GeoDiffuser: Geometry-Based Image Editing with Diffusion Models

Rahul Sajnani[1,2] Jeroen Vanbaar[2] Jie Min[2] Kapil Katyal[2] Srinath Sridhar[1,2]
[1]Brown University [2]Amazon Robotics
ivl.cs.brown.edu/research/geodiffuser

Figure 1. We introduce **GeoDiffuser**, a unified method to perform common 2D and 3D image editing tasks like object translation, 3D rotation, object removal, and re-scaling while preserving object style and inpainting disoccluded regions. Our method is a zero-shot optimization-based method that builds on top of a pre-trained diffusion model. We treat image editing as a geometric transformation of parts of the image and bake this directly into a shared attention-based edit optimization. In this figure, the top row shows natural images and the bottom row shows the edit.

Abstract

The success of image generative models has enabled us to build methods that can edit images based on text or other user input. However, these methods are imprecise, require additional information, or are limited to only 2D image edits. We present GeoDiffuser, a zero-shot optimization-based method that unifies common 2D and 3D image-based object editing capabilities into a single method. Our key insight is to view image editing operations as geometric transformations. We show that these transformations can be directly incorporated into the attention layers in diffusion models to implicitly perform editing operations. Our training-free optimization method uses an objective function that seeks to preserve object style but generate plausible images, for instance with accurate lighting and shadows. It also inpaints disoccluded parts of the image where the object was originally located. Given a natural image and user input, we segment the foreground object [27] and estimate a corresponding transform which is used by our optimization approach for editing. Figure 1 shows that GeoDiffuser can perform common 2D and 3D edits like object translation, 3D rotation, and removal. We present quantitative results, including a perceptual study, that shows how our approach is better than existing methods.

1. Introduction

Image generative models have seen significant progress recently. The most advanced diffusion-based models can now generate high-quality images almost indistinguishable from reality [45, 46, 48, 60]. These models generate images with the desired content and detail by conditioning on text prompts, sometimes in combination with additional information like segmentation masks [63]. They have proliferated in use with many commercial products incorporating them [2–4].

Although realistic image generation is an important capability, in many cases, we may also want to edit generated or existing natural images. While past work relied on computer graphics techniques for image editing [10, 26, 28, 64], recent works have put generative models to use for this problem. In particular, generative models have been shown to enable text-based edits [20, 37, 56], object stitching [14, 52], object removal [46], and interactive edits using user-defined points [38, 41, 50], 3D transforms [43] or flow [17]. However, these methods have important limitations. Text-based editing methods are imprecise for edits requiring spatial control. Object stitching and removal methods cannot easily be extended to geometric edits. Finally, interactive point-/flow-based methods require additional input such as a text prompt or optical flow.

In this paper, we present **GeoDiffuser**, a method that unifies various image-based object editing capabilities into a single method. We take the view that common user-specified image editing operations can be cast as **geometric transformations** of parts of the image. For instance, 2D object translation or 3D object rotation can be represented as a bijective transformation of the foreground object. However, naively applying this transformation on the image is unlikely to produce plausible edits, for instance, due to mismatched lighting or shadows. To overcome this problem, we use diffusion models, specifically the general editing approach (see Figure 2) enabled by DDIM Inversion [36]. Our key contribution is to **bake in the geometric transformation directly within the shared attention layers of a diffusion model to preserve style** while enabling a wide range of user-specified 2D and 3D edits. Additionally, GeoDiffuser is a zero-shot optimization-based method that operates **without the need for any additional training** and can support any diffusion model with attention layers.

Figure 1 shows common image edits performed by GeoDiffuser on natural images. Without any hyperparameter tuning, our method can perform 2D edits like object translation or removal, or 3D edits like 3D rotation and translation. Given a natural image, we first segment the object of interest [27], and optionally, extract a depth map [59] for 3D edits. For each type of edit, we first compute a geometric transform based on user input and formulate an objective function for optimization. Unlike approaches that first 'lift' an object from an image and then stitch the transformed object back into the image [26], we implicitly perform these steps by applying the transform directly to the self- and cross-attention layers. Since attention captures both local and global image interactions, our results exhibit accurate lighting, shadows and reflection while inpainting the disoccluded image regions. Moreover, our objective function incorporates terms to preserve the original style of the transformed object.

We show extensive qualitative results that demonstrate that our method can perform multiple 2D and 3D editing operations using a single approach. To evaluate our method quantitatively, we provide experiments through a perceptual study as well as metrics that measure how well the foreground and background content is preserved during the edit. Results show that our method outperforms existing methods quantitatively while being general enough to perform various kinds of edits. To sum up, our main contributions are:

- A unified image editing approach that formulates common 2D and 3D editing operations as geometric transformations of parts of the image.
- GeoDiffuser, a zero-shot optimization-based approach that incorporates geometric transforms directly within the attention layers of diffusion models enabling realistic edits while preserving object style.
- Qualitative results of 2D and 3D object edits enabled by our method without model fine-tuning (see Fig. 1).

2. Related Work

Image editing has been widely studied in computer vision and encapsulates a range of operations, such as object removal and addition [6, 52], style transfer [18, 21, 23, 25], and 2D and 3D transforms [26]. One challenge with this problem is to keep the edit consistent within the *global* context of the image. Traditional methods such as Poisson image editing [44] use gradients of the context to blend edits with existing pixels, while inpainting methods uses boundary and context to fill in pixels [58]. We discuss generative model-based and 3D-aware editing methods below.

