Customizing Text-to-Image Diffusion with Object Viewpoint Control

NUPUR KUMARI[*], Carnegie Mellon Uniersity, United States of America
GRACE SU[*], Carnegie Mellon Uniersity, United States of America
RICHARD ZHANG, Adobe Research, United States of America
TAESUNG PARK, Adobe Research, United States of America
ELI SHECHTMAN, Adobe Research, United States of America
JUN-YAN ZHU, Carnegie Mellon Uniersity, United States of America

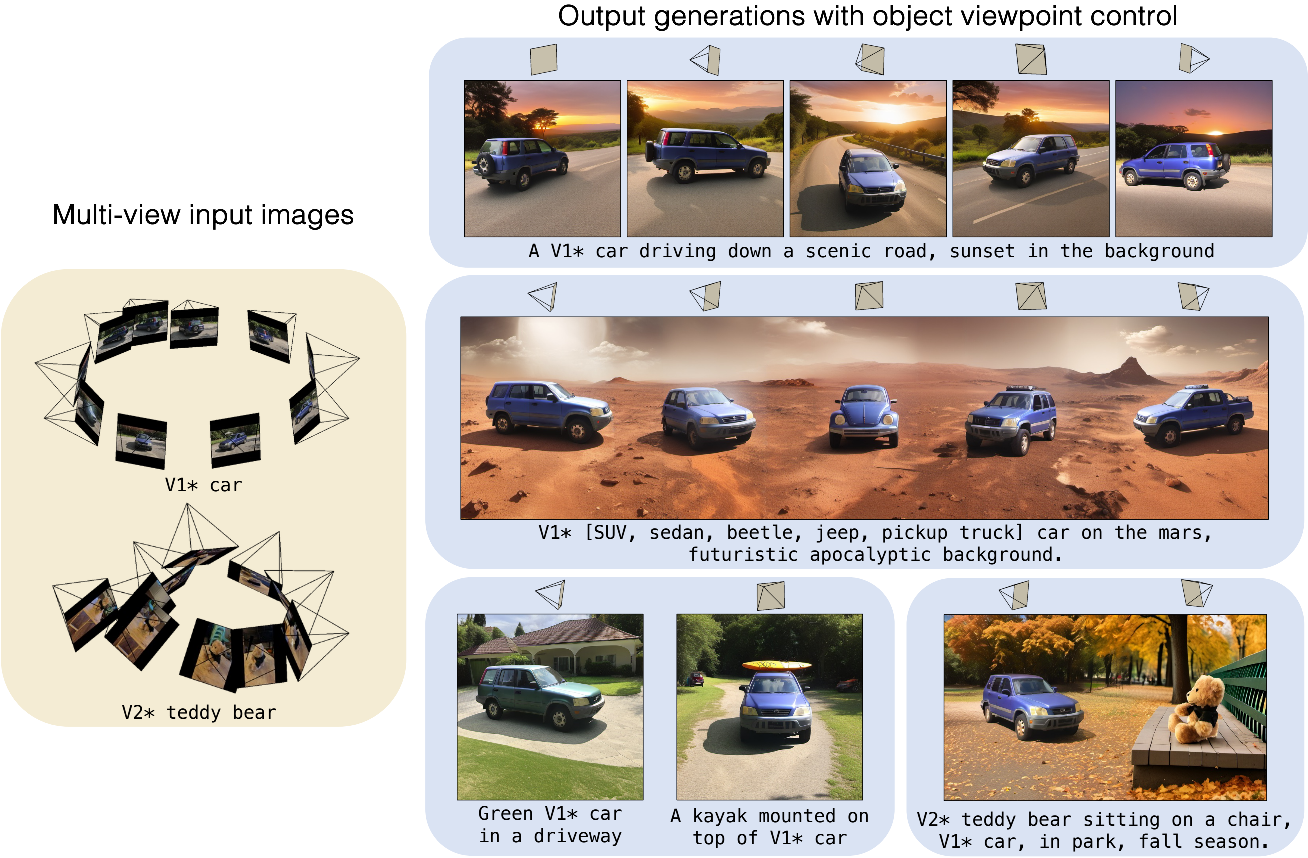

Output generations with object viewpoint control
Multi-view input images
A V1∗ car driving down a scenic road, sunset in the background
V1∗ car
V1∗ [SUV, sedan, beetle, jeep, pickup truck] car on the mars, futuristic apocalyptic background.
V2∗ teddy bear
Green V1∗ car in a driveway
A kayak mounted on top of V1∗ car
V2∗ teddy bear sitting on a chair, V1∗ car, in park, fall season.

Fig. 1. Given multi-view images of a new object (left), denoted as `V* <category name>`, we create a customized text-to-image diffusion model with object viewpoint control. The customized model allows users to specify the target viewpoint for the object while synthesizing it in novel appearances and scenes, such as `A green V* car`, or `A beetle-like V* car`. We can also generate panorama images or compose multiple concepts while controlling each object's viewpoint by using MultiDiffusion [Bar-Tal et al. 2023] with our model.

Model customization introduces new concepts to existing text-to-image models, enabling the generation of these new concepts/objects in novel contexts. However, such methods lack accurate camera view control with respect to the new object, and users must resort to prompt engineering (e.g., adding "top-view") to achieve coarse view control. In this work, we introduce a new task – enabling explicit control of the *object viewpoint* in the customization of text-to-image diffusion models. This allows us to modify the custom object's

SA Conference Papers '24, December 03–06, 2024, Tokyo, Japan
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1131-2/24/12
https://doi.org/10.1145/3680528.3687564

[*] indicates equal contribution.

Authors' Contact Information: Nupur Kumari, Carnegie Mellon Uniersity, United States of America, nkumari@andrew.cmu.edu; Grace Su, Carnegie Mellon Uniersity, United States of America, graceduansu@gmail.com; Richard Zhang, Adobe Research, United States of America, rizhang@adobe.com; Taesung Park, Adobe Research, United States of America, tapark@adobe.com; Eli Shechtman, Adobe Research, United States of America, elishe@adobe.com; Jun-Yan Zhu, Carnegie Mellon Uniersity, United States of America, junyanz@andrew.cmu.edu.

properties and generate it in various background scenes via text prompts, all while incorporating the object viewpoint as an additional control. This new task presents significant challenges, as one must harmoniously merge a 3D representation from the multi-view images with the 2D pre-trained model. To bridge this gap, we propose to condition the diffusion process on the 3D object features rendered from the target viewpoint. During training, we fine-tune the 3D feature prediction modules to reconstruct the object's appearance and geometry, while reducing overfitting to the input multi-view images. Our method outperforms existing image editing and model customization baselines in preserving the custom object's identity while following the target object viewpoint and the text prompt.

CCS Concepts: • **Computing methodologies → Computer vision**; **Image manipulation**.

Additional Key Words and Phrases: Image Generation, Diffusion Models, Deep Generative Models, Model Customization

ACM Reference Format:
Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. 2024. Customizing Text-to-Image Diffusion with Object Viewpoint Control. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 03–06, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3680528.3687564

1 Introduction

Recently, we have witnessed an explosion of works on customizing text-to-image models [Chen et al. 2023; Gal et al. 2023a; Kumari et al. 2023; Ruiz et al. 2023a]. Such methods enable a model to quickly acquire visual concepts, such as personal objects and favorite places, and re-imagine them with new environments and attributes. For instance, we can customize a model on our teddy bear and prompt it with "Teddy bear on a bench in the park." However, these methods lack precise viewpoint control, as the pre-trained model is trained purely on 2D images without ground truth camera poses. As a result, users often rely on text prompts such as "front-facing" or "side-facing", a tedious and unwieldy process to control views.

What if a user wishes to control the custom object's viewpoint while synthesizing it in a different context, e.g., the car in Figure 1? In this work, we introduce a new task: given multi-view images of an object, we customize a text-to-image model while enabling control of the object's viewpoint. During inference, our method offers the flexibility of conditioning the generation process on both a target viewpoint and a text prompt.

Neural rendering methods have allowed us to accurately control the 3D viewpoint of an *existing* scene, given multi-view images [Barron et al. 2021, 2023; Kerbl et al. 2023; Müller et al. 2022]. Similarly, we seek to imagine the object from novel views but in a *new* context. However, as pre-trained diffusion models, such as Latent Diffusion models [Rombach et al. 2022], are built upon a purely 2D representation, connecting the 3D neural representation of the object to the 2D internal features of the diffusion model remains challenging.

