.2. Datasets

In Replace-5K, 4500 pairs are for training, and 500 are for testing. When testing, we obtain masks by SAM [30]. To validate the effectiveness on real-world images, we construct additional test data by datasets of DreamBooth [50] and DreamEdit [35]. Specifically, for an image with a subject, we randomly select another image with a different subject as the reference. The target mask is aligned according to the center of the source object.

5.3. Evaluation Settings

Compared methods. 1) One-stage methods includes IP-Adapter [65], PBE [62], ObjectStitch [56], AnyDoor [8]. We fine-tune PBE, ObjectStitch, and AnyDoor on Replace-5K. 2) We also compare with a two-stage approach involving removal followed by insertion, *i.e*. Remove-Insert. Object removal is the sota inpainting method PowerPaint [74], and insertion uses the sota harmonization method [19].

Quantitative metrics. For Replace-5K, we apply the reconstruction metric PSNR and perceptual similarity metrics CLIP [48], DINO [44], LPIPS [70]. The generative metric FID [22] is also used. To evaluate the consistency between subjects, we calculate the similarity CLIP-I and DINO-I between the edited region and the reference by CLIP and DINO. For DreamBooth and DreamEdit datasets, we evaluate CLIP-I, DINO-I, and FID, where the target distribution of FID is computed from all images in the original dataset.

Table 2. Quantitative comparison on datasets of DreamBooth and DreamEdit. CILP-I, DINO-I, and FID are used to evaluate.

Method	Dreambooth			Dreamedit		
	CLIP-I↑	DINO-I↑	FID↓	CLIP-I↑	DINO-I↑	FID↓
IP-Adapter [65]	0.795	0.581	25.550	0.797	0.581	23.988
PBE* [62]	0.756	0.543	32.000	0.769	0.552	30.977
ObjectStitch* [56]	0.787	0.598	25.043	0.782	0.600	28.210
Anydoor* [8]	0.811	0.666	23.626	0.825	0.686	25.607
Remove[74]-Insert[19]	0.840	0.736	23.607	0.849	0.748	24.191
Ours	**0.845**	**0.740**	**22.919**	**0.851**	**0.749**	**23.383**

5.4. Comparison with the Existing Methods

We quantitatively and qualitatively compare with 5 methods on Replace-5K and data of DreamBooth and DreamEdit.

Quantitative comparison. The results of Replace-5K are shown in Tab. 1. Overall, we achieve best performance across 7 evaluation metrics. For the pixel-level metric PSNR, PS-Diffusion shows an improvement of 1.093. In perceptual similarity metrics, PS-Diffusion demonstrates significant improvements due to the incorporation of effects of contextual interactions, which enhance photorealism. Specifically, we improve 0.011 and 0.029 on CLIP and DINO. LPIPS and FID are decreased by 0.043 and 1.132. In addition, due to the precise control of appearance and

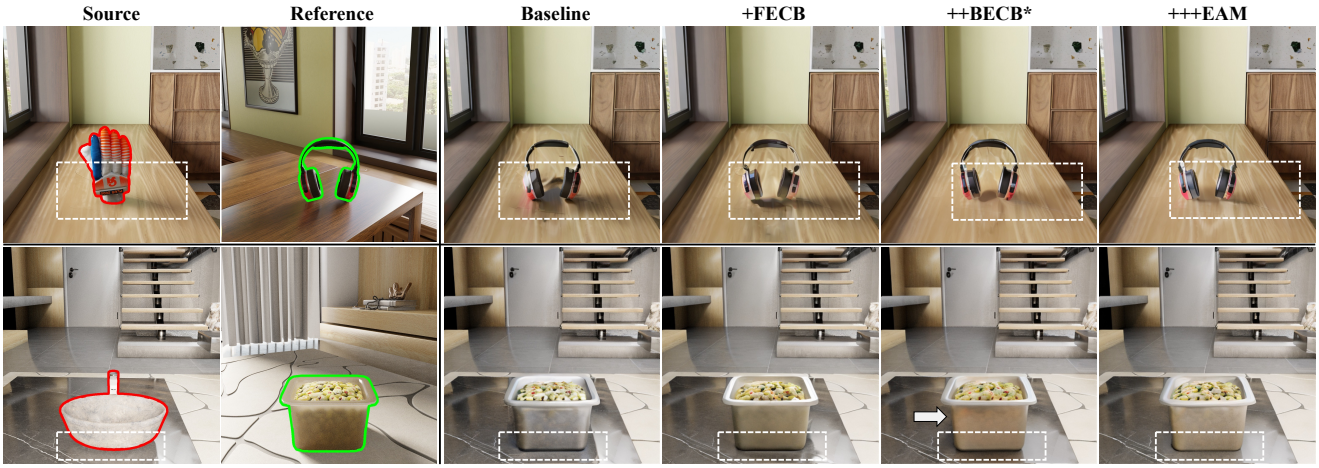

Figure 5. Qualitative ablation on the Replace-5K dataset. In ++BECB*, the features of FECB and the original branch are directly weighted.