Text-Guided Image Editing: There are several works using generative image models to edit images via changes to the text prompt. The preservation of subject identity in different settings can be achieved by textual inversion along with additional losses to finetune the generative model [47]. *Null-text* inversion is an inversion approach where a null-text embedding is optimized to match an inverted noise trajectory for a given input image along with attention reweighting [36]. Instead of an inversion process, text prompt edits can also be achieved by swap, or re-weighting of cross-attention maps derived from the visual and textual representation [20]. Edits with text prompts can also be achieved by using cross-attention from different prompts to manipulate self-attention maps [9]. Leveraging existing text-to-image models along with [8] gives the ability to generate paired data for finetuning a generative model to achieve text-guided editing results. These methods mostly produce images with style changes or enhancements, or object replacement. [15] leverage prompts and self guidance to perform 2D image edits of scaling and translation. However, it is difficult to guide the diffusion model to perform a specific 3D geometric transform based on a prompt. We extend the above approaches to build a method to handle geometric transforms without any additional training.

Non-Text-Guided Image Editing: Text-guided edits are mostly limited to appearance and style changes. Non-text-guided edits on the other hand, can achieve a variety of edits. Point-based editing approaches can perform local image edits. [50] propose a motion supervised latent optimization between the reference and target edit, to guide the denoising to obtain the edit while preserving the object identity. Stroke-based editing can edit larger image regions, or even entire images [34], by projecting strokes onto the image manifold via diffusion. For these methods, edits such as translations are however not possible. ObjectStitch [52] along with inpainting can achieve translation where the denoising diffusion is applied to a target asked region, and guided by the embedding of the object to stitch. However, object style preservation is difficult in this setting. Recent methods [38, 39] try to preserve identity and allow for translations while requiring no training. However, these are limited to 2D translations and scaling. An editing approach which first 'lifts' the object from a background is proposed in [43]. The background is inpainted and a depth-to-image generative model is used, which performs the denoising conditioned on an input depth. However, this approach needs an additional text prompt while ours does not. Additionally, we support various kinds of edits and not just 3D transforms. [17] uses flow-guidance for image editing. However, optical flow can be much harder to obtain compared to depth [59]. We present a method that performs 2D and 3D edits using precise geometric transformations while preserving identity and not requiring additional user input.

3D-Aware Editing: Some methods have addressed the 3D editing problem [26] by 'lifting' objects into 3D and use 3D meshes and scene illumination to allow for proper blending of the edited object with the existing image context. Other methods use NeRF [12,19,57,61,62] or works [30,31] learn over large-scale datasets [11], leverage geometry representations to perform edits but require multi-view images that are difficult to obtain. Edits are also directly applied to generative models, *e.g.*, [42] propose a point-based edit along with motion supervision to guide the neighboring pixels. The authors of [40] propose to represent foreground objects and background as neural feature fields, which can be edited and composited for a final output. The method of [29] addresses limitations of point-based editing in GANs, using template features rather than points for better tracking, and restricting search area around pixels to lines.

Concurrent Works: Concurrent works Instadrag [49] & MagicFixup [5] perform drag-based edits by tedious training over large-scale video data, but they don't bake in geometry of the object within the architecture. However, they may allow better inpainting and novel view synthesis of objects (discussion in supplement).

3. Background

Denoising Diffusion: We first briefly describe the concept of Denoising Diffusion Probabilistic Models (DDPM) used successfully by diffusion models for image generation [46]. Images can be considered as samples drawn from a data distribution $q(x)$. Denoising diffusion aims to learn a parameterized model $p_\theta(x)$ that approximates $q(x)$ and from which new samples can be drawn. The (forward) diffusion process iteratively adds noise to an input image x_0, with $t = 0$, according to either a fixed or learned schedule, represented by α_t with $t \in [1, T]$. At each timestep, the latent encoding is performed according to a Gaussian distribution centered around the output of the previous timestep: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha} x_{t-1}, (1-\alpha_t)\mathbf{I})$. The parameters vary over time such that $p_\theta(x_T) := \mathcal{N}(\mathbf{0}, \mathbf{I})$. Using the reparameterization trick, the noised version of input x_0 can directly be expressed as: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_0$.

The reverse process, where noise is gradually removed at each step, can be expressed as the joint distribution $p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta^{(t)}(x_{t-1}|x_t)$. Under the assumption of trainable means and fixed variances, a neural network $\hat{\epsilon}_\theta(x_t, t)$ can be trained with the objective of minimizing a variational lower bound to estimate the source noise $\epsilon_0 \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$ that determines x_t from x_0: $L_\gamma(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\| \epsilon_0 - \hat{\epsilon}_\theta(x_t, t) \|_2^2 \right]$. For more details see [33].

Conditioning and Efficiency: This formulation can be extended to the conditional case, *i.e.*, $p_\theta(x_{0:T}|y)$. The condition y could be images, (encoded) text, or something else. The computational bottleneck is the number of denoising timesteps T, however a non-Markovian variant Denoising Diffusion Implicit Models (DDIM) was introduced to reduce the number of timesteps [51]. To further reduce the computational burden, the diffusion process for images can be performed in a lower dimensional latent (feature) space, as proposed by [46]. A perceptually optimized pretrained decoder takes the latent x_1, and reconstructs the image x_0. In our work, we use a latent diffusion model together with Classifier-Free Guidance (CFG) [22] for text conditioning.

Attention: Attention was introduced as an alternative to recurrent networks and large receptive fields in convolution-based neural networks, for capturing local and global context [13, 55]. The scaled dot-product self-attention mechanism adopted in transformers has found widespread application in computer vision applications. The input is a tuple [Query (Q), Key (K), Value (V)], each with learnable parameters via linear layers. An attention layer constructs an attention map $\text{AM}(Q, K) := \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$ and then computes attention as: $\text{Attention}(Q, K, V) := \text{AM}(Q, K)V$. Here, d is the dimension of the embedding.

In addition to self-attention, the query can be derived from another input source, *e.g.*, another modality, and using the key and values from the first input, the cross-attention between the two inputs can be computed via Section 3 and Section 3. An example of cross-attention is the activation of a word in a sentence with pixels in an image.