In this work, we introduce CustomDiffusion360, a new method to bridge the gap between 3D neural capture and 2D text-to-image diffusion models by providing viewpoint control for custom objects. More concretely, given multi-view images of an object, we introduce FeatureNeRF blocks in the diffusion model U-Net's intermediate feature spaces to learn view-dependent features. To condition the generation process on a target viewpoint, we render the FeatureNeRF output from this viewpoint and merge it with the diffusion features using linear projection layers. We only train the new linear projection layers and FeatureNeRF blocks, added to a subset of transformer layers, to preserve object identity while maintaining generalization. The pre-trained model's parameters remain frozen, thus keeping our method computationally and storage efficient.

We build our method on Stable Diffusion-XL [Podell et al. 2023] and show results on various object categories, such as cars, chairs, motorcycles, teddy bears, and toys. We compare our approach with image editing [Brooks et al. 2023; Meng et al. 2022], model customization [Hu et al. 2022], and NeRF editing methods [Dong and Wang 2023]. Our method achieves high alignment with the custom object's identity and target viewpoint while adhering to the user-provided text prompt. We show that integrating the 3D object information into the text-to-image model, as done by our method, enhances performance over 2D and 3D editing baseline methods. Additionally, our method can be combined with other algorithms [Bar-Tal et al. 2023; Meng et al. 2022] for applications such as object viewpoint adjustment in the same background, panorama synthesis, and object composition.

2 Related Works

Text-based image synthesis. Large-scale text-to-image models [Gafni et al. 2022; Kang et al. 2023; Ramesh et al. 2022; Saharia et al. 2022; Yu et al. 2022] have become ubiquitous for generating photorealistic images from text prompts. This progress has been driven by the availability of large-scale datasets [Schuhmann et al. 2021] as well as advancements in model architecture and training objectives [Dhariwal and Nichol 2021; Karras et al. 2022, 2023; Peebles and Xie 2023; Sauer et al. 2023]. Among them, diffusion models [Ho et al. 2020; Song et al. 2021] have emerged as a powerful family of models that generate images by gradually denoising Gaussian noise.

Image editing with text-to-image diffusion. One of the first works, SDEdit [Meng et al. 2022], exploited the denoising nature of diffusion models, guiding generation in later denoising timesteps using edit instructions while preserving the input image layout. Since then, various works improved upon this by embedding the input image into the model's latent space [Kawar et al. 2023; Mokady et al. 2023; Parmar et al. 2023; Song et al. 2021] or using cross-attention and self-attention mechanisms for realistic and targeted edits [Cao et al. 2023; Chefer et al. 2023; Ge et al. 2023; Hertz et al. 2023; Patashnik et al. 2023]. Recently, several methods train conditional diffusion models to follow user edit instructions or spatial controls [Brooks et al. 2023; Zhang and Agrawala 2023]. However, these methods primarily focus on appearance editing, while our work enables both viewpoint and appearance control.

Model customization. While pre-trained models can generate common objects, users often wish to synthesize images with concepts from their own lives. This has given rise to the emerging technique of model personalization or customization [Gal et al. 2023a; Kumari et al. 2023; Ruiz et al. 2023a]. These methods aim at embedding a new concept, e.g., pet dog, personal car, person, etc., into the output space of text-to-image models. This enables generating new images

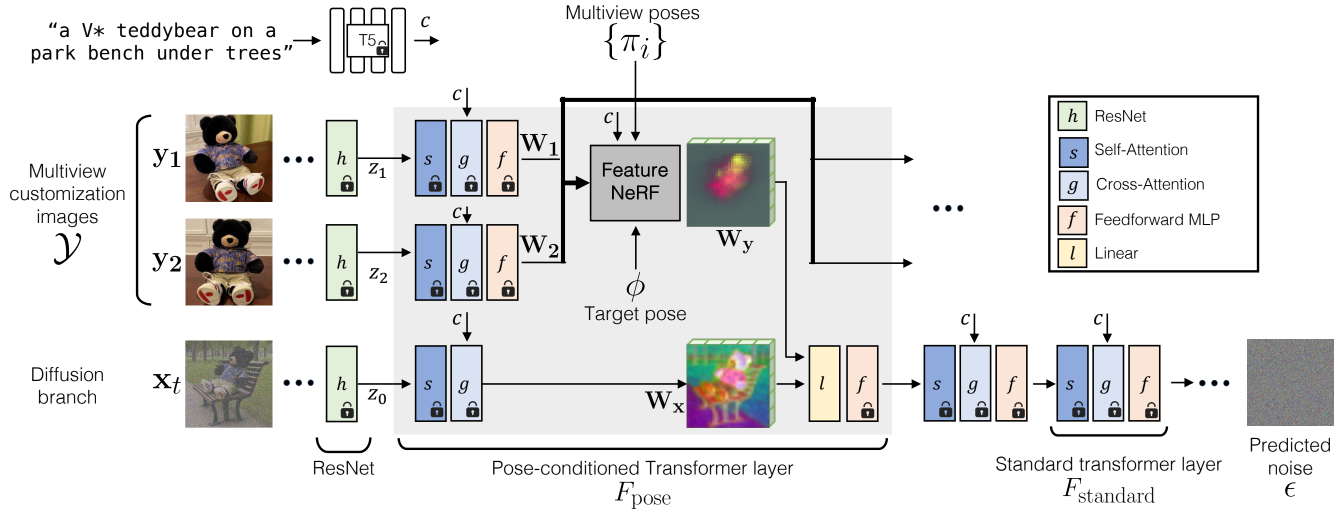


Fig. 2. **Overview.** We propose a model customization method that utilizes N reference images defining the 3D structure of an object $\mathcal{Y}$ (we illustrate with $N = 2$ views for simplicity). We modify the diffusion model U-Net with pose-conditioned transformer blocks. Our **Pose-conditioned transformer block** features a FeatureNeRF module, which aggregates features from the individual viewpoints to target viewpoint ϕ, as shown in detail in Figure 3. The rendered feature W_y is concatenated with the target noisy feature W_x and projected to the original channel dimension. We use the diffusion U-Net itself to extract features of reference images, as shown in the top row. We only fine-tune the new parameters in linear projection layer l and FeatureNerF in F_{pose} blocks.

of the concept in unseen scenarios using the text prompt, e.g., my car in a field of sunflowers. To achieve this, various works fine-tune a small subset of model parameters [Han et al. 2023; Hu et al. 2022; Kumari et al. 2023; Tewel et al. 2023] and/or optimize text token embeddings [Alaluf et al. 2023; Gal et al. 2023a; Voynov et al. 2023; Zhang et al. 2023] on the few images of the new concept with different regularizations [Kumari et al. 2023; Ruiz et al. 2023a]. More recently, encoder-based methods have been proposed that train a model on a vast dataset of concepts [Arar et al. 2023; Gal et al. 2023b; Li et al. 2023; Ruiz et al. 2023b; Shi et al. 2023; Valevski et al. 2023; Wei et al. 2023; Ye et al. 2023b], enabling faster customization. However, to our knowledge, no existing work allows for controlling the viewpoint in model customization. Given the ease of capturing multi-view images of a new concept, this work explores augmenting model customization with additional object viewpoint control.

View synthesis. Novel view synthesis aims to render a scene from unseen camera poses, given multi-view images. Recently, the success of volumetric rendering-based approaches like NeRF [Mildenhall et al. 2021] have led to numerous follow-up works with better quality [Barron et al. 2021, 2023], faster speed [Chen et al. 2022; Müller et al. 2022], and fewer training views [Deng et al. 2022; Niemeyer et al. 2022; Tancik et al. 2021; Yu et al. 2021]. Recent works learn generative models with large-scale multi-view data to learn generalizable representations for novel view synthesis [Burgess et al. 2024; Chan et al. 2023; Liu et al. 2023, 2024; Sargent et al. 2023; Wu et al. 2024; Zhou and Tulsiani 2023]. While our work draws motivation from this line of research, our goal differs - we aim to enable object viewpoint control in text-to-image personalization, rather than capturing novel views of real scenes. Concurrent to our work, ReconFusion [Wu et al. 2024] also trains a PixelNeRF [Yu et al. 2021] in the latent space of latent diffusion models for 3D reconstruction. Different from this, we learn volumetric features in the intermediate attention layers. We also focus on model customization rather than scene reconstruction. Recently, Cheng *et al.* [2024] and Höllein *et al.* [2024] propose adding camera pose conditioning in text-to-image diffusion models while we focus on model customization. CustomNet [Yuan et al. 2024], a concurrent work, also proposes to generate custom objects in a target viewpoint in a zero-shot manner. However, it focuses primarily on generating the new object in different backgrounds, whereas our method allows any new text prompt and viewpoint combination as a condition during inference.

