ZeST: Zero-Shot Material Transfer from a Single Image

Ta-Ying Cheng[1,2], Prafull Sharma[3], Andrew Markham[1], Niki Trigoni[1], and Varun Jampani[2]

[1]University of Oxford [2]Stability AI [3]MIT CSAIL

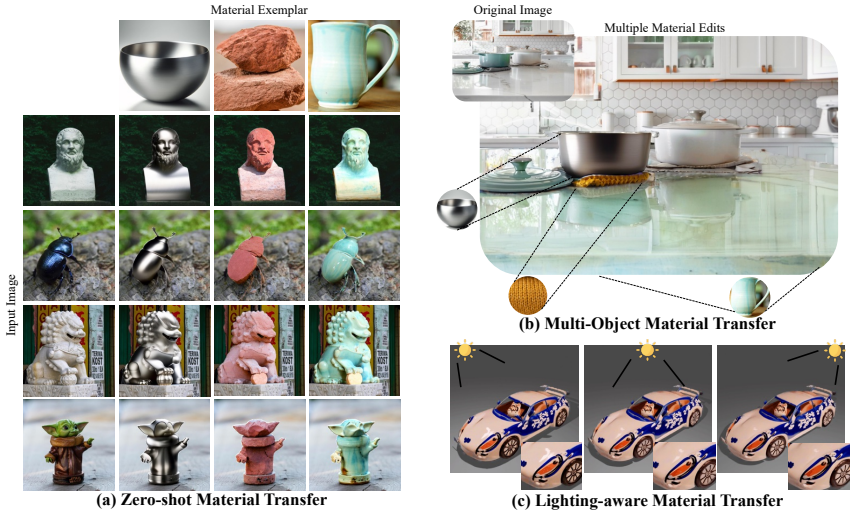

Fig. 1: Overview. We present *ZeST*, a zero-shot single-image approach to **(a)** transfer material from an examplar image to an object in the input image. **(b)** *ZeST* can easily be extended to perform multiple material edits in an single image, and **(c)** perform implicit lighting-aware edits on rendering of a textured mesh.

Abstract. We propose *ZeST*, a method for zero-shot material transfer to an object in the input image given a material exemplar image. *ZeST* leverages existing diffusion adapters to extract implicit material representation from the exemplar image. This representation is used to transfer the material using pre-trained inpainting diffusion model on the object in the input image using depth estimates as geometry cue and grayscale object shading as illumination cues. The method works on real images without any training resulting a zero-shot approach. Both qualitative and quantitative results on real and synthetic datasets demonstrate that *ZeST* outputs photorealistic images with transferred materials. We also show the application of *ZeST* to perform multiple edits and robust material assignment under different illuminations.
Project Page: https://ttchengab.github.io/zest

1 Introduction

Editing object materials in images (*e.g.,* changing a marble statue into a steel statue) is useful for several graphics and design applications such as game design, e-commerce, etc. It is a highly challenging and time-consuming task even for expert artists and graphic designers – typically requires explicit 3D geometry and illumination estimation followed by careful tuning of the target material properties (*e.g.,* metallic, roughness, transparency). Previous works try to alleviate the tedious material specification by synthesizing textures given input text prompts [39,50]. However, they are focused on texturing 3D meshes, which overlooks some of the unique challenges for material editing in 2D images, such as illumination. Another work [41] proposes fine-grained material editing on images, but it cannot directly transfer materials from a given exemplar.

In this work, we aim to make 2D-to-2D material editing practical by eliminating the need for any 3D objects as well as explicit specification of material properties. Given a single image of an object and another material exemplar image, our goal is to transfer the material appearance from the exemplar to the target object directly in 2D. See Fig. 1 for some sample input and material exemplar images. We do not assume any access to the ground-truth 3D shapes, illumination, or even the material properties, making this problem setting practical and widely applicable for material editing.

This setup is particularly challenging from two perspectives. First, an explicit approach to material transfer requires an understanding of many object-level properties in both the exemplar and the input image, such as geometry and illumination. Subsequently, we have to disentangle the material information from these properties and apply it to the new image; the entire process has several unsolved components. Second, there currently exists no real-world datasets for supervising this task. Collecting high-quality datasets presenting the same object with multiple materials and exemplars may be quite tedious.

One of the main contributions of this work in alleviating these challenges is a zero-shot approach that can implicitly transfer arbitrary material appearances from a given 2D exemplar image onto a target 2D object image, without explicitly estimating any 3D or material properties from either image. We call our approach '*ZeST*', as it does not require multiple exemplars or any training like previous works, making it easy to generalize to any images in the wild.

With *ZeST*, we propose a carefully designed pipeline that repurposes several recent advances in 2D image generation and editing for our problem setting. At a high level, we adapt the geometry-guided generation (*e.g.,* ControlNet [51]) and also exemplar-guided generation (*e.g.,* IP-Adapter [49]) to implicitly isolate and transfer material appearance from a source exemplar to the target image while applying a foreground decolored image and inpainting for illumination cues. Our key contribution is presenting a simple pipeline with careful design choices that can be used to tackle a highly challenging problem of 2D-to-2D material transfer.