The correlation between semantics and pixels for image-text cross-attention can be modified in the denoising diffusion generative image setting to adjust the appearance of a given generated image [20]. In addition, deriving masks from cross-attention to guide self-attention [9] provides the ability to change the appearance of objects while maintaining object identity.

General Editing Framework: Prior works leverage the learned capabilities of diffusion models to perform edits to a given image, rather than a generated one. A general framework (see Figure 2) that is followed in all these works is to first perform an inversion [36, 51] on the image. This inversion provides us with a noise latent that sets a good starting point to regenerate the input image as well as to

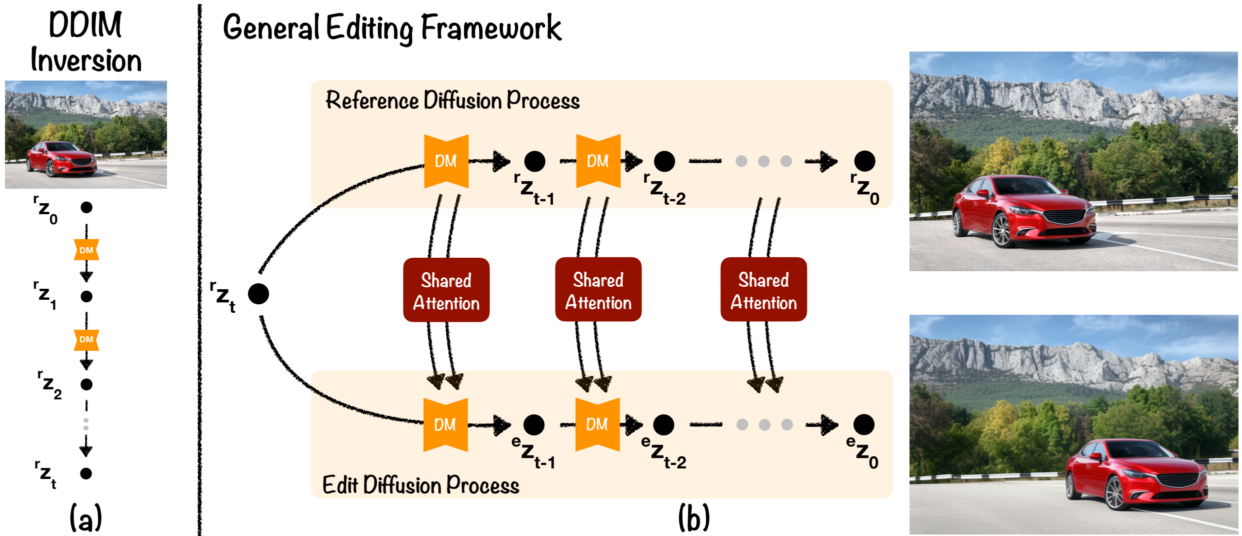

Figure 2. General image editing framework using diffusion models. **(a) DDIM Inversion:** The process of obtaining noise trajectory $\{{}^{r}z_0, {}^{r}z_1, ..., {}^{r}z_t\}$ for the reference image [51]. **(b) General Editing Framework:** The Reference Diffusion Process guides the Edit Diffusion Process to achieve the desired edit. In **GeoDiffuser**, we perform *geometric* 2D and 3D edits by transforming the shared attention layers leading to plausible edits that preserves object style, inpainting disoccluded background, and adding details (*e.g.*, the car's shadow).

edit it. Starting from the inverted noise latent, two parallel diffusion processes generate the input image as well as the edited image. The first **reference diffusion process** generates the original input and, in our work, helps preserve un-edited regions of the image. An **edit diffusion process** runs in parallel that utilizes the attention blocks from the reference process to perform the desired edit. This **shared attention** is a key insight for our proposed work. The editing framework is sketched in Figure 2 (b).

4. GeoDiffuser

The goal of GeoDiffuser is to enable editing of segmented foreground objects in either natural or generated images. We take the view that common editing operations like 2D translation, 3D object rotation or object removal can be expressed as geometric transformations of parts of the image. Naively applying this transform to segmented foreground objects typically produces poor results w.r.t. image context and does not fill in the disoccluded background. We propose to use diffusion models to realistically edit the image and preserve object style.

Supported Operations: In this paper, we focus on *geometric edits* to an image $\mathcal{I}$ specified by users through sliders that control transformations of foreground objects. In particular, we unify three kinds of edit operations that previously required separate bespoke methods: (1) **2D object edit** operations deal with realistically translating or scaling segmented objects within the image including inpainting the background where the object was originally located. (2) **3D object edit** operations deal with realistically transforming objects based on user-specified 3D rotation, translation or scaling and inpainting any disoccluded background as a result of the edit. Finally, (3) **object removal** refers to the operation of removing the segmented object completely and inpainting the disoccluded background.

In contrast with previous approaches, we formulate edits as an optimization problem based on the shared attention and leverage a pre-trained text-to-image Stable Diffusion model [46] to perform the edit. Notably, our method requires no training and can use any diffusion model with attention. Given an image $\mathcal{I}$, an object mask M_{obj}, a user-specified 2D or 3D transformation T, our goal is to edit the object in the image and inpaint any disoccluded regions introduced by the edit. To compute T for 3D edits, we use a depth map D obtained from DepthAnything [59] or simply by setting a constant depth of 0.5 m. This enables us to edit in-the-wild natural images without additional user input.

4.1. Edits via Shared Attention

Each edit operation begins by performing a DDIM inversion [51] on the given image (Figure 2 (a)). Inverting the image provides us with the latent noise trajectory that will guide the edit diffusion process. We then perform the reverse diffusion process along with the geometry-aware attention sharing mechanism as sketched in Figure 2 (b). This attention sharing mechanism along with optimizing for the image latents as well as text embeddings is the key to achieve the desired geometric edit. Figure 3 (a) depicts the process for the shared attention blocks from Figure 2 (b).