3D editing. Loosely related to our work, many works have been proposed for inserting and manipulating 3D objects within 2D real photographs by editing the image, using classic geometry-based approaches [Chen et al. 2013; Karsch et al. 2011; Kholgade et al. 2014] or generative modeling techniques [Michel et al. 2023; Xu et al. 2023; Yao et al. 2018; Yenphraphai et al. 2024; Zhang et al. 2021]. Instead of editing a single image, our work aims to "edit" the model weights of a pre-trained diffusion model. Another relevant line of work edits [Dong and Wang 2023; Haque et al. 2023] or generates [Metzer et al. 2023; Raj et al. 2023; Shi et al. 2024; Tang et al. 2023; Xu et al. 2024] a 3D scene given a text prompt or image. Unlike these methods, we do not aim to edit/generate a multi-view consistent scene. Our goal is to provide additional viewpoint control when customizing text-to-image models. This enables specifying the object viewpoint while generating new backgrounds or composing multiple objects. Additionally, we show that our method achieves greater photorealism compared to a 3D editing method for this task.

3 Method

Given multi-view images of a custom object, we aim to embed it in the text-to-image diffusion model. We construct our method in order to allow the generation of new variations of the object through text prompts while providing control of the object viewpoint. Our approach involves fine-tuning the pre-trained model while conditioning it on a 3D representation of the object learned in the diffusion model's feature space. In this section, we briefly overview the diffusion model and then explain our method in detail.

3.1 Diffusion Models

Diffusion models [Ho et al. 2020; Sohl-Dickstein et al. 2015] are a class of generative models that sample images by iterative denoising of a random Gaussian distribution. The training of the diffusion model consists of a forward Markov process, where real data $\mathbf{x}_0$ is gradually transformed to random noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ by sequentially adding Gaussian perturbations in T timesteps, i.e., $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\epsilon$. The model is trained to learn the backward process, i.e.,

$$p_\theta(\mathbf{x}_0|\mathbf{c}) = \int \Big[p_\theta(\mathbf{x}_T) \prod p_\theta^t(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})\Big] d\mathbf{x}_{1:T}, \tag{1}$$

The training objective maximizes the variational lower bound, which can be simplified to a simple reconstruction loss:

$$\mathbb{E}_{\mathbf{x}_t, t, \mathbf{c}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [w_t ||\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})||], \tag{2}$$

where $\mathbf{c}$ can be any modality to condition the generation process. The model is trained to predict the noise added to create the input noisy image $\mathbf{x}_t$. During inference, we gradually denoise a random Gaussian noise over a fixed number of timesteps. Various proposed sampling strategies [Karras et al. 2022; Lu et al. 2022; Song et al. 2021] reduce the number of sampling steps compared to the typical 1000 timesteps in training. In our work, we use the Stable Diffusion-XL (SDXL) [Podell et al. 2023] as the pre-trained text-to-image diffusion model. It is based on the Latent Diffusion Model (LDM) [Rombach et al. 2022], which is trained in an autoencoder [Kingma and Welling 2014] latent space.

3.2 Customization with Object Viewpoint Control

Model customization aims to condition the model on a new object, given N images of the object $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$, i.e., to model $p(\mathbf{x}|\mathcal{Y}, \mathbf{c})$ with text prompt $\mathbf{c}$. In contrast, we additionally condition the model on the object viewpoint, allowing more control in the generation process. Thus, given a set of multi-view images $\{\mathbf{y}_i\}_{i=1}^N$ and the corresponding camera poses $\{\pi_i\}_{i=1}^N$, our goal is to learn the conditional distribution $p(\mathbf{x}|\{(\mathbf{y}_i, \pi_i)\}_{i=1}^N, \mathbf{c}, \phi)$, where $\mathbf{c}$ is text prompt and ϕ is the camera pose corresponding to the target viewpoint. To achieve this, we fine-tune a pre-trained text-to-image diffusion model, which models $p(\mathbf{x}|\mathbf{c})$, with the additional conditioning of camera pose ϕ given posed reference images $\{\mathbf{y}_i, \pi_i\}_{i=1}^N$.

Model architecture. In Figure 2, we show the overall architecture, with an emphasis on our added pose-conditioning. Each block in the diffusion model U-Net [Ronneberger et al. 2015] consists of a ResNet [He et al. 2016], denoted as h, followed by several transformer layers [Vaswani et al. 2017]. Given the output of an intermediate ResNet layer $\mathbf{z}$, a standard transformer layer, $F_{\text{standard}}(\mathbf{z}, \mathbf{c})$, consists of a self-attention layer, denoted as s, followed by cross-attention with the text prompt, denoted as g, and a feed-forward MLP, denoted as f. We modify a subset of these transformer layers to incorporate pose conditioning as we explain next.

Pose-conditioned transformer layer. We denote the pose-conditioned transformer layer as $F_{\text{pose}}(\mathbf{z}_0, \{\mathbf{z}_i, \pi_i\})$, where $\mathbf{z}_0$ is the intermediate target feature (diffusion branch in Figure 2) and $\{\mathbf{z}_i\}$ are the input features corresponding to multi-view reference images (top two rows in Figure 2). We extract spatial features $\{\mathbf{W}_i\}$ from $\{\mathbf{z}_i\}$ using components of pre-trained U-Net itself, i.e., $F_{\text{standard}}(\mathbf{z}_i, \mathbf{c})$. To condition the diffusion branch on ϕ, we learn a radiance field, denoted as FeatureNeRF, from $\{\mathbf{W}_i, \pi_i\}$ in a feed-forward manner [Yu et al. 2021]. The predicted FeatureNeRF is then rendered from the target viewpoint ϕ to obtain view-dependent feature map $\mathbf{W}_y$.

In the main diffusion branch, we extract the intermediate feature map after the self and cross-attention layers, i.e., $\mathbf{W}_\mathbf{x} = g(s(\mathbf{z}_0), \mathbf{c})$. We concatenate $\mathbf{W}_\mathbf{x}$ with the rendered features $\mathbf{W}_y$ and then project it into the original feature dimension using a linear layer. Thus, the pose conditioned transformer layer, $F_{\text{pose}}(\mathbf{z}_0, \{\mathbf{z}_i, \pi_i\}, \mathbf{c}, \phi)$ performs:

$$\begin{gathered}\mathbf{W}_i = F_{\text{standard}}(\mathbf{z}_i, \mathbf{c}), \quad \mathbf{W}_y = \text{FeatureNeRF}(\{\mathbf{W}_i, \pi_i\}, \mathbf{c}, \phi) \\ F_{\text{pose}} = f(l(\mathbf{W}_y \oplus \mathbf{W}_\mathbf{x}))\end{gathered} \tag{3}$$

where l is a learnable weight matrix, which projects the feature into the input space of feed-forward layer f. We initialize l such that the contribution from $\mathbf{W}_y$ is zero at the start of training.

FeatureNeRF.. Here, we describe the aggregation of individual features $\mathbf{W}_i$ with poses π_i into a feature map $\mathbf{W}_y$ from pose ϕ. Given a target ray with direction $\mathbf{d}$ from target viewpoint ϕ, we sample points $\mathbf{p}$ along the ray and project it to the image plane of each given view π_i. The projected coordinate is denoted as $\pi_i^\mathbf{p}$. We then sample the feature from this coordinate in $\mathbf{W}_i$, predict a feature for the 3D point $\mathbf{p}$, and aggregate the N predicted features from each view with function ψ:

$$\begin{aligned}\mathbf{V}_i^\mathbf{p} =& \text{MLP}(\text{Sample}(\mathbf{W}_i; \pi_i^\mathbf{p}), \gamma(\mathbf{d}), \gamma(\mathbf{p})), \; i = 1, ..., N \\ \bar{\mathbf{V}}^\mathbf{p} =& \psi(\mathbf{V}_1^\mathbf{p}, ..., \mathbf{V}_N^\mathbf{p}),\end{aligned} \tag{4}$$

where γ is the frequency encoding. We use the weighted average [Reizenstein et al. 2021] as the aggregation function ψ, where a linear layer predicts the weights based on $\mathbf{V}_i$, π_i, and ϕ. For each reference view, $\mathbf{d}$ and $\mathbf{p}$ are first transformed in the view coordinate space [Yu et al. 2021]. Given the feature $\bar{\mathbf{V}}$ (superscript $\mathbf{p}$ is dropped for simplicity) for the 3D point, we predict the density and color using a linear layer:

$$(\sigma, \mathbf{C}) = \text{MLP}(\bar{\mathbf{V}}), \tag{5}$$

and also update the aggregated feature with text prompt $\mathbf{c}$ using cross-attention:

$$\hat{\mathbf{V}} = \text{CrossAttn}(\bar{\mathbf{V}}, \mathbf{c}). \tag{6}$$

We then render this updated feature volume using the predicted densities:

$$\mathbf{W}_y(r) = \sum_{j=1}^{N_f} T_j(1 - \exp(-\sigma_j \delta_j))\hat{\mathbf{V}}_j, \tag{7}$$

where r is the target ray, $\hat{\mathbf{V}}_j$ is the feature corresponding to the j^{th} point along the ray, σ_j is the predicted density of that point, N_f is the number of sampled points along the ray between the near

