# Alchemist: Parametric Control of Material Properties with Diffusion Models

Prafull Sharma[*,1,2] Varun Jampani[†,2] Yuanzhen Li[2] Dmitry Lagun[2]
Fredo Durand[2] Bill Freeman[1,2] Mark Matthews[2]
[1]MIT CSAIL [2]Google Research
www.prafullsharma.net/alchemist

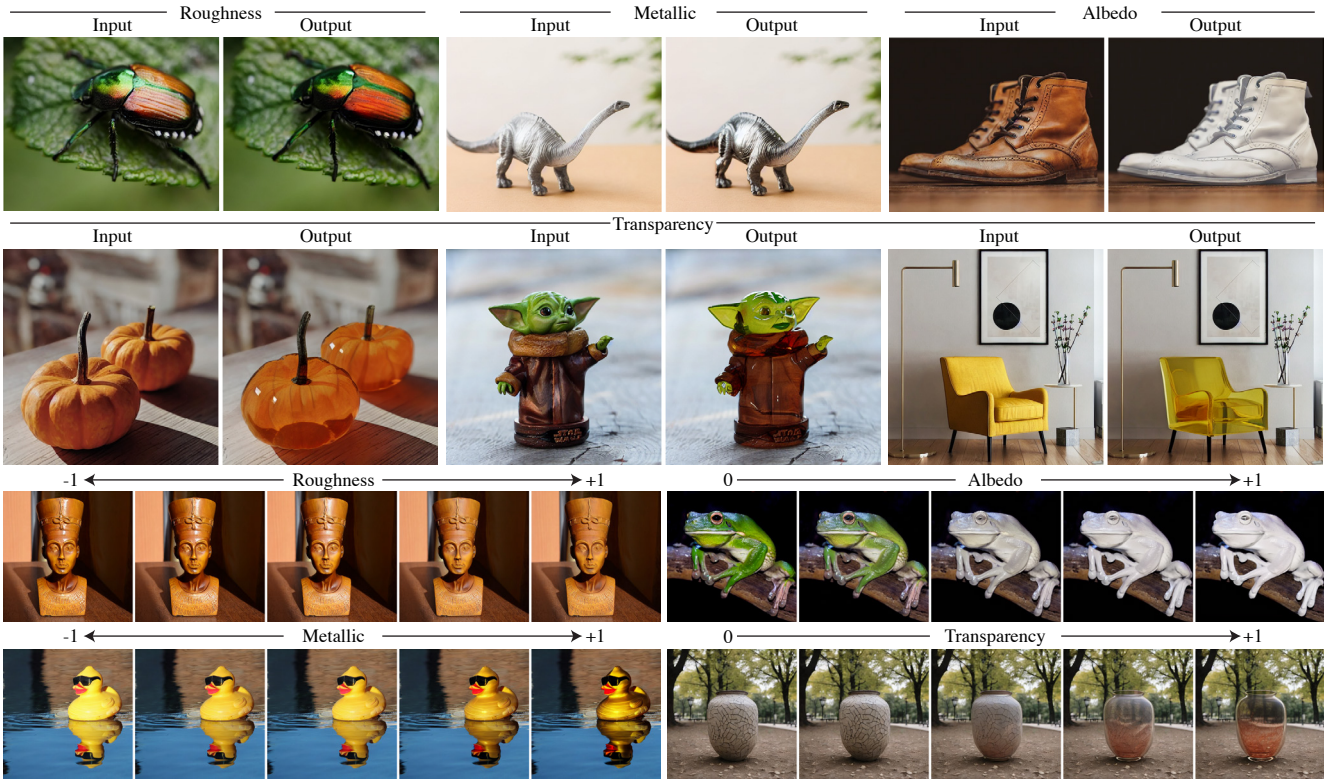


Figure 1. **Overview.** Our method, Alchemist, edits material properties of objects in input images by relative attribute strength $s$. **Top:** We set the strength $s = 1$, resulting in a beetle without specular hightlights, a dark metallic dinosaur, and boot with gray albedo. Our model generates plausible transparency including the light tint, caustics, and hallucinated plausible details behind the object. **Bottom:** We demonstrate smooth edits for linearly chosen strength values.

## Abstract

*We propose a method to control material attributes of objects like roughness, metallic, albedo, and transparency in real images. Our method capitalizes on the generative prior of text-to-image models known for photorealism, employing a scalar value and instructions to alter low-level material properties. Addressing the lack of datasets with controlled material attributes, we generated an object-centric synthetic dataset with physically-based materials. Fine-tuning a modified pre-trained text-to-image model on this synthetic dataset enables us to edit material properties in real-world images while preserving all other attributes. We show the potential application of our model to material edited NeRFs.*

## 1. Introduction

Achieving fine-grained control over material properties of objects in images is a complex task with wide commercial applications beyond computer graphics. This ability is particularly relevant in image editing, advertising, and image forensics. We propose a method for precise editing of material properties in images, harnessing the photorealistic generative prior of text-to-image models. We specifically

*This research was performed while Prafull Sharma was at Google.
†Varun Jampani is now at Stability AI.

target four key material properties: roughness, metallic, albedo, and transparency. Our results illustrate that generative text-to-image models contain a strong understanding of light transport which can be leveraged for precise control of these material properties. The physics of light transport affects the appearance of the object. How we view the objects is an interplay of physical factors such as surface geometry, illumination sources, camera intrinsics, color science, sensor linearity, and tone-mapping. However, the most significant of these factors is material properties.

In computer graphics, Bidirectional Reflectance Distribution Functions (BRDFs) [11,24,25] define material properties which led to the development of principled and physically based BRDF models [4]. Prior methods typically employed an inverse rendering approach to disentangle and estimate complex scene attributes like geometry and illumination for material modification [35]. Recent work by Subias et al. proposed a GAN-based method trained on synthetic data for perceptual material edits, focusing on metallic and roughness parameters, necessitating the masking of the targeted real-world object [79]. Our approach uses the generative prior of text-to-image models. We directly modify real-world images in pixel space, eliminating the need for auxiliary information such as explicit 3D geometry or depth maps, environment maps, and material annotations, thereby bypassing the process of accurately estimating object and scene-level properties.