Since this is a new problem setting, we created both synthetic and real-world evaluation datasets with material exemplars and object images. Extensive qualitative and quantitative evaluations demonstrate that *ZeST* excels in photo-

realism and material accuracy in the output images when compared against various baselines while being completely training-free. See Fig. 1(a) for sample results of *ZeST*. With our pipeline, artists can grab pre-designed materials as material exemplars and directly transfer them to real-world images. By using different object masks, we can also use *ZeST* to cast different materials to multiple objects present in a single image (Fig. 1 (b)). In addition, with slight alteration of the inputs, *ZeST* can perform light-aware material transfer by changing the reflections while keeping textural patterns consistent (Fig. 1 (c)); this method can have potential application when used in conjunction with 3D texture generation methods [10].

In summary, *ZeST* has several favorable properties for material editing:

- **Zero-shot, training free, single-image material transfer.** By leveraging 2D generative priors, *ZeST* works in a zero-shot manner without needing dataset finetuning. Unlike some contemporary works [50] that implicitly capture material properties using several material images, *ZeST* only needs a single material exemplar image to transfer the material in pixel space.
- **No explicit 3D, illumination or materials.** With 2D depth and segmentation estimation (which are readily available these days) and implicit material transfer, we eliminate the need for explicit specification of 3D meshes, illumination or material properties (say, in terms of BRDF).
- **Several downstream applications.** Given the simplistic and practical nature of our approach, *ZeST* can be used for several downstream graphics applications such as applying pre-designed materials to real-world images, editing multiple object materials in a single image, and perform lighting-aware material transfer given untextured mesh renderings.

2 Related Work

Diffusion Models. Denoising Diffusion Probabilistic models have emerged as the state-of-the-art for class-conditional and text-prompt conditioned image generation [18, 23–27, 43]. These models generate photorealistic images with exemplary geometry, materials, illumination, and scene composition. The models have been extended to be conditioned on input images for computational photography tasks such as super-resolution, style transfer, and inpainting.

Further work demonstrate controllable generation conditioned on text-based instructions [8,20,22,46], semantic segmentation [4], bounding box [11,30,47,48], depth [6, 53], sketch [34, 51], and image prompt [49]. Prompt-to-prompt and Prompt+ edit the input image by performing inversion followed by the introduction of new terms and reweighting the effect of terms in the input prompt [22,46]. InstructPix2Pix performs edits an input image conditioned on an instruction [7]. Ge et al. proposed rich text based image editing allowing for style assignment and specific description to specific terms in the prompt [20]. While these methods edit the image semantically and high-level descriptions, assigning specific materials using text-based approach is challenging since text acts as a limiting modality for describing textures.

A collection of reference images can be used to learn concepts which can be further included in text prompts to generate images with the learned concepts [12, 29, 40]. Spatial modalities such as depth and sketches have been used for controlling the generated images [34, 49, 51]. Pre-trained text-to-image models can be leveraged for 3D-aware image editing using language and depth cues [13, 33, 35]. The use of ControlNet has been extended by Bhat et al. to use depth for controlling the scene composition while maintaining other scene attributes [6]. Object orientation, illumination, and other object attributes can be controlled in a continuous manner using ControlNet and learned continuous tokens embedding the 3D properties [13].

Material acquisition and editing. Material acquisition and editing is an active field of research taking into account illumination and object geometry. Previous work has demonstrated material acquisition under known illumination conditions and camera [2,3,17]. Such acquisition in the wild requires localizing objects with similar materials, which has been facilitated by supervised material segmentation and leveraging pre-trained vision representation backbones [5, 31, 42, 45]. Khan et al. introduced in-image material editing using estimates of depth [28]. Recent works have employed generative adversarial networks [21] for perceptual material editing [16, 44] and physical shader-based editing using text-to-image models [41]. The use of generative models has been extended to explicitly learning materials [32] and texturing 3D meshes [9, 10, 39, 50].

In our work, we aim to use pre-trained image generation diffusion models to perform exemplar-based material transfer from a single image. We aim to use ControlNet and IP-adapter to perform material transfer in a zero-shot way without any training.

3 Method

In this section, we describe our method *ZeST* that performs exemplar-based material transfer. Recent methods perform the related problem of texture synthesis on meshes [39, 50] by finetuning a diffusion model on 3-5 material exemplar images to capture the texture/material in the latent space. On the contrary, *ZeST* only requires a single material exemplar image and a single input image, accomplishing material transfer in a zero-shot, training-free manner.

3.1 Problem Setting

Given a material exemplar image M and an input image I, we aim to output an edited image I_{gen} from I by transferring the material from the material exemplar to the object in the input image while preserving other object and scene properties (e.g. object geometry, background, lighting etc.). Performing this task requires understanding the material, geometry, and illumination from both the exemplar and the input image.