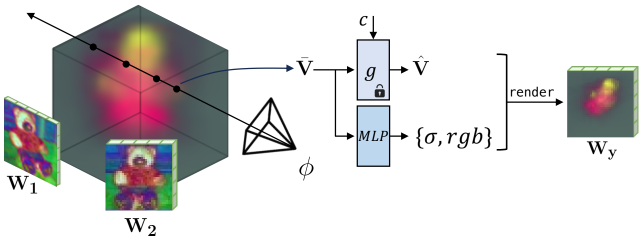


Fig. 3. **FeatureNeRF**. We predict volumetric features $\bar{\mathbf{V}}$ for each 3D point in the grid using reference features $\{\mathbf{W}_i\}$ (Eqn. 4). Given this feature, we predict the density σ and color rgb using a 2-layer MLP and use the predicted density σ to render $\hat{\mathbf{V}}$ (which has been updated with text cross-attention g). The rgb is only used to calculate reconstruction loss during training.

and far plane of the camera, and $T_j = \exp(-\sum_{k=1}^{j-1} \sigma_k \delta_k)$ handles occlusion until that point.

We build our FeatureNeRF design based on PixelNeRF [Yu et al. 2021] but update the aggregated features with text cross-attention and use learnable weighted averaging to aggregate reference view features. Through this layer, our focus is on learning 3D features that the 2D diffusion model can use rather than learning NeRF in a feature space [Kerr et al. 2023; Ye et al. 2023a].

Training loss. Our training objective includes learning 3D consistent FeatureNeRF modules, which can contribute to the final goal of reconstructing the target concept in diffusion model's output space. Thus, we fine-tune the model using the sum of training losses corresponding to FeatureNeRF and the default diffusion model reconstruction loss:

$$\mathcal{L}_{\text{diffusion}} = \sum_r M w_t ||\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})||, \tag{8}$$

where M is the object mask, with the reconstruction loss being calculated only in the object mask region. The losses corresponding to FeatureNeRF consist of RGB reconstruction loss:

$$\mathcal{L}_{\text{rgb}} = \sum_r ||M(r)(\mathbf{C}_{gt}(r) - \sum_{j=1}^{N_f} T_j(1 - \exp(-\sigma_j \delta_j))\mathbf{C})||, \tag{9}$$

and two mask-based losses as we only wish to model the object – (1) silhouette loss [Ravi et al. 2020] $\mathcal{L}_{\text{s}}$ which forces the rendered opacity to be similar to object mask, and (2) background suppression loss [Boss et al. 2021, 2022] $\mathcal{L}_{\text{bg}}$ which enforces the density of all background rays to be zero.

$$\begin{aligned} \mathcal{L}_{\text{s}} &= \sum_r ||M(r) - \sum_{j=1}^{N_f} T_j(1 - \exp(-\sigma_j \delta_j))|| \\ \mathcal{L}_{\text{bg}} &= \sum_r (1 - M(r)) \sum_{j=1}^{N_f} ||(1 - \exp(-\sigma_j \delta_j))||, \end{aligned} \tag{10}$$

Thus, the final training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda_{\text{rgb}} \mathcal{L}_{\text{rgb}} + \lambda_{\text{bg}} \mathcal{L}_{\text{bg}} + \lambda_{\text{s}} \mathcal{L}_{\text{s}}, \tag{11}$$

where M is the object mask and λ_{rgb}, λ_{bg}, and λ_{s} are hyperparameters to control the rendering quality of intermediate images vs. the final denoised image and are fixed across all experiments. We assume access to the object's mask in the image, which is used to calculate these losses. The losses for FeatureNeRF are averaged across all pose-conditioned transformer layers. We provide more training and implementation details in the supplemental material.

4 Experiments

Dataset. We select 14 custom objects from the CO3Dv2 [Reizenstein et al. 2021] and NAVI [Jampani et al. 2023] datasets. Specifically, we select 4 categories with 3 instances each from the CO3Dv2 dataset – car, chair, teddy bear, and motorcycle. From NAVI, we select 2 unique, toy-like objects. A representative image of each concept is shown in the supplemental material. We use the camera poses provided in the dataset. For each instance, we sample ∼ 100 images, using half for training and half for evaluation.

Baselines. We compare with three types of relevant baselines.

- First, we compare against 2D image editing using 3 recent, publicly available methods – LEDITS++ [Brack et al. 2024], InstructPix2Pix [Brooks et al. 2023], and SDEdit [Meng et al. 2022] with Stable Diffusion-1.5 (and SDXL in the supplemental material). As image editing methods do not inherently support viewpoint manipulation, we first render a NeRF [Tancik et al. 2023] of the input scene with the target viewpoint and then edit the rendered image.
- Secondly, we use a customization-based method, LoRA+Camera pose, where we modify LoRA [Hu et al. 2022; Ryu 2023] by concatenating the camera pose to the text embeddings, following recent work Zero-1-to-3 [Liu et al. 2023].
- Lastly, we test ViCA-NeRF [Dong and Wang 2023], a 3D editing method that trains a NeRF for each new text prompt.

We provide implementation details in the supplemental material.

Evaluation metrics. To create an evaluation set, we generate 16 prompts per object category using [ChatGPT 2022]. We instruct ChatGPT to propose four types of prompts: scene change, color change, object composition, and shape change. We then manually inspect them to remove implausible or overly complicated text prompts [Wang et al. 2023]. We evaluate (1) the customization quality of the generated image and (2) its adherence to the specific pose.

First, to measure customization quality, we use a pairwise human preference study. A successful customization is comprised of several aspects: alignment to the target concept, alignment to the input text prompt, and photorealism of the generated images. In total, we collect ∼ 1000 responses per pairwise study against each baseline using Amazon Mechanical Turk. We also evaluate our method and baselines on other standard metrics like CLIP Score [Radford et al. 2021] and DINOv2 [Oquab et al. 2023] image similarity [Ruiz et al. 2023a] to measure the text and image alignment.

To measure whether the generated custom object adheres to the specified viewpoint, we use a pre-trained model, RayDiffusion [Zhang et al. 2024], to predict the pose from the generated images and calculate its error relative to the input camera pose.