Image Inversion: For inversion, we use direct inversion [24] on the image $\mathcal{I}$ with the null prompt "". Direct inversion initializes the reference trajectory with the noising trajectory for fast and accurate reconstruction of the reference image without the need for optimizing embeddings (null-text [36]) & model weights (LoRA [50]). This inversion provides us with latents $\{{}^{\mathrm{r}}z_t, {}^{\mathrm{r}}z_{t-1} ... {}^{\mathrm{r}}z_0\}$ that preserves

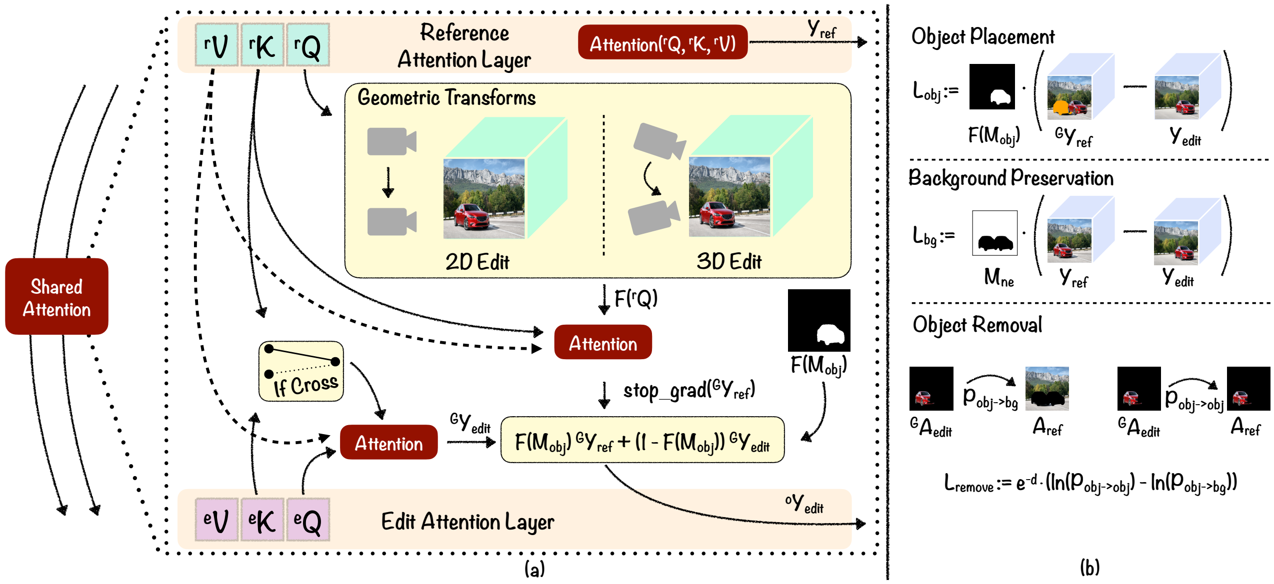

Figure 3. **(a) GeoDiffuser** attention sharing mechanism that leverages the geometric transformation $\mathcal{F}(\cdot)$ transform the reference attention ${}^{\mathcal{G}}Y_{ref}$ to guide the edit attention layer. **(b) Optimization Loss Functions** that penalize the latents and text-embeddings to perform the desired geometric edit. The orange mask highlights the region to be inpainted in the optimization.

the style of the image and guides the edit.

2D Edits: GeoDiffuser can perform 2D edits without requiring a depth map. Through a user interface, we can obtain a transformation T corresponding to a desired 2D translation or scaling. We define a 2D signal $S : [0,1]^2 \rightarrow \mathbb{R}^C$ that stores a per-pixel feature in the image. The signal S can represent the RGB values or even the features of a deep network defined at each coordinate. Given a per-pixel edit $\mathcal{F}$ defined on S, our shared attention mechanism uses $\mathcal{F}$ to transform this signal for the desired edit. In our case, this signal is the Query embedding of the attention layer.

3D Edits: 2D edits are limited as they do not leverage the geometry of objects. We can extend 2D edits to 3D by additionally incorporating depth information D monocular depth estimators [7,59] or simply a constant billboard depth map. The user specifies a 3D rigid transformation T which can then be used to compute the per-pixel edit $\mathcal{F}$ as

$$\mathcal{F}(S)[u] := S[PTD[u]P^{-1}u].$$

Here, P is the camera intrinsic matrix that is used to project points in the image and u is the coordinate location of the signal. This edit field $\mathcal{F}$ captures the 3D shape of the visible region of the object and provides an estimate of the desired location of the object. Note that if the per-pixel edit field is known, *e.g.*, from optical flow, we do not need a depth map for guidance. However, optical flow is much more challenging to obtain for a single image compared to depth maps.

Object Removal: Object removal introduces disocclusions to the background where the object was originally located. We propose an additional loss (see Section 4.2) for the optimization of the diffusion latents to handle such disocclusions. Disocclusions can also occur for 2D and 3D edits, and we consider such edits to be composites of removal and placement operations. Our proposed formulation for latent optimization thus extends to those edits as well.