In practice, estimating all the aforementioned object-level properties and further isolating material information explicitly from M is challenging since these

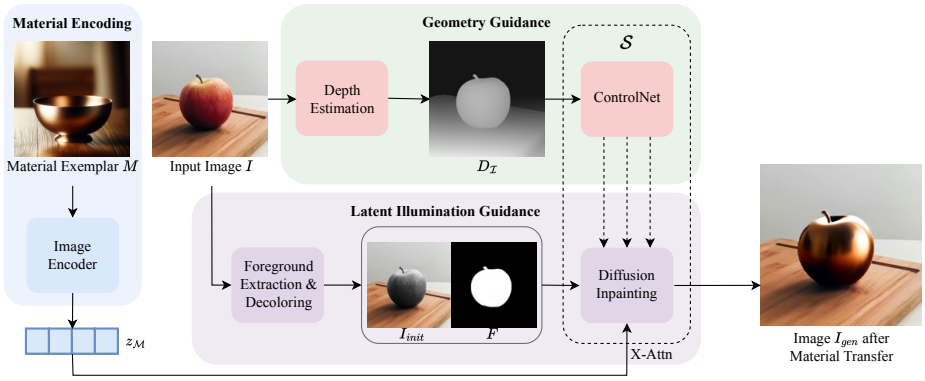

Fig. 2: *ZeST* Architecture. Given a material exemplar M and an input image I, we first encode material exemplar with an image encoder (*e.g.,* IP-Adaptor). Concurrently, we convert the input image into a depth map D_I and a foreground-grayscaled image $I_{init}$to feed into the geometry and latent illumination guidance branch, respectively. By combining the two sources of guidance with the latent features from the material encoding, *ZeST* can transfer the material properties onto the object in input image while preserving all other attributes.

properties are entangled in the pixel space. Therefore, we propose to tackle this problem in the latent space of diffusion models. Specifically, we aim to extract a latent representation z_M containing the material and texture information that we can then inject into a generative diffusion model $\mathcal{S}$ to generate I_{gen}.

3.2 *ZeST* Overview

Since there exists no synthetic/real image dataset to supervise the learning of a 2D-to-2D material transfer, we perform the material transfer in a zero-shot training-free manner. We first break down this complex task into sub-problems of (1) encoding the material exemplar, (2) geometry-guided image editing, and (3) making the generation process illumination-aware. Given the recent advances in high-fidelity diffusion models and complementary adapters for image generation, we leverage existing pre-trained modules to tackle each of the sub-problems that together compose our pipeline to perform image-prompted material editing.

Figure 2 presents an overview of our pipeline, which comprises three branches to guide the material, geometry, and lighting information, respectively. The Material Encoding branch takes the material exemplar image M as input, which is processed by the image encoder to obtain a material latent representation z_M.

Concurrently, we feed the input image I into Geometry Guidance and Latent Illumination Guidance Branch. The Geometry Guidance branch computes the depth map D_I for the image I, which is used as the input to ControlNet. The Latent Illumination Guidance branch computes a foreground mask F using I and creates a foreground-grayscale image I_{init}, which we use as input to the

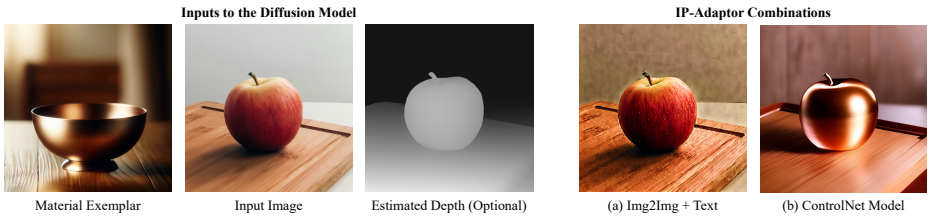

Fig. 3: The design choice of IP-Adaptor with ControlNet. Given the material exemplar and the input image, we dive into the different choices of utilizing the IP-Adaptor. In particular we realize that an Img2Img + text module (a) wouldn't properly transfer the materials properly to the main object. On the other hand, ControlNet (b) will preserve the geometry information of the given input. We thus utilize this as the starting point for geometry guidance to further explore the best illumination cues.

Diffusion Inpainting pipeline. We concatenate the embeddings from ControlNet with the inpainting diffusion model at the corresponding and inject the material embedding z_M through the cross-attention. The output of the inpainting diffusion model, I_{gen}, with the edited image containing the object in I cast with material from exemplar image M.

Our design choices to facilitate computation of material embedding, geometry guidance, and illumination cues are discussed in the following sections.

3.3 Encoding Material Exemplar

Given the material exemplar image M, this branch encodes the image into a latent representation while preserving its material properties. Previous works [39, 50] address this by finetuning a text-to-image diffusion model to encode the image into a rare token, implicitly treating the rare token as a latent representation that can be used in conjunction with other texts for image generation. However, this approach of optimizing for the material token requires the time-consuming step for every new material exemplar and usually requires 3-5 images to prevent overfitting.

We draw inspiration from the recently introduced IP-Adapter [49]. The IP adapter uses a CLIP image encoder to extract image features that can be injected into a diffusion model via the cross-attention layers. These features can be used as an additional condition to guide text prompts or other mediums for the generation. For example, one can input an image of a person and then describe "on the mountain" with text to obtain an image of the person in the mountains.

However, we realize that IP-Adaptor does not work well when combined with an Img2Img pipeline, as shown in Figure 3 (a) for our task. Moreover, adding text guidances like "changing the apple texture to golden bowl" does not produce photorealistic output and does not preserve other scene information (*i.e.* background). This problem of geometry and material entanglement within material embedding z_M remains unsolved, thus motivating the need for geometry and illumination guidance.