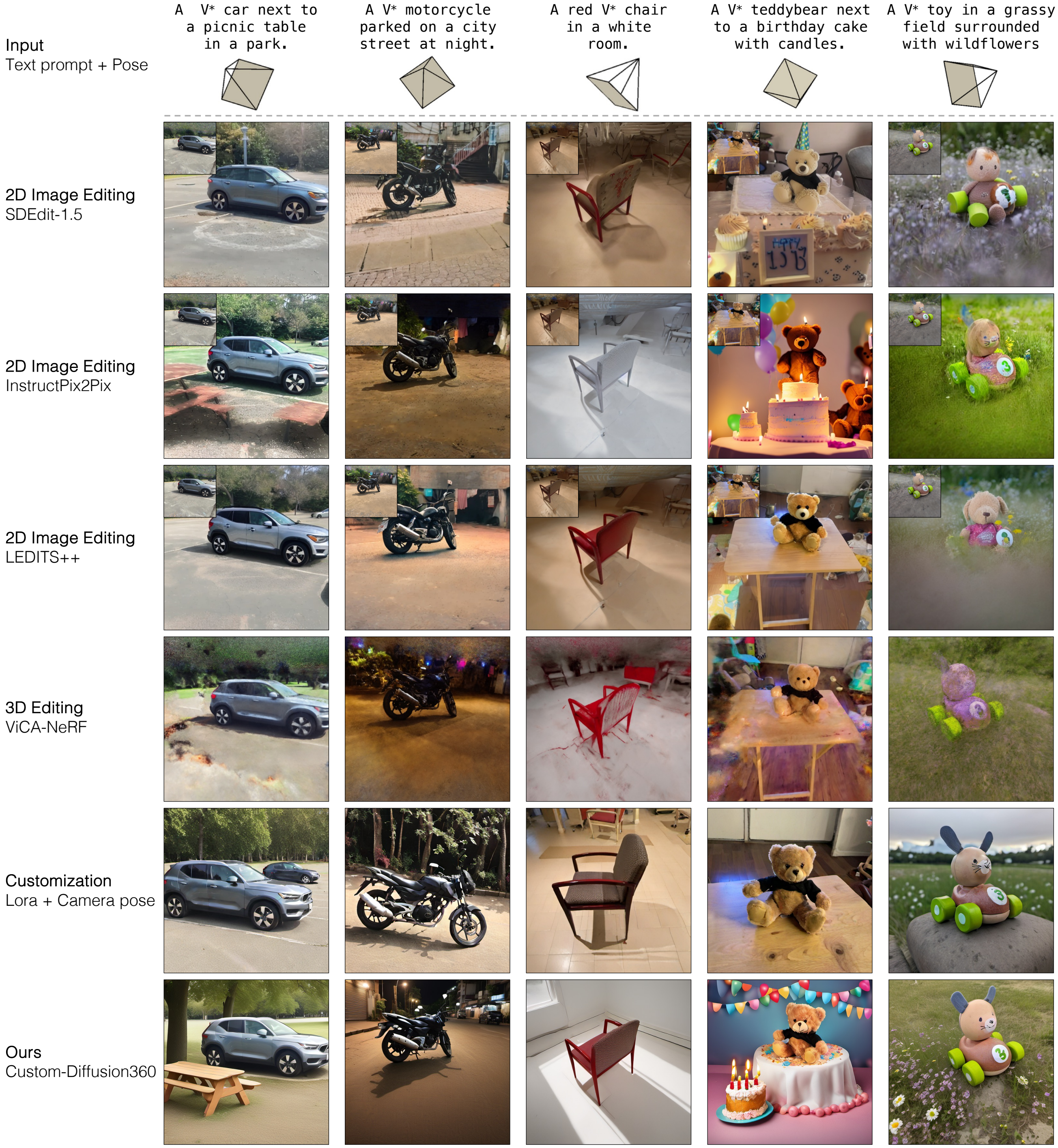
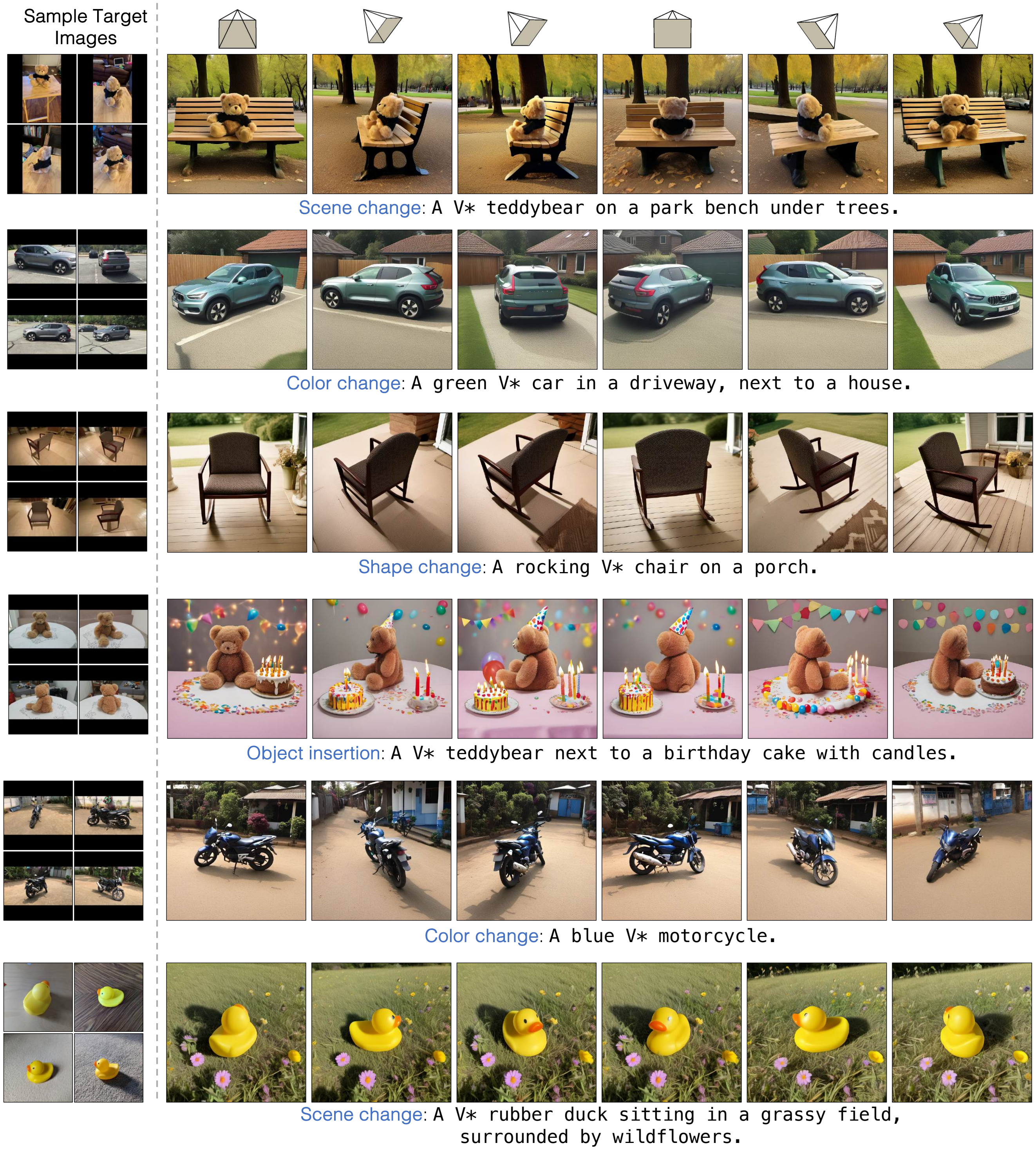

Input
Text prompt + Pose
A V* car next to a picnic table in a park.
A V* motorcycle parked on a city street at night.
A red V* chair in a white room.
A V* teddybear next to a birthday cake with candles.
A V* toy in a grassy field surrounded with wildflowers
2D Image Editing
SDEdit-1.5
2D Image Editing
InstructPix2Pix
2D Image Editing
LEDITS++
3D Editing
ViCA-NeRF
Customization
Lora + Camera pose
Ours
Custom-Diffusion360

Fig. 4. **Qualitative comparison.** Given a particular target pose, we show the qualitative comparison of our method with (1) Image editing methods *SDEdit*, *InstructPix2Pix*, and *LEDITS++*, which edit a NeRF-rendered image from the input pose, (2) *ViCA-NeRF*, a 3D editing method that trains a NeRF model for each input prompt, and (3) *LoRA + Camera pose*, our proposed baseline where we concatenate camera pose information to text embeddings during LoRA fine-tuning. Our method performs on par or better in keeping the target identity and poses while incorporating the new text prompt—e.g., putting a picnic table next to the SUV car (1st column)—and following multiple text conditions—e.g., turning the chair red and placing it in a white room (3rd column). $\mathbb{V}^*$ token is used only in ours and the LoRA + Camera pose method. Ground truth rendering from the given pose is shown as an inset in the first three rows. We show more sample comparisons in the supplement.



Scene change: A V∗ teddybear on a park bench under trees.

Color change: A green V∗ car in a driveway, next to a house.

Shape change: A rocking V∗ chair on a porch.

Object insertion: A V∗ teddybear next to a birthday cake with candles.

Color change: A blue V∗ motorcycle.

Scene change: A V∗ rubber duck sitting in a grassy field, surrounded by wildflowers.

Fig. 5. **Qualitative samples with varying object viewpoint and text prompt**. Our method learns the identity of custom objects while allowing the user to control the object viewpoint and generating the object in new contexts using the text prompt, e.g., changing the background scene or object color and shape. In each row, the images were generated with the same seed while changing the object viewpoint in a turntable manner. We show more samples in the supplement. Note that each image in a row is independently generated.