Shared Attention: A key insight of our work is that we can transform objects by merely applying the edit $\mathcal{F}$ to the query embeddings of the reference attention (Figure 3 (a)). Let ${}^{\text{r}}Q, {}^{\text{r}}K, {}^{\text{r}}V$ be the queries, keys, and values within the diffusion model of the reference denoising process and ${}^{\text{e}}Q, {}^{\text{e}}K, {}^{\text{e}}V$ be the queries, keys, and values of the corresponding attention block in the edit denoising process. The reference attention guidance ${}^{\mathcal{G}}Y_{\text{ref}}$ and edit attention guidance ${}^{\mathcal{G}}Y_{\text{edit}}$ are then given by

$${}^{\mathcal{G}}Y_{\text{ref}} := \text{Attention}(\mathcal{F}({}^{\text{r}}Q), {}^{\text{r}}K, {}^{\text{r}}V) \quad (1)$$

$${}^{\mathcal{G}}Y_{\text{edit}} := \begin{cases} \text{Attention}({}^{\text{e}}Q, {}^{\text{r}}K, {}^{\text{r}}V), \text{if SelfAttention} \\ \text{Attention}({}^{\text{e}}Q, {}^{\text{e}}K, {}^{\text{r}}V), \text{otherwise} \end{cases} \quad (2)$$

Applying transformation $\mathcal{F}$ only to the reference query embeddings ${}^{r}Q$ followed by dot product with the reference key embeddings ${}^{r}K$ in Eq. 2 provides us with correspondences between them. This edited attention map attends to the reference value embeddings and ensures that the transform only changes the geometry and preserves the appearance. We use the edit key embeddings ${}^{e}K$ in cross attention map to enable the flow of gradients to the null-initialized text embeddings for the optimizing the edit trajectory. To place the object at the desired location, the edit and reference attention guidance should approximately be the same (${}^{\mathcal{G}}Y_{\text{ref}} \approx {}^{\mathcal{G}}Y_{\text{edit}}$) for the foreground. Note that they need not be exactly the same in the case of an ill-defined edit $\mathcal{F}$. We then transform the output ${}^{\mathcal{O}}Y_{\text{edit}}$ of the edit attention layer

$${}^{\mathcal{O}}Y_{\text{edit}} := \mathcal{F}(M_{obj}) \cdot {}^{\mathcal{G}}Y_{\text{ref}} + (1 - \mathcal{F}(M_{obj})) \cdot {}^{\mathcal{G}}Y_{\text{edit}}, \quad (3)$$

where $\mathcal{F}(M_{obj})$ refers to the foreground mask after applying the transformation $\mathcal{F}$. In other words, Eq. 3 aim to preserve identity for the object in the edit at its target location, while simultaneously preserve identity and consistency for the remaining pixels (or background). See supplement for algorithm and details.

4.2. Optimization

GeoDiffuser is a zero-shot optimization-based method that operates **without the need for any additional training**. We achieve this via optimization of the latents and null-initialized text embeddings for edit guidance. The shared attention guidance provides us with a proxy of where the foreground object must be placed after the edit. However, it does not guide the inpainting of the disocclusions introduced by moving the object causing duplications. We formulate an optimization procedure to fill the disocclusions and penalize the deviation of the edit attention guidance from the reference attention guidance. The loss functions used in the optimization (shown in Fig. 3 (b)) are explained in detail next.

Background Preservation Loss: Performing shared attention guidance along with optimization could result in the un-edited regions of the image to also be changed. We introduce a background preservation loss to prevent this. Let the mask M_{ne} represent the non-editable region of the image. We define the background preservation loss as

$$\mathcal{L}_{bg} := \text{mean}(M_{\text{ne}} \cdot ||^{\mathcal{G}}Y_{\text{edit}} - Y_{\text{ref}}||_1). \tag{4}$$

Here, $Y_{\text{ref}} = \text{Attention}({}^{\text{r}}Q, {}^{\text{r}}K, {}^{\text{r}}V)$ is the attention block output for the reference de-noising process. The reference attention preserves the style of the image and constrains the optimization towards preserving the background.

Object Placement Loss: Occasionally, the optimization changes the foreground region of the image. This causes loss of detail in the foreground. To prevent this, we penalize the deviation between the edit guidance and the reference guidance within the transformed foreground mask by

$$\mathcal{L}_{obj} := \text{mean}(\mathcal{F}(M_{\text{obj}}) \cdot ||^{\mathcal{G}}Y_{\text{edit}} - {}^{G}Y_{\text{ref}}||_1). \tag{5}$$

Note, we don't use this loss for object removal.

Inpainting Loss: To inpaint the disoccluded regions of the image, we maximize the difference between the edit guidance attention map ${}^{\mathcal{G}}A_{\text{edit}} := \text{AM}({}^{\text{e}}Q, {}^{\text{r}}K)$ and the reference guidance attention map $A_{\text{ref}} := \text{AM}({}^{\text{r}}Q, {}^{\text{r}}K)$. Let $\rho_{\text{obj}\rightarrow\text{bg}}$ represent the maximum normalized correlation score for each row in the foreground mask of the attention map ${}^{\mathcal{G}}A_{\text{edit}}$ to each row in the background mask of the reference attention map A_{ref}. We can similarly compute $\rho_{\text{obj}\rightarrow\text{obj}}$ that provides us with the maximum foreground to foreground normalized correlation (see Figure 3 (b)). Our goal is to reduce $\rho_{\text{obj}\rightarrow\text{obj}}$ and increase $\rho_{\text{obj}\rightarrow\text{bg}}$. We want to inject the disoccluded region with features from the background and ensure that the diffusion process doesn't in-paint the same features. We penalize for this using

$$\mathcal{L}_{remove} := \text{mean}\left(e^{-d_{\text{obj}\rightarrow\text{bg}}}(ln(\rho_{\text{obj}\rightarrow\text{obj}}) - ln(\rho_{\text{obj}\rightarrow\text{bg}}))\right). \tag{6}$$

Here, $d_{\text{obj}\rightarrow\text{bg}}$ is the coordinate distance between the locations of the attention map. The loss weighted by coordinate distance ensures that the foreground region inpaints the region using features within its vicinity. The negative log forces the object to background correlation $\rho_{\text{obj}\rightarrow\text{bg}}$ to increase and also reduces object-object correlation forcing the inpainted region to not be filled by the same object.

Smoothness Constraint: We additionally penalize the edit attention guidance ${}^{\mathcal{G}}Y_{\text{edit}}$ for smoothness by penalizing the absolute value of its gradients using $\mathcal{L}_s$.