size, we enhance subject consistency between the target and reference. CLIP-I and DINO-I are improved by 0.013 and 0.006. Moreover, we perform best on data of DreamBooth and DreamEdit, in Tab. 2. Although Remove-Insert sometimes approaches PS-diffusion on CLIP-I and DINO-I due to harmonization keeping the foreground unchanged, the risk of copy-paste artifacts results in higher FID.

Qualitative comparison. The results are illustrated in Fig. 4. First, our method not only achieves realistic contextual interactions, but also keeps highly consistent appearances of subjects. For one-stage methods [8, 56, 62], while recognizing effects after fine-tuning, they often lack consistency and plausibility, and only maintain subjects' semantic similarity. For example, AnyDoor* generates the wrong shadow direction in the first row and produces shadows that should not appear in the second row. Moreover, we precisely control the object's location and size. In the first row of Fig. 4, the size of the can is improperly influenced by the size of the clock in one-stage methods, whereas PS-Diffusion distinguishes areas that need reconstruction as background. For the two-stage approach, Remove-Insert is limited to the masked area, failing to generate effects outside masks. In addition, unsatisfactory removal in the first and third rows degrade the visual quality. In contrast, PS-Diffusion successfully edits shadows and reflections on the background while completely removing the source objects.

5.5. Ablation Results

Effectiveness of the proposed components. As illustrated in Tab. 3 and Fig. 5, we conduct an ablation analysis of each module on Replace-5K. PSNR, CLIP, DINO, LPIPS, and FID are measurements. We have the following observations: 1) In disentangled control signals, illumination-dependent properties in FECB guide the editing of lighting-related effects. BECB corrects the effect area outside the mask. Hence, they bring improvements in quantitative and qualitative results. 2) Compared to directly weighting two ControlNet branches, EAM attains better performance. As shown in the second to last column, direct weighting confuses the learning of interaction and subject. This leads to excessive alteration of original pixels in the background (first and second line) or abnormal changes in the object's appearance (second line). As shown in the last column, EAM alleviates this issue by decoupling attention of editing objects and contextual interactions, thereby improving performance. 3) By integrating all modules, the model achieves the best performance, demonstrating the complementarity of disentangled control and attention.

Table 3. Ablation on Replace-5K. When EAM is not applied, the features of BECB and the original branch are directly weighted.

FECB	BECB	EAM	PSNR↑	CLIP↑	DINO↑	LPIPS↓	FID↓
			21.480	0.935	0.905	0.190	6.250
√			21.799	0.940	0.913	0.185	5.751
√	√		22.010	0.942	0.917	0.181	5.513
√	√	√	**22.583**	**0.945**	**0.921**	**0.174**	**5.109**

Visualization of the effect map. To verify the effectiveness of BECB, we visualize the effect map. As shown in the effect map of Fig. 6, when objects are edited, potential areas where effects projected onto the background, such as shadows and reflections, are activated. In addition, we present the ablation results with and without the rough mask estimated from the source image in the LAB color space. As shown in the last two columns, due to the close relationship between the effects of the same scene on different objects, the rough mask improves the accuracy of the effect maps.

Effect attention coefficient. We conduct ablation experiments on the effectiveness of the hyperparameter ω in EAM. The quantitative and qualitative visualization results are presented in Fig. 7. The incorporation of EAM leads to a significant enhancement in the PSNR, proving its effectiveness. However, in the qualitative outcomes, it is observed that an excessively high ω gradually disrupts the original

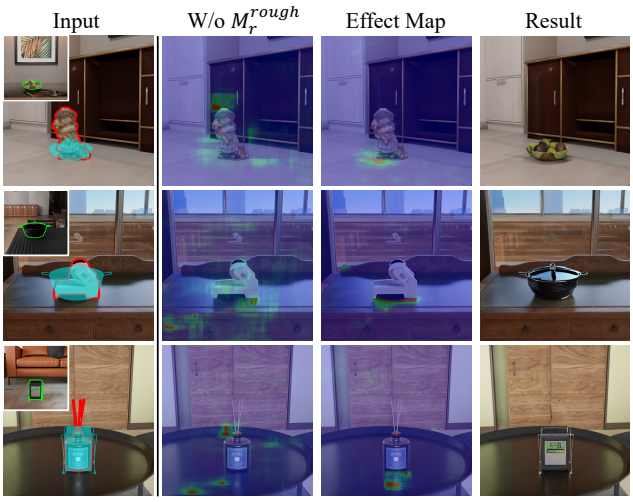

Figure 6. Visualization of the effect map in BECB. We also show the ablation results without M_e^{rough} estimated in LAB color space.