Table 1. **Human preference evaluation**. Our method is preferred over almost all baselines for text alignment, image alignment to the target concept, and photorealism. We find that LoRA + Camera pose overfits the training images, as shown in Figure 4.

Method	Text Alignment	Image Alignment	Photorealism
SDEdit	40.6 ± 2.7%	36.1 ± 2.8%	33.1 ± 3.2%
vs. Ours	**59.4** ± 2.7%	**63.9** ± 2.8%	**66.9** ± 3.2%
InstructPix2Pix	44.8 ± 2.6%	29.3 ± 2.2 %	27.6 ± 2.6%
vs. Ours	**55.2** ± 2.6%	**70.7** ± 2.2 %	**72.4** ± 2.6%
LEDITS++	32.5 ± 2.5%	35.9 ± 2.5%	26.2 ± 2.8%
vs. Ours	**67.5**± 2.5%	**64.1** ± 2.5%	**73.8** ± 2.8%
ViCA-NeRF	27.1 ± 2.8%	24.4 ± 3.3%	12.9 ± 2.7%
vs. Ours	**72.9** ± 2.8%	**75.6** ± 3.3 %	**87.1** ± 2.7 %
LoRA + Camera pose	32.3 ± 2.7%	**66.9** ± 2.5 %	**52.5** ± 2.8%
vs. Ours	**67.7** ± 2.7%	33.1 ± 2.5%	47.5 ± 2.8%

Table 2. **Accuracy of object viewpoint condition** in generated images by ours and the LoRA + Camera pose baseline method. We observe that the baseline usually overfits the training images and does not respect the target viewpoint condition with new text prompts.

Method	**Angular error ↓**	**Camera center error↓**
Ours	**14.19**	**0.080**
LoRA + Camera pose	41.14	0.305

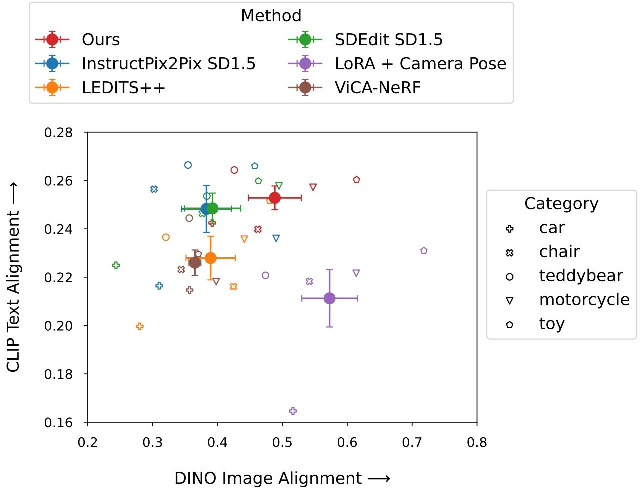


Fig. 6. **Quantitative comparison**. We show CLIP scores (higher is better) vs. DINO-v2 scores (higher is better). We plot the performance of each method on each category and the overall mean and standard error (highlighted). Our method results in higher CLIP text alignment while maintaining visual similarity to target concepts, as indicated by DINO-v2 scores. The text alignment of our method compared to SDEdit and InstructPix2Pix is only marginally better as these methods incorporate the text prompt but at the cost of photorealism, as we show in Table 1.

More details about the evaluation and prompts are provided in the supplemental material.

4.1 Results

Generation quality and adherence. First, we measure the quality of the generation – adherence to the text prompt, the identity preservation to the customized objects, and photorealism – irrespective of the object viewpoint. Recall that for each concept, we curate 16 prompts. For each prompt, we generate 3 images at each viewpoint, covering 6 target viewpoints, resulting in 288 images per concept. Table 1 shows the pairwise human preference for our method vs. baselines. Our method is preferred over all baselines except LoRA + Camera pose, which we observe to overfit on training images, thus producing higher image alignment. Figure 6 shows the CLIP vs. DINO scores for all methods. Ideally, a method should have both a high CLIP score and a DINO score, but often, there is a trade-off between text and image alignment. Our method has on-par or better text alignment relative to the baselines, while having better image alignment. We observe that image-editing baselines often require careful hyperparameter tuning for each image. We select the best-performing hyperparameters and keep them fixed across all experiments. The camera pose corresponding to the target object viewpoint is uniformly sampled from ∼ 50 validation poses not used during training. We also randomly perturb the camera position or focal length. We show sample training and perturbed validation camera poses for the car object in the supplemental material.

Accuracy of object viewpoint. Previously, we evaluated our method purely on image customization benchmarks. Next, we evaluate the accuracy of the object viewpoint conditioning. Table 2 shows the mean angular error and camera center error between the generated object's pose, predicted using RayDiffusion [Zhang et al. 2024], and the input pose. We only compare with LoRA + Camera pose, as only this baseline takes the camera pose for the target object viewpoint as input. We observe that it often overfits training images and fails to generate the object in the correct viewpoint with new text prompts. We evaluate this metric only on the objects from the CO3Dv2 dataset with validation camera poses, as RayDiffusion has been trained on CO3Dv2 and struggles with other unique objects.

Qualitative comparison. We show the qualitative comparison of our method with the baselines in Figure 4. We observe that image editing methods can fail to generate photorealistic results. In the case of LoRA + Camera pose, it fails to generalize and overfits to the training views (5th row). Finally, the 3D editing-based method ViCA-NeRF maintains 3D consistency but generates blurred images, especially for text prompts that change the background. Figure 5 shows more samples with different text prompts and object viewpoints for our method.

Additional comparison to customization + 3D-aware image editing. We further compare against a two-stage approach that first generates an image of the custom object using LoRA+DreamBooth [Ruiz et al. 2023a; Ryu 2023] and then edits the object to a target viewpoint using two recent 3D-aware image editing methods, Image Sculpting [Yenphraphai et al. 2024] and Object3DIT [Michel et al. 2023]. For each prompt, we generate 3 images, then edit and rotate the object to 6 different viewpoints. This results in 288 images per concept, similar to our evaluation setting. We compare against this on only the three car objects since Image-Sculpting uses Adobe