Geometry Editing Optimization: We edit images by penalizing the null-initialized text embeddings and images latents during generation using the final loss $\mathcal{L} := w_{bg}\mathcal{L}_{bg} + w_{obj}\mathcal{L}_{obj} + w_r\mathcal{L}_{remove} + w_s\mathcal{L}_s$.

In our experiments, we found that the inpainting loss $\mathcal{L}_{remove}$ is hard to optimize and changes every image differently. To combat this, we devise an adaptive optimization scheme that increases the weight w_r of the removal loss if the loss is more than -1.8 and reduce the loss weight if the removal loss is lower than -6. All our experiments are performed on an Nvidia RTX3090 with a run time of 30 seconds (for removal) up to 45 seconds (for 2D & 3D edits) on image resolution of 512. We also penalize depth smearing artefacts of the foreground using amodal loss $\mathcal{L}_{amodal}$. See supplement for algorithm & more details.

5. Results & Experiments

In this section, we present visual examples of our editing results and quantitative results of visual metrics of editing quality and a perceptual study.

Dataset: To measure the efficacy of our method we collected a dataset of real images from Adobe Stock images [1] to ensure we exclude generative AI images. We collect 70 images corresponding to the prompts *dog, car, cat, bear, mug, lamp, boat, plane, living-room, peaceful scenery*. We also test on real in-the-wild images from [16] and generated images from [43]. For many images in our dataset, we show multiple 2D and 3D edits demonstrating the general editing capabilities of GeoDiffuser.

Baselines: Since there is no extant method that performs all types of edits that we support, we compare each edit type to a different baseline. For the object removal, we compare with a state-of-the-art off-the-shelf **LaMa** image inpainting model [53], dilating object masks to make LaMa work better. For the 3D edit operations, we benchmark ourselves against **Zero123-XL** [31], **FreeDrag** [29], **DragonDiffusion** [38, 39], **Diffusion Self Guidance** [15] and **Diffusion Handles** [43]. Please see supplement for more details about each baseline. Moreover, our tests with **Ob-**

	MD ↓	Warp Error ↓	Clip Similarity ↑
3D Edits			
Diffusion Self Guidance [15]	92.067	0.243	0.809
Dragon Diffusion [38, 39]	66.108	0.226	0.953
FreeDrag [29]	31.451	0.182	**0.977**
Zero123-XL + Lama [30, 53]	19.010	0.157	0.961
Diffusion Handles [43]	10.837	0.114	0.890
GeoDiffuser (Ours)	**7.304**	**0.091**	0.967
2D Edits			
Diffusion Self Guidance [15]	155.149	0.297	0.806
FreeDrag [29]	64.716	0.259	0.962
Zero123-XL + Lama [30, 53]	20.000	0.135	0.929
Dragon Diffusion [38, 39]	38.070	0.151	0.957
GeoDiffuser (Ours)	**5.579**	**0.098**	**0.963**

Table 1. Our method adheres to the desired edit having the least **Mean Distance** and **Warp Error** compared to Dragon Diffusion, FreeDrag, Diffusion Self Guidance, and Diffusion Handles.

ject 3DIT [35] on real images produced poor results so we exclude it. For 2D edits, we compare with **Dragon Diffusion** [38, 39]. Since above methods require prompts while our method does not, we manually added text descriptions to our data. We benchmark our method against baselines using community accepted metrics of Mean Distance, Clip Similarity Score, and Warp Error. We additionally test our method for inpainting and editing with a perceptual study.

5.1. Quantitative Evaluation

Metrics: We detail the metrics here to evaluate our edits quantitatively against baselines for edit adherence and style preservation. We performed a total of 102 edits on our dataset: 36 2D edits and 66 3D edits. Metrics such as FID and Image Fidelity (IF) [38, 50] are not suitable for evaluating geometric edits because there could be large visual difference (e.g., large translation) and they do not measure disocclusion inpainting quality.

Therefore, we use three other metrics to better evaluate methods: (1) The **Mean Distance (MD)** metric computes interest points on the foreground of the image using SIFT [32] and finds correspondences between the input and edited image using DiFT [54]. We then measure the distance between the correspondence estimated by DiFT and the edit specified by the user. This metric measures how well each approach transforms the foreground object. (2) the **Warp Error (WE)** metric forward warps the foreground region of the input image to the edited image and compute the absolute difference between their pixels for the transformed foreground. This metric measures how well each approach adheres to the edit. (3) the **CLIP Similarity (CS)** metric computes the CLIP image embedding [57] of the input and edited image and measures the cosine similarity. A good editing approach preserves the image context with high **CS** and adheres to the edit with low **WE** & **MD**.

Table 1 shows quantitative comparison for 2D and 3D edits of our method with the baselines. GeoDiffuser (**MD(2D):** 5.579 & **MD(3D)**: 7.304) outperforms baselines FreeDrag, Diffusion Handles, Zero123-XL, Dragon Diffusion, and Diffusion Self Guidance in **MD** metrics and Warp Error for both 2D and 3D edits. Dragon Diffusion does not perform well in these tasks since their method fails to inpaint disocclusions or preserve the foreground. Zero123-XL baseline performs better but since it is not trained on real-world images, it does not preserve the foreground object resulting in incorrect DiFT correspondences. All methods seem to preserve the context of the scene with a clip score above 0.920 apart from Diffusion Handles and Diffusion Self Guidance that struggle to preserve the image style in 3D edits with clip score of 0.890 & 0.809 respectively. But Diffusion Handles better adheres to the edit as it uses the depth to project activation functions of the SD model. Diffusion Self Guidance consistently underperforms as it most often does not move objects and does not preserve the object appearance. For 3D edits, FreeDrag has a marginally high clip similarity score of 0.977 compared to ours (**CS:** 0.967). However, at times **CS** is higher when the foreground is not removed appropriately. FreeDrag struggles to optimize for large edits and occasionally produces improper object inpainting. See Figure 4 and supplement for visual comparison and a detailed implementation, timing, and performance analysis for all baselines.