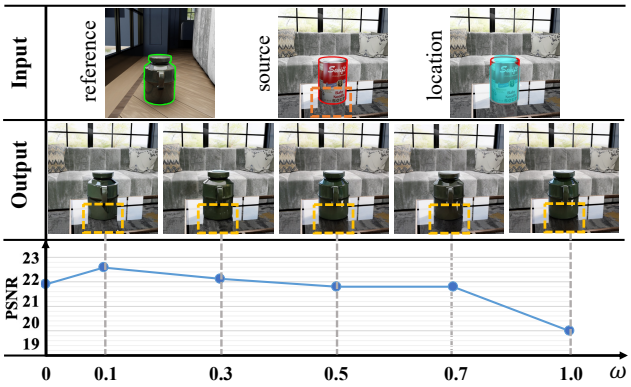

Figure 7. Ablation experiments of ω in EAM. As ω increases, the region of the effect (orange dotted frame) is implausibly deepened. PSNR is evaluated on the entire test set.

pixels of the background. This not only reduces the physical plausibility but also leads to inconsistent effects (the reflection of the object on the desktop in the source image is soft). Hence, PSNR gradually decreases. In PS-Diffusion, ω is set to 0.1 to achieve the best performance.

5.6. More Applications

The high controllability of PS-Diffusion enables precise image editing based on user-provided masks. Importantly, the adherence to physical laws enhances photorealism. Hence, PS-Diffusion can be extended to various precise and photorealistic object-level editing applications: object insertion, object movement, object scaling, and object rotation.

To be specific, for object insertion, the original object mask is identical to the target mask. In the object movement and scaling, the reference subject is the target in the source image. During object scaling, the target mask is scaled to the specified size. For object rotation, the object in the reference image is rotated utilizing some object rotation methods [25, 64]. The results are depicted in Fig. 8, demonstrating the effectiveness and scalability of PS-Diffusion.

Figure 8. The application of PS-Diffusion in other object-level editing tasks, *i.e.* object insertion, scaling, movement, and rotation.

5.7. Limitations and Future Work

To discuss limitations, we evaluate hard samples under complex lighting or specular reflections. FID on these samples is suboptimal compared to the entire dataset. Specifically, the FID of 8.781 on hard samples is higher than 5.109 on the entire dataset. Moreover, we find failure cases under strong specularity or extreme lighting. For example, the edited objects in the mirror are ambiguous or unedited. It is challenging as it involves geometry and the estimation of material and lighting. Without priors, generative models struggle to recognize reflective or diffuse objects. In the future, we will explore a more exquisite lighting estimator in FECB and the priors of diffuse and normal maps in BECB.

6. Conclusion

In this paper, we propose PS-Diffusion for photorealistic subject-driven image editing. We first disentangle the control signals of contextual interactions based on their affecting regions. To be specific, for those within the mask, we condition the diffusion model with effects estimated by intrinsic image decomposition. An additional background effect control branch identifies regions of effect outside the mask. Moreover, the effect attention module in the denoising U-Net is introduced to separate the learning of effects of contextual interactions and subject features. Extensive experiments demonstrate the effectiveness and scalability.

7. Acknowledgments

This work was supported by the Natural Science Foundation of Tianjin, China (No.24JCZXJC00040), Shenzhen Science and Technology Program (No. JCYJ20240813114229039), the National Natural Science Foundation of China (No. 623B2056, 624B2072), the Fundamental Research Funds for the Central Universities, the Supercomputing Center of Nankai University (NKSC).