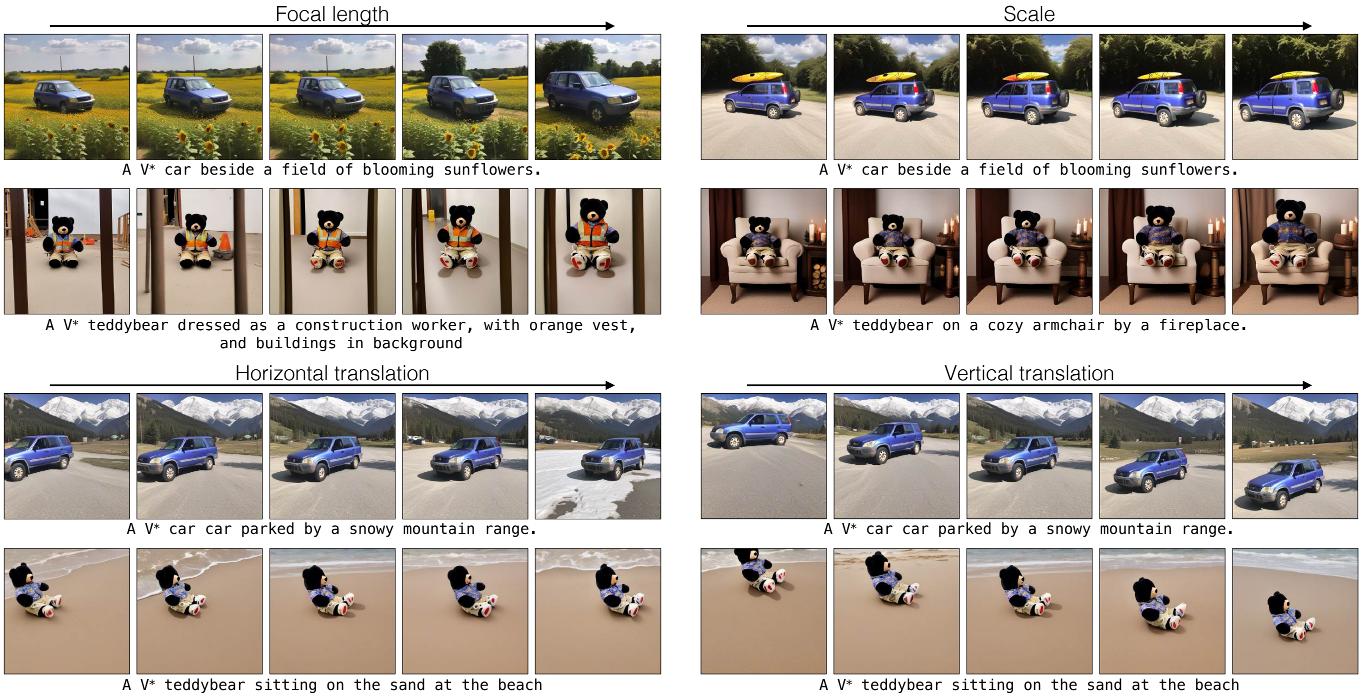
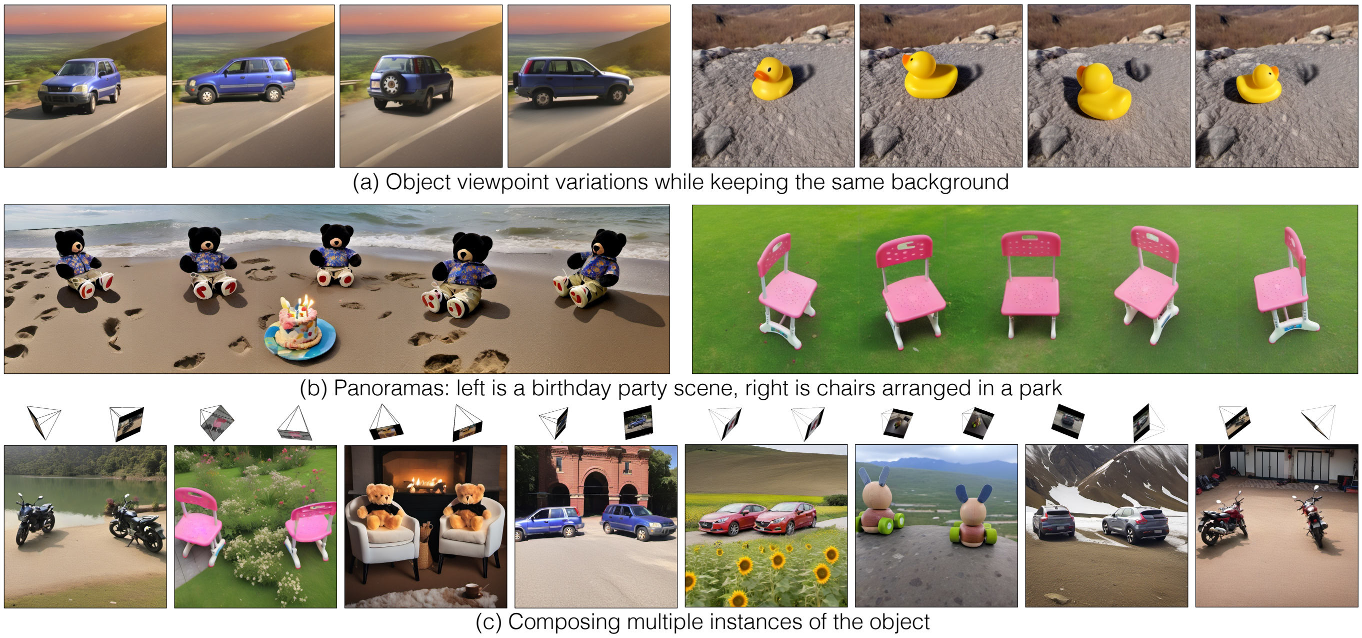

Focal length
A V* car beside a field of blooming sunflowers.
A V* teddybear dressed as a construction worker, with orange vest, and buildings in background
Scale
A V* car beside a field of blooming sunflowers.
A V* teddybear on a cozy armchair by a fireplace.
Horizontal translation
A V* car car parked by a snowy mountain range.
A V* teddybear sitting on the sand at the beach
Vertical translation
A V* car car parked by a snowy mountain range.
A V* teddybear sitting on the sand at the beach

Fig. 7. **Extrapolating object viewpoint from training viewpoints**. Our method can generalize to different viewpoints, including those not within the training distribution. *Top left:* We vary the focal length from ×0.8 to ×1.4 of the original focal length. *Top right:* We vary the camera position towards the image plane along the z axis. *Bottom row:* We vary the camera position along the horizontal and vertical axis.



(a) Object viewpoint variations while keeping the same background

(b) Panoramas: left is a birthday party scene, right is chairs arranged in a park

(c) Composing multiple instances of the object

Fig. 8. **Applications**. 1^{st} *row*: Our method can be combined with other image editing methods as well. We use SDEdit with our method to in-paint the car and rubber duck from different viewpoints while keeping the same background. 2^{nd} *row*: We can generate interesting panorama shots by controlling the object viewpoint independently in each grid. 3^{rd} *row*: We can also compose the radiance field predicted by FeatureNeRF to control the relative pose while generating multiple instances of the object.

Photoshop's generative fill [Adobe 2023] as one of the intermediate steps, which requires manually inpainting each image. The CLIP scores for Image Sculpting and Object3DIT are 0.26 and 0.27, respectively, compared to our score of 0.25. However, their DINO scores at 0.24 and 0.40 are substantially lower than our 0.48. As shown in Figure 9, both methods lead to lower-fidelity results. Object3DIT struggles in many scenarios due to its training on a synthetic dataset, and Image Sculpting's performance is highly dependent on single image-to-3D methods like Zero-1-to-3 [Liu et al. 2023] used in its pipeline.

Generalization to novel viewpoints. Since our method learns a 3D radiance field, we can also extrapolate to unseen object viewpoints at inference time as shown in Figure 7. We generate images while varying the camera distance from the object (scale), focal length, or camera position along the horizontal and vertical axis.

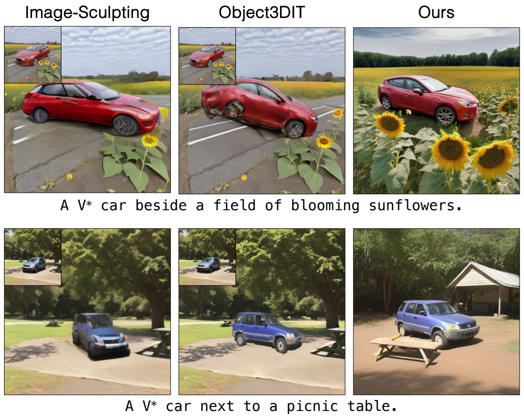
Applications. Our method can be combined with existing image editing methods as well. Figure 8a shows an example where we use SDEdit [Meng et al. 2022] to generate the object with different viewpoints while keeping the same background. We can also generate interesting panoramas using MultiDiffusion [Bar-Tal et al. 2023], where the object viewpoint in each grid is controlled by our method, as shown in Figure 8b. Moreover, since we learn a 3D consistent FeatureNeRF for the new object, we can compose multiple instances of the object [Song et al. 2023], with each instance in a different viewpoint. Figure 8c shows an example of two teddy bears facing each other and sitting on armchairs. Here, we additionally use DenseDiffusion [Kim et al. 2023] to modulate the attention maps and guide the generation of each object instance to only appear in the corresponding region predicted by FeatureNeRF.

We show more results and ablation experiments in the supplemental material, including the role of mask-based losses, the importance of text cross-attention in FeatureNeRF, and performance with predicted camera viewpoints.

5 Discussion and Limitations

We introduce a new task of customizing text-to-image models with object viewpoint control. Our method learns view-dependent object features in the intermediate feature space of the diffusion model and conditions the generation on them. This enables synthesizing the object with varying object viewpoints while controlling other aspects through text prompts.