3.4 Geometry Guidance via Depth Estimation

Since decoupling geometry and material properties in images is challenging and requires additional training data, we provide an alternative solution where we enforce a stronger geometry prior to the diffusion model to overwrite the structural information present in z_M. To this end, we adopt a depth-based ControlNet to provide geometry guidance from the input image I. We observe that the geometry information from the depth map D_I overwrites the geometry information encoded in the z_M (see Figure 3 (b)). Note that with the geometry enforced by using depth-based ControlNet, we can successfully transfer the golden material of the bowl to the apple.

While the use of ControlNet with IP-Adaptor is introduced in the original IP-Adaptor paper [49], we employ it for a different purpose contrary to applying new structural control over an object in the image (*e.g.,* changing a person's pose). After extensively comparing various components for encoding the material exemplar and input image (analysis in Section 4.2), we find the depth-based guidance from pre-trained ControlNet helps us preserve the original geometry of the object for the task of material transfer.

While the addition of ControlNet helps preserve the geometry, we observe that the results suffer from inconsistency in preserving the illumination and background from the input image. This is evident in Figure 3, where the background and the lighting changes differ from the input.

3.5 Latent-space Illumination Guidance

Our final branch is primarily responsible for preserving the illumination and background in the input image. We propose two-fold guidance for illumination in the latent space during generation – an inpainting module and a foreground decoloring process. In addition to the attached IP-Adaptor and ControlNet, we adopt an inpainting diffusion model $\mathcal{S}$ instead of a standard generator. Specifically, our ControlNet-inpainting procedure takes in four conditions for image generation:

$$I_{gen} = \mathcal{S}(z_M, D_I, I_{init}, F), \tag{1}$$

where z_M is the material encoding, D_I is the depth map computed for input image I, I_{init} is the initial image to denoise from, and F is the foregound mask of target object in I which we are editing.

We conduct an ablation on the various versions of I_{init}, as shown in Figure 4. Specifically, we test out the following settings: (1) using the original input image, (2) initializing the foreground with random noise, and (3) using the foreground grayscaled image. Intuitively, directly letting $I_{init} = I$ (Setting (1)) would be a preferable option as I encompasses implicit lighting information (from the object's shading and the surrounding environment) while conveniently enforces all other parts of the image other than the object to remain the same. In practice, however, we found that using the original image inevitably introduces a strong prior of the base color from the input object (e.g. orange color of pumpkin), which would be entangled with the material base color from M in the output

Fig. 4: Ablating input for illumination guidance. To validate our design choice of the foreground-grayscale image for initializing inpainting, we compare the generated results against using the original image and random noise as inputs. The original image presents a strong base color prior that perturbs the generation, while the random image neglects shading information, leading to wrong lighting in both examples.

image. This artifact is sustained even when we significantly extend the number of denoising steps. On the other hand, when initializing I_{init} with random noise, the method indeed removes the base color prior but also removes the shading information causing incorrect illuminations in the synthesized object (e.g., the left side of the synthesized pumpkin is darker, but light is coming from the left). In our proposed pipeline, we perform grayscale operations in the pixel space for the object region (3). This provides a balanced solution of removing the strong color priors from the input image while keeping the shading cues for the inpainting diffusion model.

Thus, we propose to initialize I_{init} as:

$$I_{init} = F \odot I_{gray} + (1 - F) \odot I, \tag{2}$$

which converts the color of foreground object in the image to grayscale. $(1-F)\odot I$ implicitly preserves the lighting direction, intensity, and color information, and $F \odot I_{gray}$ preserves the object's shading information without base color prior.

3.6 Implementation Details

We implement our method using Stable Diffusion XL Inpainting [36] with the corresponding version of depth-based ControlNet [51] and IP-Adaptor [49]. We use Dense Prediction Transformers for depth estimation [38] and Rembg[1] for foreground extraction. Our method is implemented in PyTorch and runs on a single Nvidia A-10 GPU with 24 GB of RAM. For all Dreambooth approaches, we use the official LoRA-Dreambooth provided by Diffusers.

4 Experiments

We evaluate the efficacy of our method against various baselines. We also present several examples of downstream applications using our method.