Manipulating material properties in pixel space using a pre-trained text-to-image model presents two main challenges. First, the scarcity of real-world datasets with precisely labeled material properties makes generalizing from supervised training difficult. Second, text-to-image models are trained with textual descriptions like "gold," "wood," or "plastic," which often lack fine-grained details about the material. This issue is compounded by the inherent disconnect between the discrete nature of words and the continuous nature of material parameters.

To overcome the first challenge, we render a synthetic dataset featuring physically-based materials and environment maps, thus addressing the need for fine-grained annotations of material properties. For the second challenge, we employ extra input channels to an off-the-shelf diffusion model, refining this model with an instruction-based process inspired by InstructPix2Pix [3]. Despite being trained on only 500 synthetic scenes comprising 100 unique 3D objects, our model effectively generalizes the control of material properties to real input images, offering a solution to the issue of continuous control.

To summarize, we present a method that utilizes a pre-trained text-to-image model to manipulate fine-grained material properties in images. Our approach offers an alternative to traditional rendering pipelines, eliminating the need for detailed auxiliary information. The key contributions of our method are as follows:

1. We introduce an image-to-image diffusion model for parametric control of low-level material properties, demonstrating smooth edits of roughness, metallic, albedo and transparency.
2. We render a synthetic dataset of fine-grained material edits using 100 3D objects and randomized environment maps, cameras, and base materials.
3. Our proposed model generalizes to real images despite being trained on synthetic data.

## 2. Related Work

**Diffusion models for image generation.** Denoising Diffusion Probabalistic Models (DDPMs) have been an active focus of the research community [12, 14, 27–29, 32, 33, 77] for their excellent photorealistic image generation capabilities from text prompts [55,64,66,70]. Image-to-image tasks are possible by modifying the denoising network to accept image inputs, allowing style-transfer [76], inpainting, uncropping, super-resolution, and JPEG compression [69]. Furthermore, the generative priors of 2D diffusion models have been utilized towards novel-view synthesis, 3D generation, and stylistic 3D editing [6, 8, 18, 22, 30, 42, 44, 61, 63, 71, 72, 74, 81, 85, 89, 91, 95]. Our image-to-image method leverages and further controls this learned prior of DDPMs.

**Control in generative models.** Controlling generative model output remains an active area of study with many works proposing text-based methods [1, 3, 5, 10, 20, 26, 34, 36, 43, 52, 60, 80, 82, 84]. Other works proposed alternative control inputs such as depth maps, sketches [83, 90], paint strokes [50], identity [47, 88], or photo collections [40,67,68,73]. Prompt-to-Prompt [26], $\mathcal{P}+$ [84], and Null-text inversion [52] present editing techniques based on reweighting of cross-attention maps. ControlNet [92] and T2I-Adapter [54] demonstrate control through spatial inputs defining mid-level information. Generated images from diffusion models can also incorporate new subjects from an image collection using a small number of exemplars [7, 19, 40, 67, 68, 73, 87]. While these works control high and mid-level information about objects, control of low-level properties such as materials remains a challenge for them, leading us to our present line of study.

**Material understanding and editing.** Editing materials in images is a significant challenge, requiring a strong understanding of image formation. Human vision research has extensively explored how attributes like albedo, roughness, illumination, and geometry affect object perception [15, 16, 16, 17, 49, 53, 56–58, 78].

Image based material editing was introduced by Khan et al. presenting simple material operations using depth estimates [35]. Subsequent works demonstrated disentangle-

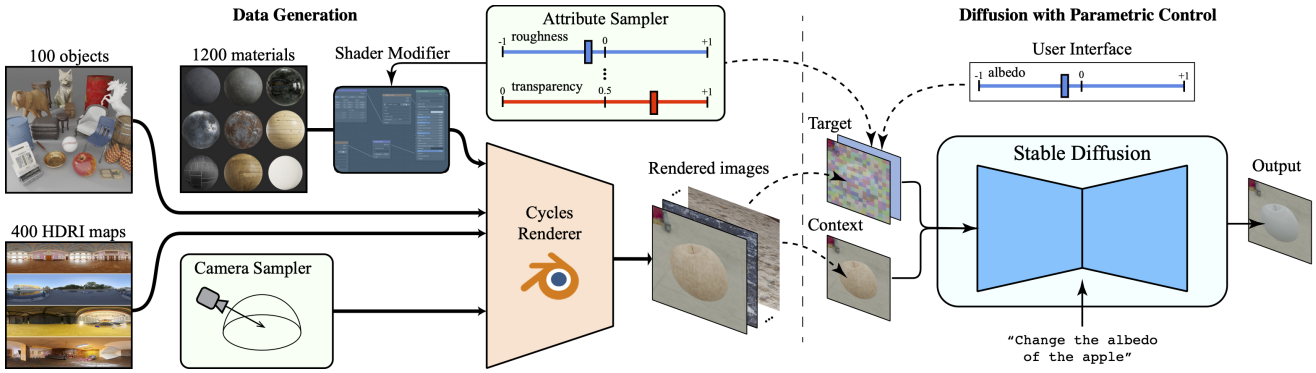


Figure 2. **Method.** We generate a synthetic dataset by taking each of 100 objects, applying randomized materials and illumination maps, and modifying the shading network according to randomly sampled attribute strength $s$. Each object is rendered from 15 randomized cameras (see Section 3 for details). During training we provide the $s = 0$ image as context and randomly choose a target image of known attribute strength. At test time we provide the user-input context image and edit strength.

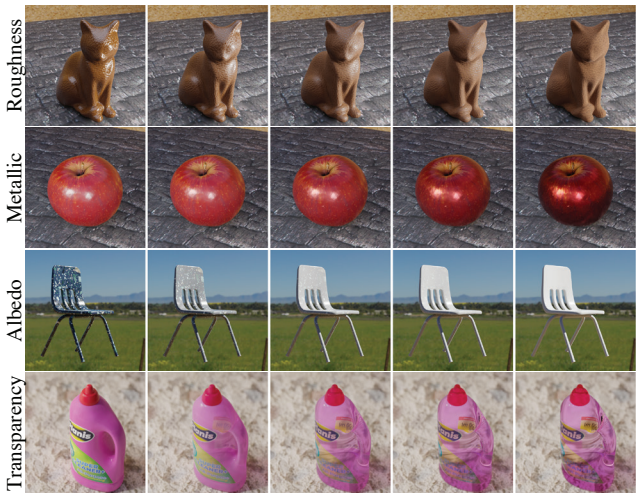


Figure 3. **Synthetic dataset.** Samples from our synthetic dataset illustrating appearance change for a linear attribute change.

ment of material and lighting with a statistical prior [45], editing gloss appearance [2, 48], intrisic image decomposition [41], and 2D editing of material exemplars [96]. We forego these "decompositional" approaches and instead leverage the largely self-supervised prior of DDPMs for direct editing in pixel-space.