Perceptual Study: In addition to quantitative evaluations, we perform a perceptual study with 53 participants to compare our inpainting and editing results against prior works. This was setup as a forced choice questionnaire where participants had to select one of two options as containing the best edit result. Of the two randomly presented options, one was ours and the other was a baseline. Participants preferred our inpainting over LaMa [53] 94.06% of the time. They also preferred our geometric edits over Zero123-XL 86.48% of the time for realism and 88.48% of the time for edit adherance. See supplement for more information.

	MD ↓	Warp Error ↓	Clip Similarity ↑
Timesteps for Geometric Attention Sharing (Geometry Guidance)			
t=30	8.363	0.0998	0.872
t=37	7.158	0.0950	0.932
GeoDiffuser (t = 45)	**6.785**	**0.0934**	**0.966**
Adaptive Optimization			
w/o Adaptive Optimization	9.164	0.0944	**0.966**
GeoDiffuser (with Adaptive Optimization)	**6.785**	**0.0934**	**0.966**
Loss Functions			
w/o Background Preservation Loss	**6.736**	0.0958	0.954
w/o Removal Loss	57.600	0.0941	0.965
w/o Object Placement Loss	7.397	0.0986	0.963
GeoDiffuser	6.785	**0.0934**	**0.966**

Table 2. **Metric Ablations**: Increasing the number of time steps for geometric attention sharing and adaptive optimization both improve the Mean Distance, Warp Error, and Clip Similarity score. Removing removal loss introduces duplication of objects and removing background preservation changes the scene background.

Ablations: We present quantitative ablations of our design choices in Table 2. Increasing the number of time steps

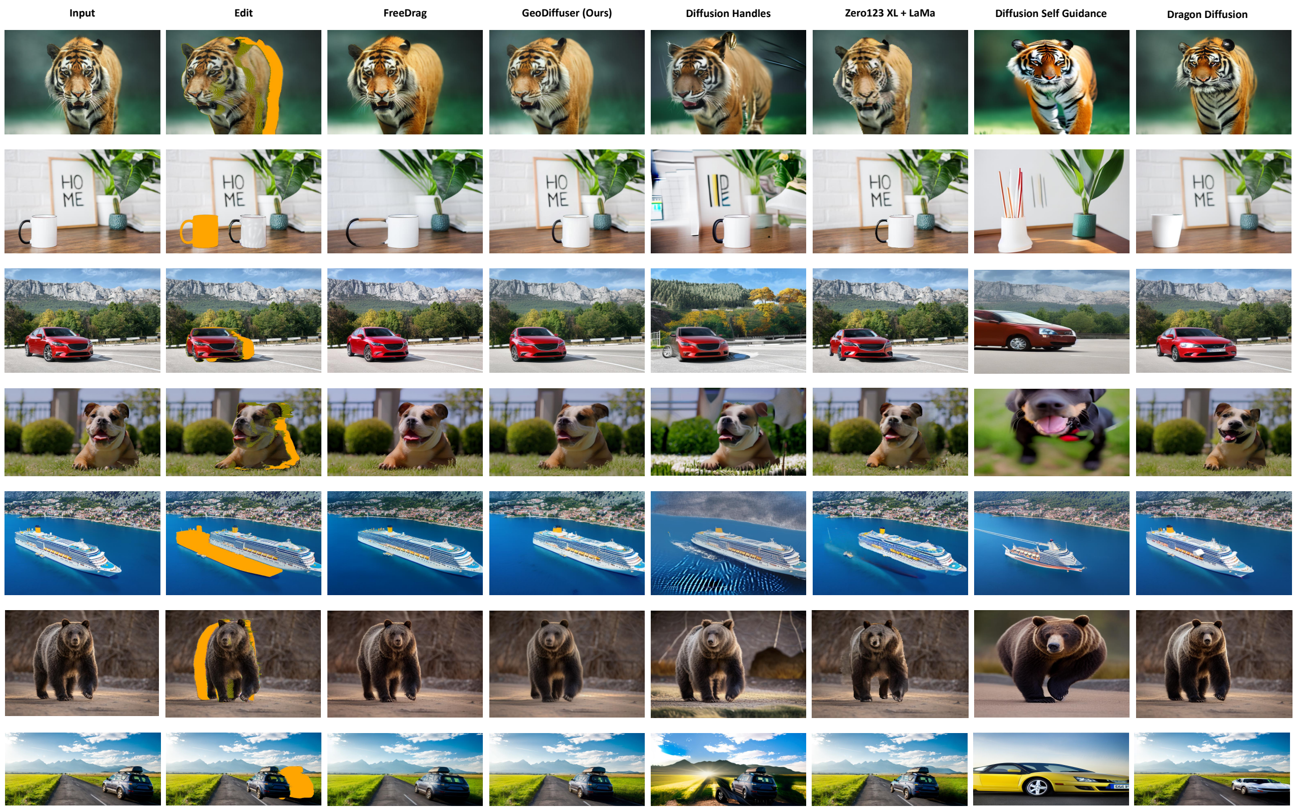

Figure 4. We perform the same edit using prior works and compare with out work. We show the intended 3D edit in column 2 where we highlight the region to be inpainted with orange and the region foreground inpainting region with green. Our work GeoDiffuser best adheres to the intended edit and ensures preservation of the scene without requiring prompts. Diffusion Handles requires an inpainting model and a depth trained diffusion model to perform the same edit with prompts but still fails to preserve the appearance of the scene. FreeDrag is slow and does not adhere well to the edit. Dragon Diffusion and Diffusion Self Guidance do not preserve the appearance of the object and do not rotate objects accurately. Please see supplement for a detailed analysis of all prior works.

for geometric attention sharing provides geometric guidance for more accurate edits with lower **MD** and **WE** (Table 2, Figure 5). Without adaptive optimization, we need image specific tuning for loss weights which is not scalable. Removing placement loss reduces the foreground edit accuracy increasing the **MD** and **WE**. Background preservation loss improves scene preservation with improved global consistency and high **CS**. Without removal loss there exist duplicates within the edited image that lead to incorrect correspondences while computing **MD** resulting in very high errors. Please see supplement for more visual ablations.