[1] `https://github.com/danielgatis/rembg`

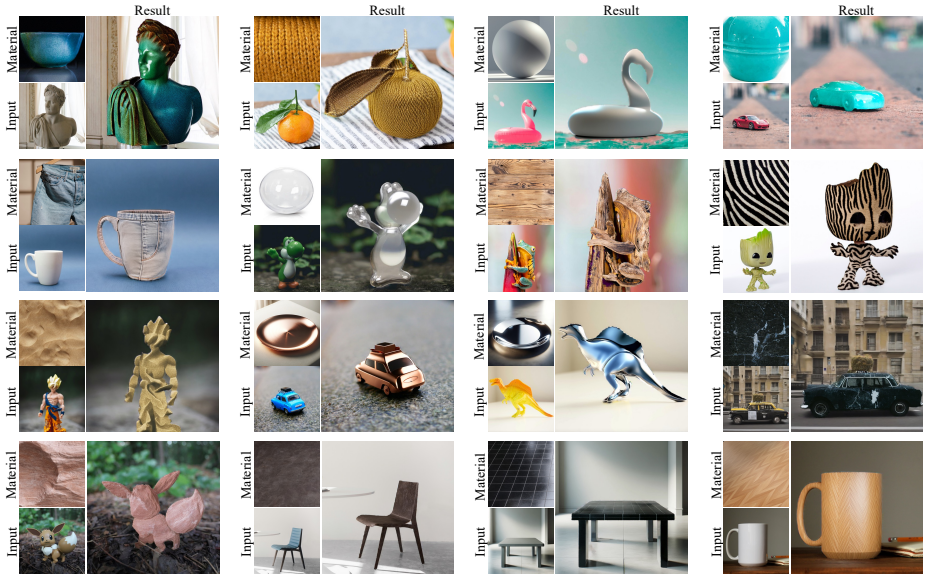

Fig. 5: Qualitative results on diverse materials. We present results of material transfer from a diverse set of material exemplar images. Even when perturbed by lighting and complex geometry, *ZeST* can still isolate the material information from the exemplar image and transfer to various objects while preserving the original geometry and illumination conditions. Note the change in specular regions as shinier materials are chosen in the case of the car made of brass and the dinosaur made of shiny steel.

4.1 Datasets

As the first to propose this problem, we create two datasets for comparison and evaluation. The real-world datasets provide us an understanding of our model's robustness, while the synthetic dataset is used for standard quantitative metrics.

Real-World Dataset. We curate a dataset comprising of 30 diverse material exemplars and 30 input images, collected from copyright-free image sources (*i.e.* Unsplash) and images generated by DALLE-3. All of these images are object-centric, where there exists a main object in the foreground to which we are extracting the material from or applying the material onto.

Synthetic Dataset. To perform quantitative evaluation, we use Blender to create a synthesized dataset of 9 materials randomly initialized by adjusting the base color, metallic, and roughness, and 20 meshes of different categories from Objaverse [15] rendered at three random viewpoints each, generating 540 ground-truth renderings. We render spheres assigned with each material individually and use the rendered image the material exemplar and pre-textured mesh rendering as input for all methods.

While *ZeST* is completely training-free, other methods of learning materials (e.g., Dreambooth) require further fine-tuning for every exemplar given. This

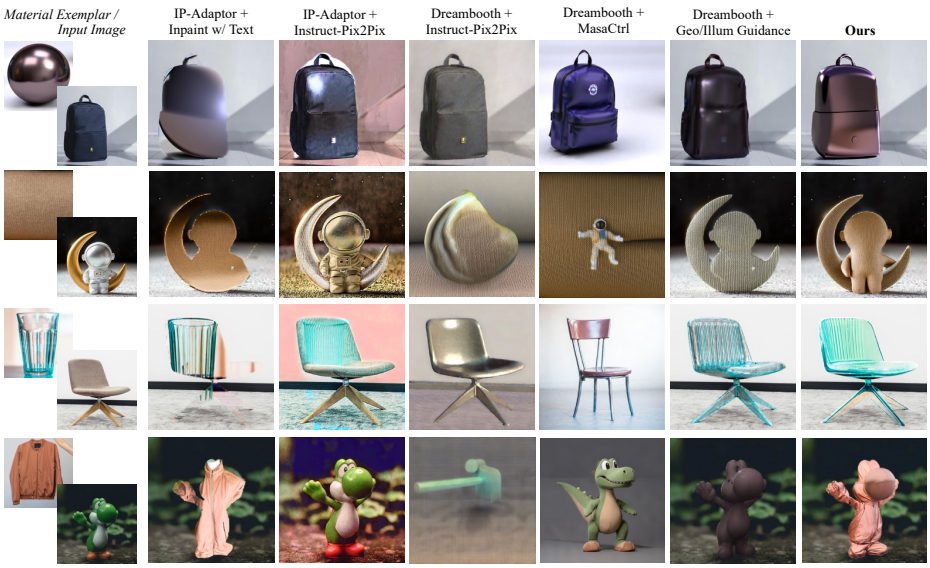

Fig. 6: Qualitative comparisons against baselines. Given the material exemplar and input image in the first column, we compare our method to five different baselines. Without any geometry guidance, all image editing baselines fail to impose the correct geometry of the input image. On the other hand, using Dreambooth with our geometry and illumination guidance often contains albedo shifts, potentially due to information loss when encoding material properties into a word token.

makes it infeasible to scale up the two datasets. Both our datasets are of comparable sizes to previous works on finetuning diffusion models [40, 50].

4.2 Qualitative Results

Material transfer results on real images. To demonstrate the application of *ZeST* on a wide range of materials and objects, we present examples of material transfer in Figure 5. The first three rows present results on real-world images, while the fourth row shows results using PBR materials [1]. Based on the examples, we observe that the material is properly disentangled from the geometry in the material exemplar and follows the shape of the object in the input image. This is particularly evident in the results of the orange, frog, and Groot toy figure, where the material is completely flat. We also notice accurate shadings in the bust and table examples when comparing them against their inputs. In the car and toy dinosaur examples, the reflections from the exemplars are isolated from the textural patterns and cast reasonably based on the illumination cues.