Generative models, particularly Generative Adversarial Networks (GANs) [21], have been investigated for their ability to alter material perception, focusing on gloss and metallic properties [13, 79]. The application of semantic and material editing in NeRFs has also been explored using text-prompts and semantic masks [23, 94].

# 3. Method

There is no existing object-centric dataset that precisely varies only single material attributes. Curating such a real world dataset would be infeasible due to the difficulty of creating physical objects in this way with known parameters. Therefore, we opt to render a synthetic dataset, giving us full control of material attributes. Using this data, we propose a method to perform material attribute control given a context image, instruction, and a scalar value defining the desired relative attribute change. The method is based on latent diffusion model for text-to-image generation with modification that allows us to condition the network on the relative attribute strength.

## 3.1. Datasets

We render our dataset with the Cycles renderer from Blender [9], using publicly available 3D assets, physically-based materials, and environment maps. Each scene begins with one of 100 unique object meshes from polyhaven.com. Each of these is paired with five randomly chosen materials of the 1200 available from ambientcg.com, and illuminated with one of 400 environment maps. The material is a Principled BRDF shader node, the base shader in Blender. The base configuration of the material is kept as a control defined as 0 strength change for each of the attributes. This control serves as the context input image to the method against which relative changes in roughness, metallic, albedo, and transparency are applied, sampling 10 random relative values for each attribute, the details of which are described below. Finally, we render 15 images of each setup using different camera viewpoints and intrinsics. This creates a wide combination of scenes with diversity in material, lighting, and background conditions. Samples from the rendered dataset are presented in Figure 3.

**Roughness and Metallic.** For both roughness and metallic properties, we operate in an additive framework. In the case when the material has the associated map for roughness or metallic, we use an additive node yielding a parametric control between [-1, 1]. For materials where either of these spatial maps are missing, we control the attribute control directly as a constant map, assuming the base 0.5 as the control state of the attribute. Note that these values are

clamped between [0, 1] so in some cases, further increasing or decreasing the roughness does not result in any change in the rendered image. We account for this by under-sampling such images where the gradient of change is constant.

Reducing the roughness value results in a surface that reflects light more uniformly and sharply, giving it a glossy or shiny appearance. On the other hand, increasing the roughness value leads to a more diffused light reflection, making the surface appear matte or dull. Low metallic value results in appearance predominantly determined by the base color, as in the case of plastic and wood. Increasing the metallic leads to the surface absorbing more of the incoming light, resulting in a darker appearance of the object.

**Albedo.** We implement a color mixing between the original albedo map of the object and a spatially constant gray (RGB = 0.5) albedo map. The parametric controller operates between 0 and 1, where 0 corresponds to the original albedo, and 1 corresponds to completely gray albedo. This can be considered as detexturing the albedo and can be interesting when combined with roughness and metallic parameters to achieve a mirror-like or a shading-only image.

**Transparency.** We introduce the ability to control transparency by controlling the transmission value in the BSDF shader node. The attribute value is chosen to be in range [0, 1]. For a chosen transmission value t, we choose to reduce the roughness and metallic component in an additive manner by t, and also add a white overlay to the albedo to increase the intensity of the appeared color. For the value of 0, we expect the same opaque object and at 1, we would get a transparent version of the object, making it appear as if it was made of glass. Note that we made the choice to retain the effect of the albedo resulting in a fine tint on the object.

### 3.2. Parametric Control in Diffusion Models

The rendered synthetic data is used to finetune an image-to-image diffusion model conditioned on relative attribute strength and a generic text instruction providing parametric control over material properties. We operate in latent space using Stable Diffusion 1.5, a widely adopted text-to-image latent diffusion model.

Diffusion models perform sequential denoising on noisy input samples, directing them towards the dataset distribution by maximizing a score function [75]. A noising process is defined over timesteps $t \in T$, resulting in a normal distribution at $T$. We operate in latent space by using a pre-trained variational encoder $\mathcal{E}$ and ddecoder $\mathcal{D}$ [38], a potent aid to conditional image generation [66]. Training draws an image sample $\mathbf{I}$ from the dataset, encodes it into a latent $z = \mathcal{E}(\mathbf{I})$, then noises it at $t$ as $z_t$. A denoising network $\epsilon_\theta$ predicts the added noise given the latent $z_t$, diffusion time $t$, and conditioning variables.

Our image-to-image model is conditioned on an input image to be edited, provided as $\mathcal{E}(I_c)$ concatenated to the latent being denoised $z_t$. Text conditioning is provided via cross-attention layers using a generic prompt, $p =$ "*Change the* `<attribute_name>` *of the* `<object_class>`." Since textual CLIP embeddings [62] do not encode fine-grained information well [59], prompt-only conditioning of $s$ expressed textually (i.e. "*Change the roughness of the apple by 0.57.*") yields inconsistent output. To facilitate relative attribute strength conditioning, we also concatenate a constant scalar grid of edit strength $s$.

We initialize the weights of our denoising network with the pre-trained checkpoint of InstructPix2Pix [3], providing an image editing prior and understanding of instructive prompts. During training (Fig. 2), we minimizes the loss:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(\mathbf{I}),\mathcal{E}(I_c),\mathbf{s},p,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\Theta(z_t, t, \mathcal{E}(I_c), \mathbf{s}, p)||_2^2\right] \tag{1}$$

We always provide the $s = 0$ image as context $I_c$ , and draw an edited image $I_e$ at random for noising. Since we always render an $s = 0$ sample, and other $s$ are chosen with stratified sampling, our distribution has a slight bias towards zero. Since many edit strengths may have little effect (i.e. we cannot lower the roughness of an object with 0 roughness), we find that providing too many of these examples biases the network towards inaction. We therefore downweight such null examples, defined as $||I_c - I_e||^2 < \tau$, by $w_{null}$ via rejection sampling. In practice we set $w_{null} = 0.80$, $\tau = 0.05$. We train with `fp16` precision for 10k steps using Adam [37] and learning rate of `5e-5`. We use the text encoders and noise schedule from Stable Diffusion.