References

[1] Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR (2022) 2

[2] Benenson, R., Ferrari, V.: From colouring-in to pointillism: revisiting semantic segmentation supervision. arXiv preprint arXiv:2210.14142 (2022) 5

[3] Brock, A.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018) 2

[4] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 2

[5] Cao, T., Kong, J., Zhao, X., Yao, W., Ding, J., Zhu, J., Zhang, J.D.: Product2img: Prompt-free e-commerce product background generation with diffusion model and self-improved lmm. In: ACM MM (2024) 2

[6] Careaga, C., Aksoy, Y.: Intrinsic image decomposition via ordinal shading. ACM Transactions on Graphics **43**(1), 1–24 (2023) 3, 4, 5

[7] Careaga, C., Miangoleh, S.M.H., Aksoy, Y.: Intrinsic harmonization for illumination-aware compositing. arXiv preprint arXiv:2312.03698 (2023) 4

[8] Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: CVPR (2024) 2, 3, 4, 5, 6, 7

[9] Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022) 2

[10] Czapp, Á.T., Jani, M., Domián, B., Hidasi, B.: Dynamic product image generation and recommendation at scale for personalized e-commerce. In: Proceedings of the 18th ACM Conference on Recommender Systems (2024) 2

[11] Eftekhar, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: CVPR (2021) 5

[12] Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable image generation. Advances in Neural Information Processing Systems **36**, 16222–16239 (2023) 2

[13] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 2

[14] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR (2021) 2

[15] Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: CVPR (2022) 5

[16] Gryka, M., Terry, M., Brostow, G.J.: Learning to remove soft shadows. ACM Transactions on Graphics (TOG) **34**(5), 1–15 (2015) 5

[17] Gu, J., Zhao, N., Xiong, W., Liu, Q., Zhang, Z., Zhang, H., Zhang, J., Jung, H., Wang, Y., Wang, X.E.: Swapanything: Enabling arbitrary object swapping in personalized visual editing. arXiv preprint arXiv:2404.05717 (2024) 2

[18] Gu, Y., Xu, H., Xie, Y., Song, G., Shi, Y., Chang, D., Yang, J., Luo, L.: Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In: CVPR (2024) 2

[19] Guerreiro, J.J.A., Nakazawa, M., Stenger, B.: Pct-net: Full resolution image harmonization using pixel-wise color transformations. In: CVPR (2023) 2, 6

[20] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 2

[21] Hess, R.: The essential Blender: guide to 3D creation with the open source suite Blender. No Starch Press (2007) 5

[22] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017) 6

[23] Hu, H., Chan, K.C., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al.: Instruct-imagen: Image generation with multi-modal instruction. In: CVPR (2024) 2

[24] Hua, M., Liu, J., Ding, F., Liu, W., Wu, J., He, Q.: Dreamtuner: Single image is enough for subject-driven generation. arXiv preprint arXiv:2312.13691 (2023) 2

[25] Huang, Z., Wen, H., Dong, J., Wang, Y., Li, Y., Chen, X., Cao, Y.P., Liang, D., Qiao, Y., Dai, B., et al.: Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In: CVPR (2024) 8

[26] Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. In: CVPR (2024) 2

[27] Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. arXiv preprint arXiv:2403.06976 (2024) 2, 3

[28] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 2

[29] Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: CVPR (2024) 2

[30] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: CVPR (2023) 6

[31] Kocsis, P., Philip, J., Sunkavalli, K., Nießner, M., Hold-Geoffroy, Y.: Lightit: Illumination modeling and control for diffusion models. In: CVPR (2024) 3

[32] Kulkarni, A., Tsai, E., Chen, K., Wang, Z., Cloninger, A., Saab, R.: From pixels to pictures: Understanding the internal representation of latent diffusion models 5

[33] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) 2

[34] Li, C., Yang, B., Wu, Z., Chen, G., Yu, Y., Zhou, S.: Shadow removal based on diffusion segmentation and super-resolution models. In: CVPR (2024) 3

[35] Li, T., Ku, M., Wei, C., Chen, W.: Dreamedit: Subject-driven image editing. arXiv preprint arXiv:2306.12624 (2023) 2, 3, 6

[36] Li, Z., Lv, X., Yu, W., Liu, Q., Lin, J., Zhang, S.: Face shape transfer via semantic warping. Visual Intelligence **2**(1), 26 (2024) 2

[37] Liu, C., Li, X., Ding, H.: Referring image editing: Object-level image editing via referring expressions. In: CVPR (2024) 2

[38] Liu, Q., You, J., Wang, J., Tao, X., Zhang, B., Niu, L.: Shadow generation for composite image using diffusion model. In: CVPR (2024) 3

[39] Lo, L., Yeo, C.Y., Shuai, H.H., Cheng, W.H.: Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In: CVPR (2024) 3