Qualitative comparisons. Since our work is the first to perform material transfer in latent space, we modified existing methods to compare against. Specifically, since existing image-guided texture synthesis methods utilize Dreambooth for their first step to encode the textures from images into word tokens [14,39,50],

we set Dreambooth as the backbone for learning material properties and combine with text-guided image editing techniques for comparison, including MasaCtrl and Instruct-Pix2Pix, and using *ZeST* but swapping out the IP-Adaptor with text. While our method is training-free, Dreambooth requires finetuning for every material exemplar given. We also explore alternative options to combine with IP-Adaptor, including text-guided inpainting and Instruct-Pix2Pix with the prompt "Change the texture of the object".

We present qualitative comparisons against the baselines on four exemplar and input images in Figure 6. By using Inpainting with Text prompt instead of ControlNet, the model ignores the geometry of the original input when casting the materials. In both cases when using Instruct-Pix2Pix (with IP-Adaptor or Dreambooth), the geometry of all objects is better preserved, but the model fails to capture the material property from the material exemplar image. The combination of Dreambooth and MasaCtrl fails to preserve the geometry of the object in the input image and misattributes the material. The closest baseline to ours is Dreambooth with our proposed geometry and illumination guidance; however, we observe that the word encoding process results in some information loss as evident in the color shifts of the backpack and the astronaut figure. Furthermore, the method requires additional training for every material exemplar, whereas *ZeST* takes roughly 15 seconds to generate the image.

Our method, *ZeST*, performs the task effectively by retaining the object geometry, scene illumination, and attributing the material correctly. Additionally, note that *ZeST* adapts to more challenging material exemplar images, such as transparent materials (glass cup in Figure 6 Row 3) and images with other minor objects (additional hand in Figure 6 Row 4).

4.3 Quantitative Comparisons

We follow previous work [41, 50] and use the synthetic images to compare all methods in terms of PSNR, LPIPS [52], and CLIP similarity score [37] against ground truth renderings. We also incorporate another DreamSim [19], a more recent metric that is more similar to human references. We grab IP-Adaptor + Instruct-Pix2Pix and Dreambooth + our geometry and illumination guidance as baselines, as they are the strongest (and only) performers from our qualitative comparisons that can roughly edit the material based on the geometry.

Table 1 (left) presents our results. We see a dramatic improvement when shifting from the instruct-pix2pix pipeline to our geometry and illumination guidance. While using Dreambooth performs similarly to our IP-Adaptor in the synthetic dataset, it requires a fine-tuned model for each material exemplar, making it unfeasible to scale up. In addition, we show in the next section that our method excels in real-world datasets.

User Study. We also create a user study with 16 participants to understand the capability of our model given real-world materials tested on real images. Each subject is shown 5 random samples from the 900 combinations generated from the dataset with our method and against the two strongest baselines: Dreambooth + ControlNet-Inpainting and IP-Adaptor + Instruct-Pix2Pix. We ask

Table 1: Quantitative Comparisons and User Study. We grab the strongest baselines in our qualitative comparisons for additional studies. Left: We measure the PSNR, LPIPS [52], CLIP similarity score [37], and DreamSim [19] in a quantitative study on the synthetic dataset of 540 exemplar-input combinations. Right: We perform a user study to evaluate the material fidelity and photorealism of the edited images from each method. We randomly sample 5 out of 900 real-world exemplar-input combinations for each of the 16 participants.

	PSNR↑	LPIPS↓	CLIP↑	DreamSim↓
IP-Adaptor + Instruct-Pix2Pix	17.08	0.099	0.740	0.390
DB + Our Geo/Illum. Guidance	25.52	0.058	0.874	0.238
Ours	**25.59**	**0.053**	**0.883**	**0.198**

	Fidelity↑	Photorealism↑
IP-Adaptor + Instruct-Pix2Pix	1.48	3.23
DB + Our Geo/Illum. Guidance	3.25	3.41
Ours	**4.05**	**3.78**

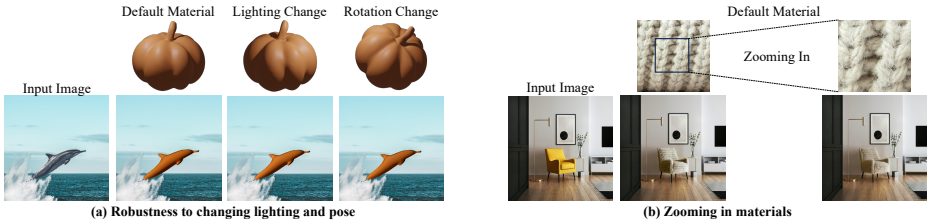

Fig. 7: Robustness to lighting and object pose. We present two types of robustness testing. **(a)**: Robustness to changing the material exemplar lighting and pose. **(b)**: Zooming into the material exemplar. Our model yields highly similar results in both, showing the capability to adapt to these external changes.

each subject to rate each image from 1 to 5 based on (1) material fidelity: how close the material in the generated image is compared to the original exemplar and (2) photorealism: how realistic the generated image is. Our results are summarized in Table 1 (right).

Our results show significant improvements from the two baselines in both material fidelity and photorealism of the edited image. The score improvements are also greater in real-world scenarios compared to synthetic ones. This could be the result of information loss during finetuning and overfitting to the exemplar background, which is less significant under controlled synthetic scenarios.

4.4 Robustness of the Model

In addition to the diverse set of results presented in Figure 5, we extensively test out the behavior of *ZeST* with special cases of material exemplar images.