At test time we provide a held out image as context $I_c$, edit strength $s$, and prompt $p$ for the object class of the input image. We denoise for 20 steps using the DPM-solver++ based noise scheduler [46].

**Multi-attribute editing.** We edit multiple attributes in a single diffusion pass by concatenating more than one edit strength, drawn from $\{s_a, s_r\, s_m\}$ giving us $[z_t|\mathcal{E}(I_c)|\mathbf{s}]$ as the final UNet input, where $|$ is concatenation.

**Classifier-free guidance.** Ho et al. [29] proposed classifier-free guidance (CGF) to improve visual quality and faithfulness of images generated by diffusion models. We retain the same CFG setup as InstructPix2Pix for both image and prompt conditioning. We do not however impose CFG with respect to the relative attribute strengths $\mathbf{s}$. We want the network to be faithful to edit strength and forced to reason about incoming material attributes. As $s$ can be 0 by definition of the problem itself, and downweighted as described above, we did not find CFG on $\mathbf{s}$ necessary.

We will release the dataset generation pipeline, image renderings with metadata, and the training code.

## 4. Results

We present qualitative analysis demonstrating the generalization capability of our model to real images despite

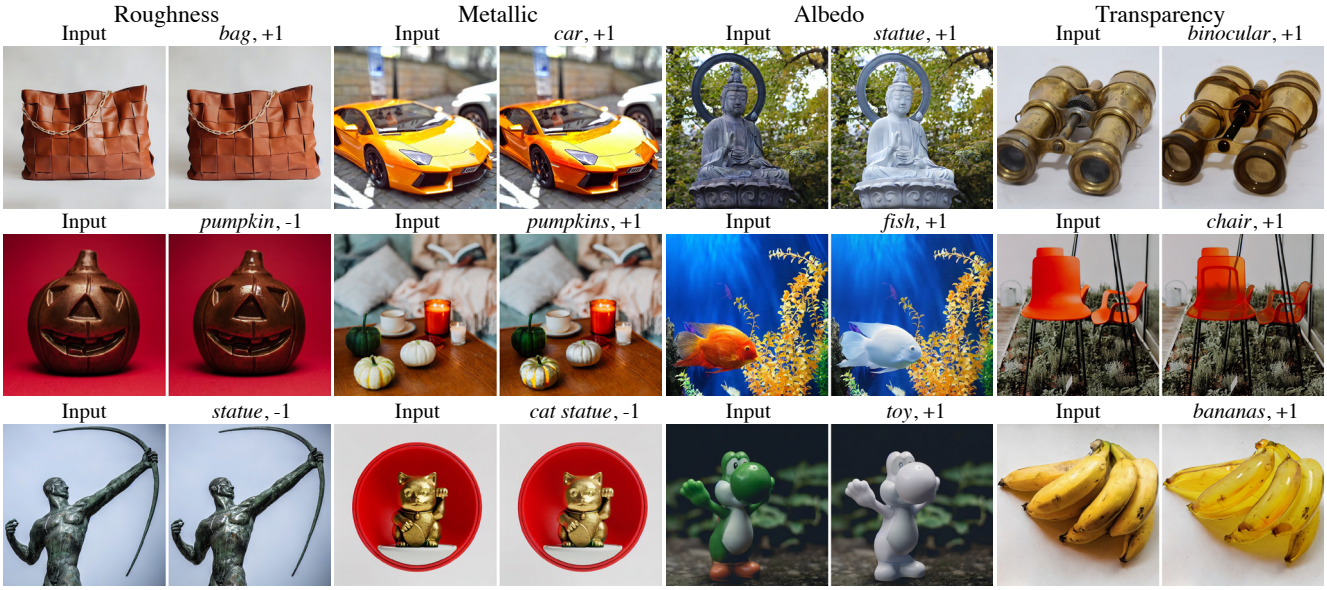


Figure 4. **Single-attribute editing results.** Outputs from our model trained on individual attributes. Left of columns are held-out input and right are model output (*object class*, $s$). Increased "Roughness" replaces specular highlights on the bag with base albedo. "Metallic" varies contributions from albedo and shine in regions of the pumpkins and cat. Surfaces are changed to a flat grey "Albedo" revealing object illumination. "Transparency" preserves object tint while inpainting background and hallucinating plausible hidden structures and caustics.

being trained on synthetic data. Comparisons to baselines show the effectiveness of our model for fine-grained material editing, further supported by a user study. We extend the use of our model to NeRF material editing on the DTU dataset [31, 51].

### 4.1. Results on Real Images

We demonstrate the effectiveness of our technique through editing material attributes, one at a time, for real unseen images, in Figure 4. For each of the material attributes we use a separate model trained only on that attribute. We observe that the model outputs preserve geometry and take the global illumination into account.

**Roughness.** As the roughness is increased, the output shows removal of the specular highlights replaced by estimate of the base albedo. The highlights are amplified when the roughness is reduced as shown in the case of the pumpkin and statue.

**Metallic.** The increase in the metallic component of the car and pumpkin results in dampening of the base albedo and increase in the shine on the surface. The effect is reverse for the cat statue when the metallic strength was reduced. Our method shows similar behavior to the Principled BRDF shaders, which present perceptually subtle effects when tuning the metallic value.

**Albedo.** As the relative strength for the albedo is turned to 1, we observe the albedo of the Buddha statue, fish, and toy go to gray. This is not a trivial in-image desaturation operation as the network maintains the highlights, shadows, and the light tint from the plausible environment map.

**Transparency.** The transparent renditions of the binocular and the chair demonstrate the prior over 3D geometry of the objects, using which it generates the appropriately tinted glass-like appearance and in-paints background objects. With the edit of the banana, we can see the caustics underneath and the preservation of the specular highlights.