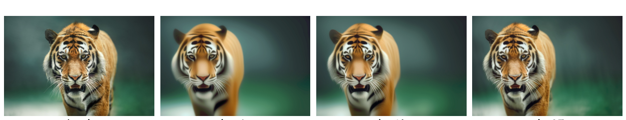

Figure 5. **Geometry Guidance**: Increasing steps t for geometric attention sharing better preserves object style (translation edit).

Qualitative Results: We show more qualitative comparisons of 3D edits performed by GeoDiffuser against baselines in Figure 4 and supplement. Note how GeoDiffuser not only removes / transforms objects but also their reflection and shadows.

6. Conclusion

GeoDiffuser is a unified method to perform common 2D and 3D object edits on images. Our approach is a zero-shot optimization-based method that uses diffusion models to achieve these edits. The key insight is to formulate image editing as a geometric transformation and incorporate it directly within the shared attention layers in a diffusion model-based editing framework. Results show that our single approach can handle a wide variety of image editing operations, producing better results compared to prior work.

Limitations & Future Work: While we can handle background disocclusions, we cannot yet handle foreground object disocclusions resulting from large 3D rotations that requires accurate novel view synthesis of in-the-wild objects which is a very difficult problem. Our method also occasionally generates artifacts due to downsampled attention masks and is limited by the capabilities of the base diffusion model (see supplement for details). We plan to address these limitations in future work.

Acknowledgements: Part of this work was done during Rahul's internship at Amazon. This work was additionally supported by NSF grant CNS #2038897, ONR grant N00014-22-1-259, and an AWS Cloud Credits award.

References

[1] Adobe stock. https://stock.adobe.com/. 6

[2] Amazon titan image generator, multimodal embeddings, and text models are now available in amazon bedrock — aws news blog. https://aws.amazon.com/blogs/aws/amazon-titan-image-generator-multimodal-embeddings-and-text-models-are-now-available-in-amazon-bedrock/. (Accessed on 03/04/2024). 1

[3] Dall·e 2. https://openai.com/dall-e-2. (Accessed on 03/04/2024). 1

[4] Gemini - chat to supercharge your ideas. https://gemini.google.com/. (Accessed on 03/04/2024). 1

[5] Hadi Alzayer, Zhihao Xia, Xuaner Zhang, Eli Shechtman, Jia-Bin Huang, and Michael Gharbi. Magic fixup: Streamlining photo editing by watching dynamic videos. *ArXiv*, abs/2403.13044, 2024. 3

[6] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 5

[8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[9] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing, 2023. 2, 3

[10] Tao Chen, Zhe Zhu, Ariel Shamir, Shi-Min Hu, and Daniel Cohen-Or. 3-sweep: Extracting editable objects from a single photo. *ACM Transactions on graphics (TOG)*, 32(6):1–10, 2013. 1

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3

[12] Jiahua Dong and Yu-Xiong Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[14] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 1

[15] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6, 7

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[17] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. *arXiv preprint arXiv:2401.18085*, 2024. 1, 3

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 3

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 3

[21] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*, 2023. 2

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3

[23] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 2

[24] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 4

[25] 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

[26] Natasha Kholgade, Tomas Simon, Alexei Efros, and Yaser Sheikh. 3d object manipulation in a single photograph using stock 3d models. *ACM Transactions on graphics (TOG)*, 33(4):1–12, 2014. 1, 2, 3

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2

[28] Jean-François Lalonde, Derek Hoiem, Alexei A Efros, Carsten Rother, John Winn, and Antonio Criminisi. Photo clip art. *ACM transactions on graphics (TOG)*, 26(3):3–es, 2007. 1

[29] Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, Yi Jin, and Jinjin Zheng. Freedrag: Feature dragging for reliable point-based image editing. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 6860–6870, June 2024. 3, 6, 7

[30] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3, 7

[31] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3, 6

[32] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004. 7

[33] Calvin Luo. Understanding diffusion models: A unified perspective, 2022. 3

[34] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[35] Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. 2, 3, 4

[37] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1

[38] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models, 2023. 1, 2, 6, 7

[39] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. *arXiv preprint arXiv:2402.02583*, 2024. 2, 6, 7

[40] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[41] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1

[42] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3

[43] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7695–7704, June 2024. 1, 3, 6, 7

[44] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, jul 2003. 2

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

[46] 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 (CVPR)*, pages 10684–10695, June 2022. 1, 3, 4

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022. 2

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[49] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent Y. F. Tan, and Jiashi Feng. Instadrag: Lightning fast and accurate drag-based image editing emerging from videos. *ArXiv*, abs/2405.13722, 2024. 3

[50] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 1, 2, 4, 7

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 4

[52] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 1, 2

[53] 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. 6, 7

[54] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 7

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Infor-*

mation Processing Systems, volume 30. Curran Associates, Inc., 2017. 3

[56] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13, 2023. 1

[57] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3, 7

[58] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[59] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 2, 3, 4, 5

[60] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2023. 1

[61] Lu Yu, Wei Xiang, and Kang Han. Edit-diffnerf: Editing 3d neural radiance fields using 2d diffusion model. *arXiv preprint arXiv:2306.09551*, 2023. 3

[62] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 3

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[64] Youyi Zheng, Xiang Chen, Ming-Ming Cheng, Kun Zhou, Shi-Min Hu, and Niloy J Mitra. Interactive images: Cuboid proxies for smart image manipulation. *ACM Trans. Graph.*, 31(4):99–1, 2012. 1