Relighting and rotating the object in the material exemplar image. A good material extractor should be agnostic to small lighting and rotation changes of the same object used as the material exemplar. To evaluate this, we render a random material and cast it onto an irregular-shaped pumpkin (another example is in the Appendix). We then render three samples of the pumpkin, a default lighting orientation, a change in lighting direction pitch by 120 degrees, and a random rotation, as shown in 7 (a). The transferred materials onto the dolphin

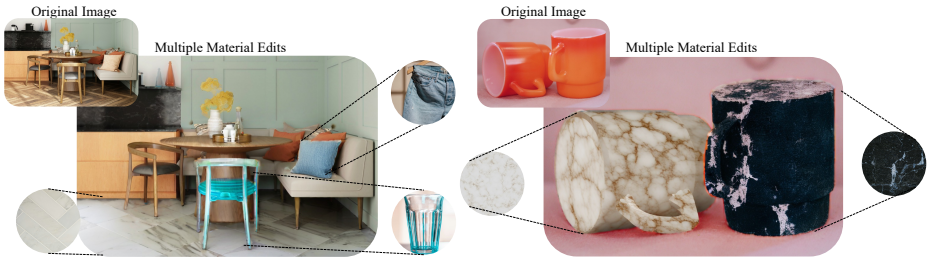

Fig. 8: Multiple Material Transfers in a Single Image. By replacing the foreground extraction with an open-vocabulary segmentation module (*e.g.,* SAM) to obtain multiple masks, *ZeST* can be applied iteratively to cast different material properties to different objects in a single RGB image.

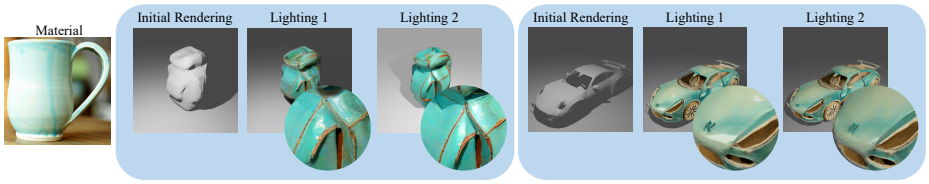

Fig. 9: Lighting-aware Image Editing. Given a rendering of a untextured mesh, we can alter *ZeST* slightly to achieve lighting-aware material edit. It can be seen from both examples where the reflection can be disentangled from the object texture.

remain roughly consistent across all samples, showing that our method is fairly resistant to these changes at a small scale.

Effect of image scale of material exemplar image. To examine the effect of the scale of the material exemplar, we first use an image of a woolen cloth material with a distinctive repeating pattern and apply our method to an image of a chair. Then, we zoom into the exemplar image manually to the edge only very few repeated patterns are left. Our results in Figure 7 (b) show that while the scale of the material is drastically different, the model automatically re-adjusts the patterns into a reasonable size to be cast onto the input image.

4.5 Applications

Applying multiple materials to multiple objects. By replacing the foreground extraction with a segmentation module (*e.g.,* SAM) to obtain multiple masks, *ZeST* can be used to iteratively change multiple materials in a single image. Figure 8 presents two examples of editing multiple objects in a single image. As evident in the transparent glass chair where the wooden table behind is roughly visible, *ZeST* generalizes to complex scenes with multiple objects.

Lighting-aware Material Transfer. Given a material exemplar image and an untextured mesh rendered under multiple illumination conditions, *ZeST* can also perform lighting-aware material transfer. Specifically, we first generate the

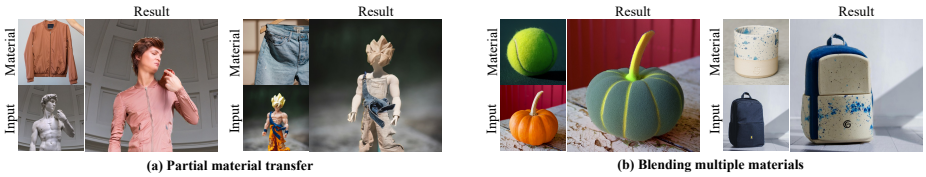

Fig. 10: Limitations. Our method primarily fails in two modes. **(a)** The model sometimes picks the most "probable" areas to transfer the material, instead of casting the material on the entire object. **(b)** If two textures are present in the exemplar image (e.g., foreground and background of the tennis ball, the glazed top and bottom logo of the cup), the model sometimes combine both materials when performing the edit.

materials and textures of the image under Lighting 1 using *ZeST*. Then, by fixing the same seed during generation and using the generating image given the first lighting as the input to the second, we can enforce consistency in the material and texture generated (details of implementation in Appendix) while changing the reflections. We show examples of transferring the glazed cup material to two mesh renders in Figure 9. *ZeST* successfully disentangles the reflections while keeping most textural patterns consistent between the two images. This technique could potentially be applied jointly with other 3D texture synthesis works [10] and be helpful to applications such as e-commerce design.

4.6 Limitations

Since *ZeST* operates majorly in the latent space, the model sometimes exhibits uncontrollable behaviors based on its image understanding. Figure 10 presents two forms of more frequent failure cases: (a) Partial material transfer: the material is only transferred to parts instead of the entirety of the object. We hypothesize that the failure stems from the entanglement of material properties and the exemplar's identity, as the material is only applied to where it seems the most probable (*e.g.,* only apply the jacket material to the statue's body). (b) Blending multiple materials: since the current IP-Adaptor does not have a module to extract regions of an image for material transfer, *ZeST* sometimes mixes up multiple materials in the exemplar image during transfer.