### 4.2. Baseline Comparisons

We compare our method, Alchemist, to the GAN-based in-image material editing of Subias et al. [79], Prompt-to-Prompt [26] with Null-text Inversion (NTI) [52], and InstructPix2Pix [3] in Figure 5. Furthermore, we fine-tuned the InstructPix2Pix prompt-based approach with our synthetic dataset. Subias et al.'s method results in exaggerated material changes as their objective is perceptual, not physically-based material edits. Null-text inversion and InstructPix2Pix change the global image information instead of only the object of interest: lighting changes for roughness and albedo edits, or a geometry change for metallicity edit. When InstructPix2Pix is trained on our dataset with a prompt-only approach, we observe the model exaggerating the intended effect, yielding artifacts on the panda for metallic and the water for the transparency case. The model also changes the albedo of the sand when only asked to make the change to the crab. Our method faithfully edits only the object of interest, introducing the specular hightlights on the leg of the cat statue, dampening the albedo for the metallic panda, changing the albedo of the crab to gray while retaining the geometry and illumination effects, and turning the dolphin transparent with plausible refractive effects.

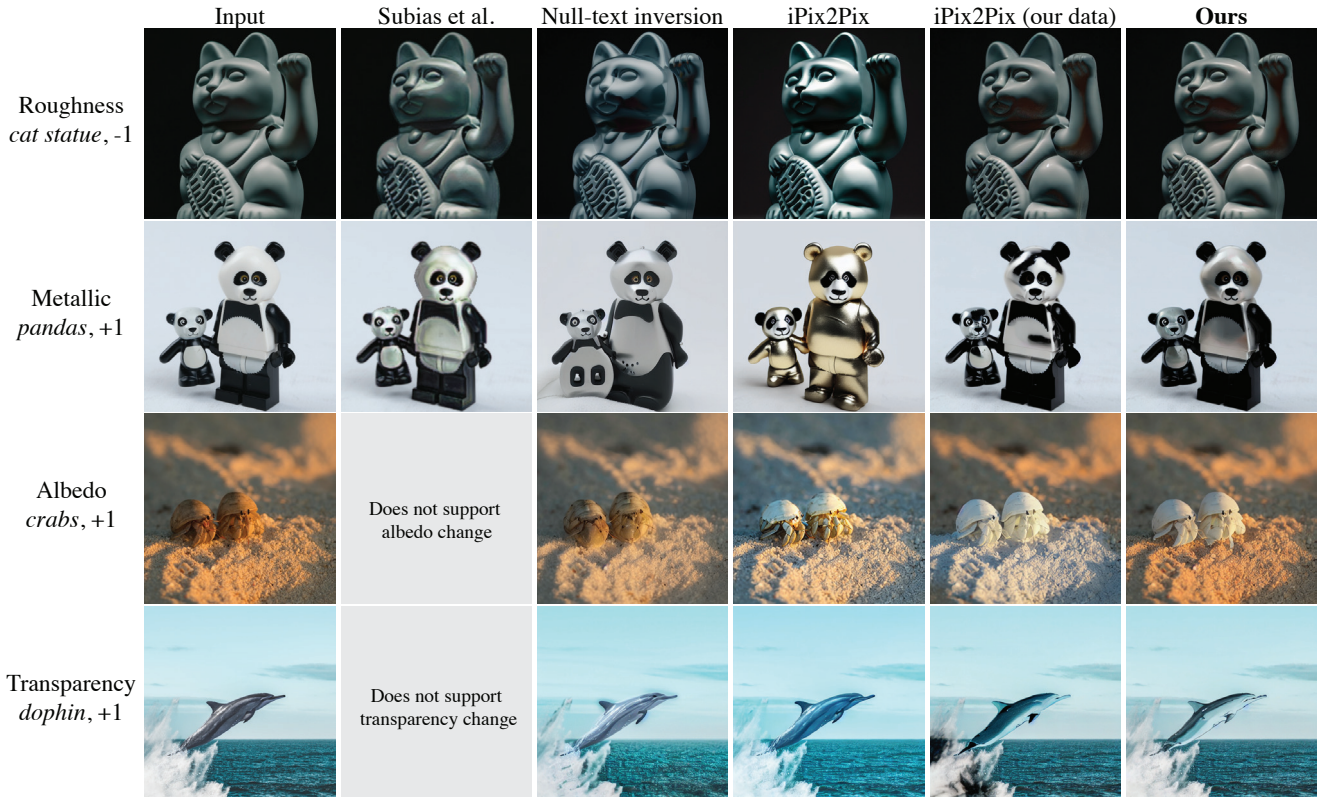


Figure 5. **Qualitative comparison.** Comparison of Alchemist with baseline methods. We increase each of the attributes shown on the left.

| | InstructPix2Pix w/ our data | | | **Our Method** | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| Roughness | 30.9 | 0.89 | 0.13 | 31.5 | 0.90 | 0.09 |
| Metallic | 31.0 | 0.89 | 0.10 | 31.1 | 0.89 | 0.09 |
| Albedo | 26.9 | 0.88 | 0.14 | 27.2 | 0.88 | 0.10 |
| Transparency | 26.9 | 0.85 | 0.13 | 27.1 | 0.85 | 0.13 |

Table 1. **Quantitative analysis.** Metrics for the prompt-only InstructPix2Pix trained on our data and our proposed method computing the PSNR, SSIM [86], and LPIPS [93] on a held-out unseen synthetic rendered dataset of 10 scenes with 15 cameras.

Specific configurations for each baseline is presented in the supplement.

**Quantitative Comparisons.** Due to the lack of an existing dataset for quantitative analysis, we rendered 10 held-out scenes with unseen 3D objects, materials, and environment maps. Each scene was rendered from 15 different viewpoints with 10 random scalar values for each attribute. We present the average PSNR, SSIM [86], and LPIPS [93] of edits against GT for prompt-only InstructPix2Pix and Alchemist in Table 1. While the PSNR and SSIM scores are quite close, our model does better in terms of the LPIPS score. We also note that prompt-only InstructPix2Pix fine-tuned on our data does not yield smooth transitions as the relative strength is linearly changed, visible in video results presented in the supplement. Note that the image reconstruction metrics are not commonly used for evaluation of probabilistic generative models. Samples of the test data and model outputs are presented in the supplement.