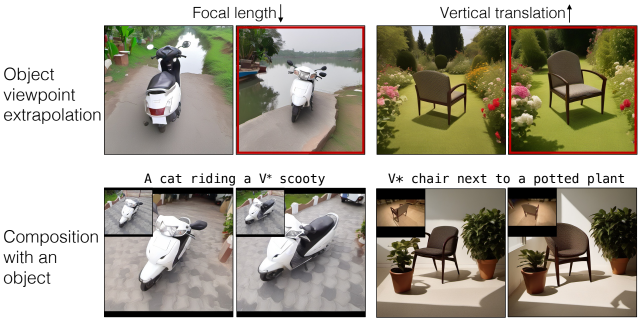
Limitations. Though our method outperforms existing image editing and model customization approaches, it still has several limitations. As we show in Figure 10, our method occasionally struggles at generalizing to extreme viewpoints that were not seen during training and resorts to either changing the object identity or generating the object in a seen viewpoint. We expect this to improve by adding more viewpoint variations during training. Our method also sometimes struggles to follow the input viewpoint condition when the text prompt adds multiple objects to the scene. We hypothesize that in such challenging scenarios, the model is biased towards generating object-centric front views, as seen in its original training data. Also, we fine-tune the model for each custom object, which takes computation time (~ 40 minutes). Exploring pose-conditioning in a zero-shot, feed-forward manner [Chen et al. 2023; Gal et al. 2023b] may help reduce the time and computation. Finally, we focus on enabling viewpoint control for rigid objects. Future work includes extending this conditioning to handle dynamic objects that change the pose in between reference views. One potential way to address this is using a representation based on dynamic and non-rigid NeRF methods [Fridovich-Keil et al. 2023; Pumarola et al. 2021; Song et al. 2023].



A V* car beside a field of blooming sunflowers.

A V* car next to a picnic table.

Fig. 9. **Comparison to 3D-aware editing methods**. We first generate the image of the custom object using LoRA+DreamBooth (shown as an inset) and then use the 3D-aware editing method to edit and rotate the object to a target viewpoint. We show qualitative samples generated by our method (3rd column) with approximately the same target viewpoint as input. Object3DIT and Image-Sculpting lead to lower fidelity edits than images generated by our method (3rd column) with the target viewpoint directly as the input condition.



A cat riding a V* scooty

V∗ chair next to a potted plant

Fig. 10. **Limitations**. Our method can occasionally fail when the target object viewpoint deviates far from the training images, e.g., reducing the focal length too much (top left) or rendering the object off-center (top right), as the pre-trained model is often biased towards generating the object in the center. Also, it can fail to follow the input text prompt or the exact object viewpoint when multiple objects are composed in a scene (bottom row).

Acknowledgments

We are thankful to Kangle Deng, Sheng-Yu Wang, and Gaurav Parmar for their helpful comments and discussion and to Sean Liu, Ruihan Gao, Yufei Ye, and Bharath Raj for proofreading the draft.

This work was partly done by Nupur Kumari during the Adobe internship. The work is partly supported by Adobe Research, the Packard Fellowship, the Amazon Faculty Research Award, and NSF IIS-2239076. Grace Su is supported by the NSF Graduate Research Fellowship (Grant No. DGE2140739).

References

Adobe. 2023. Generative Fill. https://www.adobe.com/products/photoshop/generative-fill.html.

Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *ACM Transactions on Graphics (TOG)* (2023).

Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. 2023. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers.*

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning (ICML).*

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV).*

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV).*

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021. Nerd: Neural reflectance decomposition from image collections. In *IEEE International Conference on Computer Vision (ICCV).*

Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan Barron, Hendrik Lensch, and Varun Jampani. 2022. Samurai: Shape and material from unconstrained real-world arbitrary image collections. In *Conference on Neural Information Processing Systems (NeurIPS).*

Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. 2024. LEDITS++: Limitless Image Editing using Text-to-Image Models. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

James Burgess, Kuan-Chieh Wang, and Serena Yeung-Levy. 2024. Viewpoint Textual Inversion: Discovering Scene Representations and 3D View Control in 2D Diffusion Models. *European Conference on Computer Vision (ECCV)* (2024).

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *IEEE International Conference on Computer Vision (ICCV).*

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. 2023. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *IEEE International Conference on Computer Vision (ICCV).*

ChatGPT. 2022. ChatGPT. https://chat.openai.com/chat.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)* (2023).

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV).*

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

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2023. Subject-driven text-to-image generation via apprenticeship learning. In *Conference on Neural Information Processing Systems (NeurIPS).*

Ta-Ying Cheng, Matheus Gadelha, Thibault Groueix, Matthew Fisher, Radomir Mech, Andrew Markham, and Niki Trigoni. 2024. Learning Continuous 3D Words for Text-to-Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Conference on Neural Information Processing Systems (NeurIPS).*

Jiahua Dong and Yu-Xiong Wang. 2023. ViCA-NeRF: View-Consistency-Aware 3D Editing of Neural Radiance Fields. In *Conference on Neural Information Processing Systems (NeurIPS).*

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV).*

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023a. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR).*

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023b. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* (2023).

Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. Expressive text-to-image generation with rich text. In *IEEE International Conference on Computer Vision (ICCV).*

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *IEEE International Conference on Computer Vision (ICCV).*

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *IEEE International Conference on Computer Vision (ICCV).*

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR).*

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS).*

Lukas Höllein, Aljavz Bovzivc, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. 2024. Viewdiff: 3d-consistent image generation with text-to-image models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR).*

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. 2023. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *Conference on Neural Information Processing Systems (NeurIPS).*

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. In *Conference on Neural Information Processing Systems (NeurIPS).*

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. 2023. Analyzing and Improving the Training Dynamics of Diffusion Models. *arXiv preprint arXiv:2312.02696* (2023).

Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. 2011. Rendering synthetic objects into legacy photographs. *ACM Transactions on graphics (TOG)* (2011).

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023).

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. 2023. Lerf: Language embedded radiance fields. In *IEEE International Conference on Computer Vision (ICCV).*

Natasha Kholgade, Tomas Simon, Alexei Efros, and Yaser Sheikh. 2014. 3D object manipulation in a single photograph using stock 3D models. *ACM Transactions on graphics (TOG)* (2014).

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. 2023. Dense Text-to-Image Generation with Attention Modulation. In *IEEE International Conference on Computer Vision (ICCV).*

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR).*

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Dongxu Li, Junnan Li, and Steven CH Hoi. 2023. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Conference* *on Neural Information Processing Systems (NeurIPS)*.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE International Conference on Computer Vision (ICCV)*.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2024. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-nerf for shape-guided generation of 3d shapes and textures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. 2023. Object 3dit: Language-guided 3d-aware image editing. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* (2021).

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* (2022).

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. In *TMLR*.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-level Shape Variations with Text-to-Image Diffusion Models. In *IEEE International Conference on Computer Vision (ICCV)*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *IEEE International Conference on Computer Vision (ICCV)*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *IEEE International Conference on Computer Vision (ICCV)*.

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

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501* (2020).

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Devi Labatut, Patrikh, Yanivck Taigman, and David Novotny. 2021. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023a. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2023b. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949* (2023).

Simo Ryu. 2023. LoRA-Stable Diffusion. https://github.com/cloneofsimo/lora.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. 2023. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994* (2023).

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning (ICML)*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411* (2023).

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2024. Mvdream: Multi-view diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*.

Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. 2023. Total-Recon: Deformable Scene Reconstruction for Embodied View Synthesis. In *IEEE International Conference on Computer Vision (ICCV)*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*.

Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. 2021. Learned initializations for optimizing coordinate-based neural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653* (2023).

Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. 2023. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. $P+$: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).

Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. 2023. Evaluating Data Attribution for Text-to-Image Models. In *IEEE International Conference on Computer Vision (ICCV)*.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *IEEE International Conference on Computer Vision (ICCV)*.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. ReconFusion: 3D Reconstruction with Diffusion Priors. (2024).

Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, and Sergey Tulyakov. 2023. DisCoScene: Spatially Disentangled Generative Radiance Fields for Controllable 3D-aware Scene Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. 2024. DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model. In *International*

Conference on Learning Representations (ICLR).

Shunyu Yao, Tzu Ming Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman, and Josh Tenenbaum. 2018. 3d-aware scene manipulation via inverse graphics. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023b. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

Jianglong Ye, Naiyan Wang, and Xiaolong Wang. 2023a. FeatureNeRF: Learning Generalizable NeRFs by Distilling Foundation Models. In *IEEE International Conference on Computer Vision (ICCV)*.

Jiraphon Yenphraphai, Xichen Pan, Sainan Liu, Daniele Panozzo, and Saining Xie. 2024. Image sculpting: Precise object editing with 3d geometry control. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. In *International Conference on Machine Learning (ICML)*.

Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. 2024. CustomNet: Object Customization with Variable-Viewpoints in Text-to-Image Diffusion Models. In *ACM Multimedia*.

Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. 2024. Cameras as Rays: Sparse-view Pose Estimation via Ray Diffusion. In *International Conference on Learning Representations (ICLR)*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*.

Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. 2021. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *International Conference on Learning Representations (ICLR)*.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Transactions on Graphics (TOG)* (2023).

Zhizhuo Zhou and Shubham Tulsiani. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.