5 Conclusion

We present *ZeST*, a zero-shot, training-free method for exemplar-based material-editing. *ZeST* is built completely using readily available pre-trained models and demonstrates generalizable and robust results on real images. We curate synthetic and real image datasets to evaluate the performance of our approach. We also demonstrate downstream applications like multiple edits in a single image and material-aware relighting. *ZeST* serves as a strong starting point for future research in image-to-image material transfer, implying opportunities of leveraging pre-trained image diffusion models for complex graphic designing tasks.

References

1. https://www.textures.com/browse/pbr-materials/114558
2. Aittala, M., Weyrich, T., Lehtinen, J.: Practical svbrdf capture in the frequency domain. ACM Trans. Graph. **32**(4), 110–1 (2013)
3. Aittala, M., Weyrich, T., Lehtinen, J., et al.: Two-shot svbrdf capture for stationary materials. ACM Trans. Graph. **34**(4), 110–1 (2015)
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
5. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479–3487 (2015)
6. Bhat, S.F., Mitra, N.J., Wonka, P.: Loosecontrol: Lifting controlnet for generalized depth conditioning. arXiv preprint arXiv:2312.03079 (2023)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
8. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
9. Cao, T., Kreis, K., Fidler, S., Sharp, N., Yin, K.: Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4169–4181 (2023)
10. Chen, D.Z., Siddiqui, Y., Lee, H.Y., Tulyakov, S., Nießner, M.: Text2tex: Text-driven texture synthesis via diffusion models. arXiv preprint arXiv:2303.11396 (2023)
11. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5343–5353 (2024)
12. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
13. Cheng, T.Y., Gadelha, M., Groueix, T., Fisher, M., Mech, R., Markham, A., Trigoni, N.: Learning continuous 3d words for text-to-image generation. arXiv preprint arXiv:2402.08654 (2024)
14. Corneanu, C., Gadde, R., Martinez, A.M.: Latentpaint: Image inpainting in latent space with diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4334–4343 (2024)
15. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
16. Delanoy, J., Lagunas, M., Condor, J., Gutierrez, D., Masia, B.: A generative framework for image-based editing of material appearance using perceptual attributes. In: Computer Graphics Forum. vol. 41, pp. 453–464. Wiley Online Library (2022)
17. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Flexible svbrdf capture with a multi-image deep network. In: Computer graphics forum. vol. 38, pp. 1–13. Wiley Online Library (2019)
18. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

19. Fu*, S., Tamir*, N., Sundaram*, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. NeurIPS (2023)
20. Ge, S., Park, T., Zhu, J.Y., Huang, J.B.: Expressive text-to-image generation with rich text. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7545–7556 (2023)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
24. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
25. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
26. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134 (2023)
27. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022)
28. Khan, E.A., Reinhard, E., Fleming, R.W., Bülthoff, H.H.: Image-based material editing. ACM Transactions on Graphics (TOG) **25**(3), 654–663 (2006)
29. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
30. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
31. Liang, Y., Wakaki, R., Nobuhara, S., Nishino, K.: Multimodal material segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19800–19808 (2022)
32. Lopes, I., Pizzati, F., de Charette, R.: Material palette: Extraction of materials from a single image. arXiv preprint arXiv:2311.17060 (2023)
33. Michel, O., Bhattad, A., VanderBilt, E., Krishna, R., Kembhavi, A., Gupta, T.: Object 3dit: Language-guided 3d-aware image editing. Advances in Neural Information Processing Systems **36** (2024)
34. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
35. Pandey, K., Guerrero, P., Gadelha, M., Hold-Geoffroy, Y., Singh, K., Mitra, N.: Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. arXiv preprint arXiv:2312.02190 (2023)
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
38. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
39. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
40. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
41. Sharma, P., Jampani, V., Li, Y., Jia, X., Lagun, D., Durand, F., Freeman, W.T., Matthews, M.: Alchemist: Parametric control of material properties with diffusion models. arXiv preprint arXiv:2312.02970 (2023)
42. Sharma, P., Philip, J., Gharbi, M., Freeman, B., Durand, F., Deschaintre, V.: Materialistic: Selecting similar materials in images. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023)
43. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
44. Subias, J.D., Lagunas, M.: In-the-wild material appearance editing using perceptual attributes. In: Computer Graphics Forum. vol. 42, pp. 333–345. Wiley Online Library (2023)
45. Upchurch, P., Niu, R.: A dense material segmentation dataset for indoor and outdoor scene parsing. In: European Conference on Computer Vision. pp. 450–466. Springer (2022)
46. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
47. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. arXiv preprint arXiv:2402.03290 (2024)
48. Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023)
49. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
50. Yeh, Y.Y., Huang, J.B., Kim, C., Xiao, L., Nguyen-Phuoc, T., Khan, N., Zhang, C., Chandraker, M., Marshall, C.S., Dong, Z., et al.: Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. arXiv preprint arXiv:2401.09416 (2024)
51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
53. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)