**User study.** To further the comparison between the baseline and our method, we conducted a user study presenting N=14 users with pairs of edited images. For each image pair, the users were asked to choose between the two based on: (1) *Photorealism*, and (2) *"Which edit do you prefer?"*. For both questions, the users were presented with the instruction, i.e. for transparency, the instruction was stated as "the method should be able to output the same object as transparent retaining the tint of the object". Each user was presented with a total of 12 image pairs (3 image results for each of the 4 attributes).

Our method was chosen as the one with more photo-realistic edits (**69.6%** vs. 30.4%) and was strongly preferred overall (**70.2%** vs. 29.8%). This is likely due to the apparent exaggeration exhibited by InstructPix2Pix trained on our data with prompt-only approach, leading to saturated effects making it less photo-realistic.

**Smoothness in Parametric Control.**

We demonstrate that our model achieves fine grained control of material parameters by linearly varying the strength of a single attribute, as shown in Figure 6. Observe that the model generates plausible specular highlights on the headphone instead of naively interpolating pixel values to the extrema and introduces more dampening of the

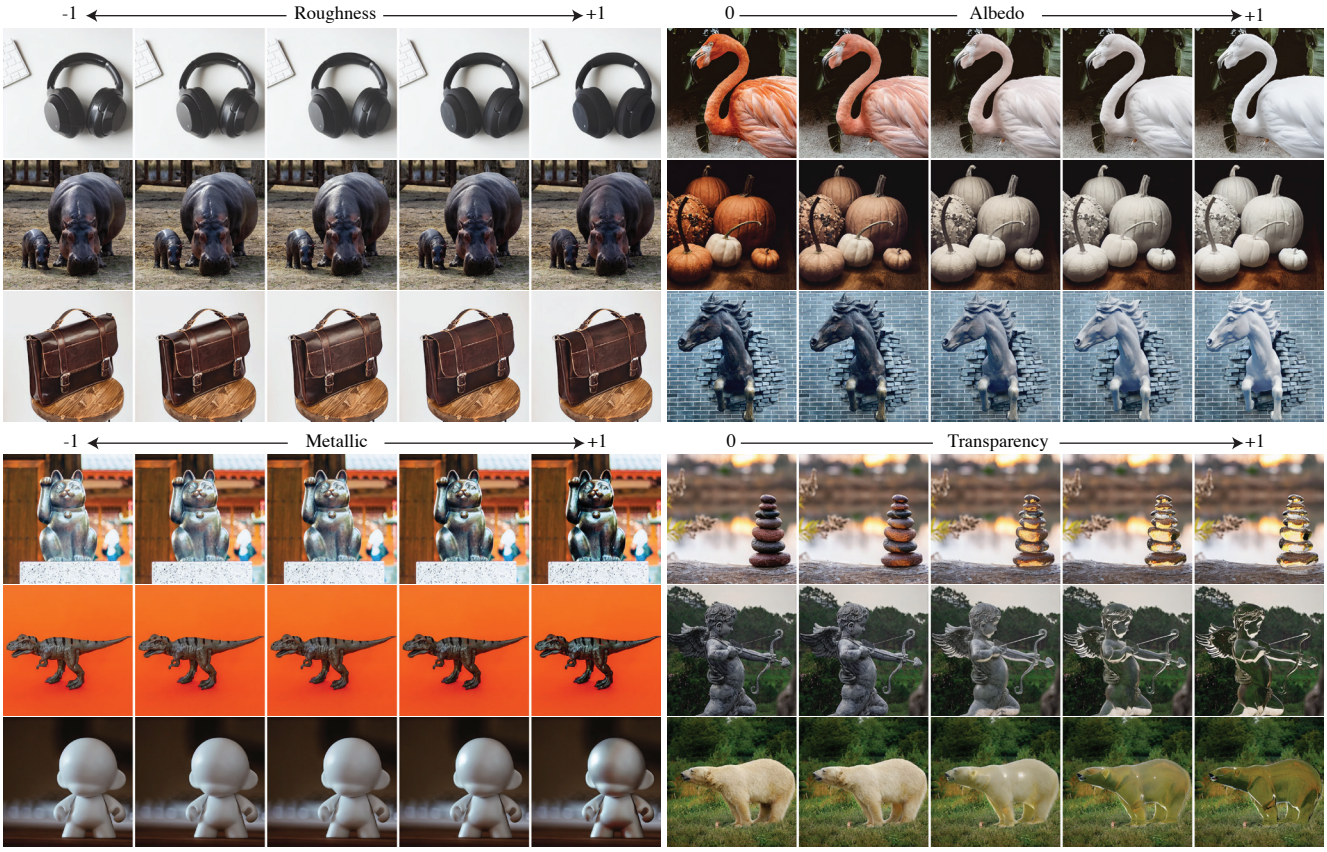


Figure 6. **Slider results.** Alchemist produces edits smoothly with attribute strength. We note that the outputs for linear change in the input relative strength in InstructPix2Pix prompt-only trained on our data results in non-smooth transitions. Refer to the supplement videos.

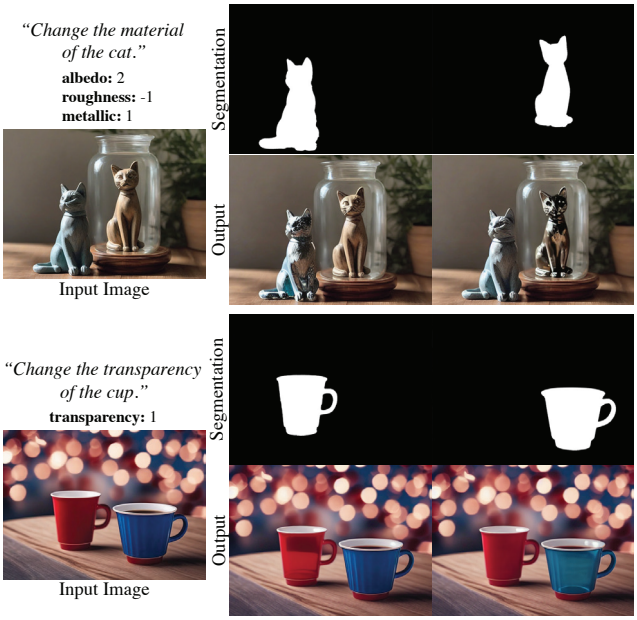


Figure 7. **Spatial Localization.** Edit results (**bottom**) when the scalar strength input is masked by the shown segmentation (**top**). The image is only altered in the segmented region, becoming either shiny (*cat*), or semi-transparent (*cup*).

albedo on the cat to give it a metallic look. For transparency, the model preserves the geometry while refracting the light through the object to produce a transparent look. The InstructPix2Pix model trained on our data did not yield such smooth results as the relative strength of the attributes were changed in text format. Please refer to the supplementary for video results.