[40] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: CVPR (2022) 2

[41] Michel, O., Bhattad, A., VanderBilt, E., Krishna, R., Kembhavi, A., Gupta, T.: Object 3dit: Language-guided 3d-aware image editing. In: NeurIPS (2024) 2

[42] Murali, S., Govindan, V.: Shadow detection and removal from a single image using lab color space. Cybernetics and information technologies **13**(1), 95–103 (2013) 4

[43] Ning, S., Wang, D., Qin, Y., Jin, Z., Wang, B., Han, X.: Picture: Photorealistic virtual try-on from unconstrained designs. In: CVPR (2024) 2

[44] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 4, 6

[45] Pan, Y., Mao, C., Jiang, Z., Han, Z., Zhang, J.: Locate, assign, refine: Taming customized image inpainting with text-subject guidance. arXiv preprint arXiv:2403.19534 (2024) 2, 4

[46] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023) 2

[47] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2

[48] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 6

[49] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 4, 5

[50] 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) 2, 3, 6

[51] Sajnani, R., Vanbaar, J., Min, J., Katyal, K., Sridhar, S.: Geodiffuser: Geometry-based image editing with diffusion models. arXiv preprint arXiv:2404.14403 (2024) 2

[52] Sheng, Y., Liu, Y., Zhang, J., Yin, W., Oztireli, A.C., Zhang, H., Lin, Z., Shechtman, E., Benes, B.: Controllable shadow generation using pixel height maps. In: ECCV (2022) 2

[53] Sheng, Y., Zhang, J., Benes, B.: Ssn: Soft shadow network for image compositing. In: CVPR (2021)

[54] Sheng, Y., Zhang, J., Philip, J., Hold-Geoffroy, Y., Sun, X., Zhang, H., Ling, L., Benes, B.: Pixht-lab: Pixel height based light effect generation for image compositing. In: CVPR (2023) 2

[55] Song, X., Cui, J., Zhang, H., Chen, J., Hong, R., Jiang, Y.G.: Doubly abductive counterfactual inference for text-based image editing. In: CVPR (2024) 2

[56] Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: Objectstitch: Generative object compositing. arXiv preprint arXiv:2212.00932 (2022) 3, 6, 7

[57] Tarrés, G.C., Lin, Z., Zhang, Z., Zhang, J., Song, Y., Ruta, D., Gilbert, A., Collomosse, J., Kim, S.Y.: Thinking outside the bbox: Unconstrained generative

object compositing. arXiv preprint arXiv:2409.04559 (2024) 2, 3, 4
[58] Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: CVPR (2017) 2
[59] Winter, D., Cohen, M., Fruchter, S., Pritch, Y., Rav-Acha, A., Hoshen, Y.: Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. arXiv preprint arXiv:2403.18818 (2024) 2, 3, 4, 5
[60] Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3121–3138 (2022) 2
[61] Yan, Y., Zhou, Z., Wang, Z., Gao, J., Yang, X.: Dialoguenerf: Towards realistic avatar face-to-face conversation video generation. Visual Intelligence **2**(1), 24 (2024) 2
[62] 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: CVPR (2023) 2, 3, 5, 6, 7
[63] Yang, S., Chen, X., Liao, J.: Uni-paint: A unified framework for multimodal image inpainting with pre-trained diffusion model. In: ACM MM (2023) 2
[64] Yang, Y., Huang, Y., Wu, X., Guo, Y.C., Zhang, S.H., Zhao, H., He, T., Liu, X.: Dreamcomposer: Controllable 3d object generation via multi-view conditions. In: CVPR (2024) 8
[65] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 2, 6
[66] Yenphraphai, J., Pan, X., Liu, S., Panozzo, D., Xie, S.: Image sculpting: Precise object editing with 3d geometry control. In: CVPR (2024) 2
[67] Zhan, G., Zheng, C., Xie, W., Zisserman, A.: What does stable diffusion know about the 3d scene? arXiv preprint arXiv:2310.06836 (2023) 5
[68] Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W., Niu, L.: Controlcom: Controllable image composition using diffusion model. arXiv preprint arXiv:2308.10040 (2023) 2
[69] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 3
[70] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 6
[71] Zhang, X., Guo, J., Yoo, P., Matsuo, Y., Iwasawa, Y.: Paste and harmonize via denoising: Subject-driven image editing with frozen pre-trained diffusion model. In: ICASSP (2024) 3
[72] Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV (2016) 2
[73] Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&m vto: Multi-garment virtual try-on and editing. In: CVPR (2024) 2
[74] Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594 (2023) 2, 6