### 4.3. Specializations

**Spatial localization.** Attending to a specific instance of a class when multiple objects of the same class are present in the image is a difficult task using just language instruction. We explore the possibility of changing the material attribute of a specific instance of the class by only attributing the scalar value in the segmented region, assuming a known segmentation map from an instance segmentation method such as Segment Anything [39]. Though the network was not trained for this specific task, we find that the network *does* respect the localization of the relative attribute strength, though requires over-driving to values beyond 1. We observe that mask-based editing works in such cases, i.e. changing the material properties of specific instance of cat and cup, as shown in Figure 7.

**Multi-attribute changes.** To enable control over multiple

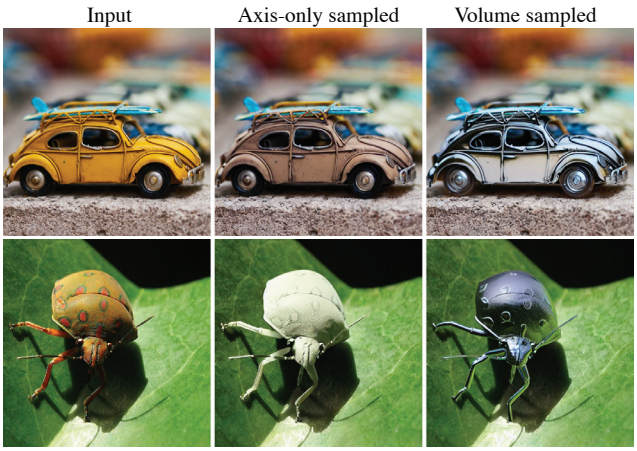


Figure 8. **Multi-Attribute Editing.** Comparison between an "axis-only sampled" model trained on images where only one of $\{s_a, s_r, s_m\}$ is $\neq 0$, vs. a "volume sampled" one where all $\{s_a, s_r, s_m\}$ may be $\neq 0$. We show edits with $(s_a, s_r, s_m) = (1, -1, 1)$. The former tends to edit only a single attribute, while the latter successfully achieves the desired "silver" appearance.

attributes in a single diffusion pass, we train our network on two versions of the dataset to be able to vary albedo ($s_a$), roughness ($s_r$), and metallic ($s_m$). In the *axis-only sampled* version, we keep the context image at the baseline, and vary a single attribute at a time, such that only one of $\{s_a, s_r, s_m\}$ is non-zero for any given training target image. In the *volume sampled* version, $\{s_a, s_r, s_m\}$ are all allowed to be non-zero, effectively sampling the 3D volume of material attributes. In both data-sets, we keep the number of input images the same.

We present the qualitative analysis of the joint control in Figure 8. We find that the "one attribute at a time" model fails to compose the three attributes, generally showing bias towards one of the attributes. The model trained on the volume of these attributes successfully generalizes, demonstrating excellent ability to edit multiple attributes at once. We find this essential to producing a strong metallic appearance on objects, as the Principled BSDF shader requires a white, non-rough, and highly metallic surface to produce this look.

**Material editing of NeRFs.** We test the efficacy of per-frame editing using our method for two-step material control in neural radiance field (NeRF) reconstruction. We use a selection of scenes from the DTU MVS [31] dataset and edit them to have reduced albedo or higher specular reflections. We train a NeRF with the vanilla configuration based on [65] (complete details in the supplement).

In the results presented in Figure 9, we observe highly plausible renderings from held-out views showing 3D structure with the intended albedo, roughness, and metallic change. Please refer to the supplement for rendered video of NeRFs trained on edited data.

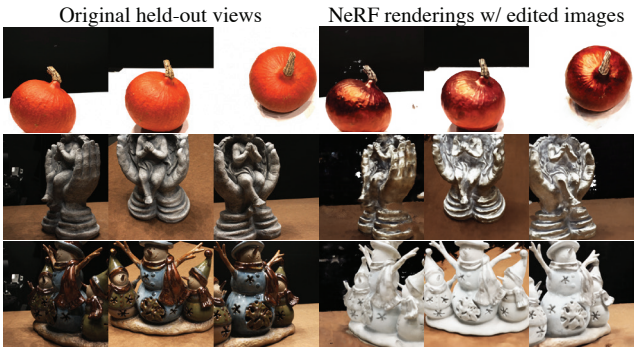


Figure 9. **NeRF Results. Left:** Original test views from DTU. **Right:** We edit training views of each scene, train a NeRF, then render held-out test views. The respective edits $(s_a, s_r, s_m)$ from top to bottom are: *scan30*: $(0, -0.5, 0.5)$, *scan118*: $(1, -1, 1)$ and *scan69*: $(1, 1, 0)$.


Figure 10. **Limitations.** Alchemist sometimes fails to achieve the desired result. **Left:** A shiny surface remains on the teapot after a roughness edit. **Right:** The stem of a candy-cane is omitted.

## 5. Discussion

Our model generalizes to editing fine-grained material properties in real images, despite being trained solely on synthetic data. We believe that our method could extend to a wide range of material alterations achievable with a shader. However, our approach does have limitations, such as producing minimal perceptual changes for roughness and metallic attributes, and occasionally yielding physically unrealistic transparency, as illustrated in Figure 10. The model lacks a complete 3D world model and is unable inpaint to maintain physical consistency as seen in the candy-cane example. As is typical with generative models, our method generates plausible interpretations that are true to the given instructions, but it does not necessarily replicate the exact outcomes of a traditional graphics renderer.

## 6. Conclusion

We present a method that allows precise in-image control of material properties, utilizing the advanced generative capabilities of text-to-image models. Our approach shows that even though the model is trained on synthetic data, it effectively edits real images, achieving seamless transitions as the relative strength of the desired attribute is varied. Beyond image editing, we demonstrate the applicability to NeRF allowing for editable materials in NeRFs. We believe that our work can further impact downstream applications and allow for improved control over low-level properties of objects.

## 7. Acknowledgements

We would like to thank Forrester Cole, Charles Herrmann, Junhwa Hur, and Nataniel Ruiz for helpful discussions. Thanks to Shriya Kumar and Parimarjan Negi for proofreading the submission.

## References

[1] 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

[2] Ivaylo Boyadzhiev, Kavita Bala, Sylvain Paris, and Edward Adelson. Band-sifting decomposition for image-based material editing. *ACM Transactions on Graphics (TOG)*, 34(5):1–16, 2015. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 4, 5

[4] Brent Burley. Physically based shading at disney. In *ACM SIGGRAPH 2012 Courses*. ACM, 2012. 2

[5] 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. *arXiv preprint arXiv:2304.08465*, 2023. 2

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 2

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2

[8] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis. *arXiv preprint arXiv:2308.11473*, 2023. 2

[9] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3

[10] Yuren Cong, Martin Renqiang Min, Li Erran Li, Bodo Rosenhahn, and Michael Ying Yang. Attribute-centric compositional text-to-image generation. *arXiv preprint arXiv:2301.01413*, 2023. 2

[11] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982. 2

[12] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 2

[13] Johanna Delanoy, Manuel Lagunas, J Condor, Diego Gutierrez, and Belén Masia. A generative framework for image-based editing of material appearance using perceptual attributes. In *Computer Graphics Forum*, volume 41, pages 453–464. Wiley Online Library, 2022. 3

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[15] Katja Doerschner, Huseyin Boyaci, and Laurence T Maloney. Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision*, 10(4):8–8, 2010. 2

[16] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014. 2

[17] Roland W Fleming, Ron O Dror, and Edward H Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of vision*, 3(5):3–3, 2003. 2

[18] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *ArXiv*, abs/2302.01133, 2023. 2

[19] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[20] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[22] 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. 2

[23] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[24] Xiao D He, Patrick O Heynen, Richard L Phillips, Kenneth E Torrance, David H Salesin, and Donald P Greenberg. A fast and accurate light reflection model. *ACM SIGGRAPH Computer Graphics*, 26(2):253–254, 1992. 2

[25] Xiao D He, Kenneth E Torrance, Francois X Sillion, and Donald P Greenberg. A comprehensive physical model for light reflection. *ACM SIGGRAPH computer graphics*, 25(4):175–186, 1991. 2

[26] 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. 2, 5

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[28] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[29] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 4

[30] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *ArXiv*, abs/2303.11989, 2023. 2

[31] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 5, 8

[32] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2

[33] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 2

[34] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[35] Erum Arif Khan, Erik Reinhard, Roland W Fleming, and Heinrich H Bülthoff. Image-based material editing. *ACM Transactions on Graphics (TOG)*, 25(3):654–663, 2006. 2

[36] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[38] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[39] 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. 7

[40] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[41] Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. Material editing using a physically based rendering network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2261–2269, 2017. 3

[42] 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. *ArXiv*, abs/2303.11328, 2023. 2

[43] Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *European Conference on Computer Vision*, pages 89–106. Springer, 2020. 2

[44] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[45] Stephen Lombardi and Ko Nishino. Reflectance and natural illumination from a single image. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 582–595. Springer, 2012. 3

[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 4

[47] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2

[48] Yusuke Manabe, Midori Tanaka, and Takahiko Horiuchi. Glossy appearance editing for heterogeneous material objects. *Journal of Imaging Science and Technology*, 65(6):60406–1, 2021. 3

[49] Phillip J Marlow, Juno Kim, and Barton L Anderson. The perception and misperception of specular surface reflectance. *Current Biology*, 22(20):1909–1913, 2012. 2

[50] 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. *arXiv preprint arXiv:2108.01073*, 2021. 2

[51] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[52] 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. 2, 5

[53] Isamu Motoyoshi and Hiroaki Matoba. Variability in constancy of the perceived surface reflectance across different illumination statistics. *Vision Research*, 53(1):30–39, 2012. 2

[54] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[55] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[56] Shin'ya Nishida and Mikio Shinya. Use of image-based information in judgments of surface-reflectance properties. *JOSA A*, 15(12):2951–2965, 1998. 2

[57] Gaë Obein, Kenneth Knoblauch, and Françise Viéot. Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of vision*, 4(9):4–4, 2004. 2

[58] Maria Olkkonen and David H Brainard. Perceived glossiness and lightness under real-world illumination. *Journal of vision*, 10(9):5–5, 2010. 2

[59] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023. 4

[60] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 2

[61] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[63] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023. 2

[64] 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. 2

[65] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2022. 8

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 4

[67] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2

[68] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[69] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[70] 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. 2

[71] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects, 2023. 2

[72] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 2

[73] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[74] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 2

[75] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 4

[76] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2

[77] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[78] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021. 2

[79] J Daniel Subias and Manuel Lagunas. In-the-wild material appearance editing using perceptual attributes. In *Computer Graphics Forum*, volume 42, pages 333–345. Wiley Online Library, 2023. 2, 3, 5

[80] Ming Tao, Bing-Kun Bao, Hao Tang, Fei Wu, Longhui Wei, and Qi Tian. De-net: Dynamic text-guided image editing adversarial networks. *arXiv preprint arXiv:2206.01160*, 2022. 2

[81] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. *ArXiv*, abs/2304.12439, 2023. 2

[82] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2

[83] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIG-*

*GRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[84] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[85] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 2

[86] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[87] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[88] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[89] Jiale Xu, Xintao Wang, Yan-Pei Cao, Weihao Cheng, Ying Shan, and Shenghua Gao. Instructp2p: Learning to edit 3d point clouds with text instructions. *arXiv e-prints*, 2023. 2

[90] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

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

[92] 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. 2

[93] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[94] Xingchen Zhou, Ying He, F. Richard Yu, Jianqiang Li, and You Li. Repaint-nerf: Nerf editting via semantic masks and diffusion models. *arXiv e-prints*, 2023. 3

[95] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. *SIGGRAPH Asia*, 2023. 2

[96] Károly Zsolnai-Fehér, Peter Wonka, and Michael Wimmer. Photorealistic material editing through direct image manipulation. In *Computer Graphics Forum*, volume 39, pages 107–120. Wiley Online Library, 2020. 3