MARBLE: Material Recomposition and Blending in CLIP-Space

Ta Ying Cheng*
University of Oxford

Prafull Sharma
MIT CSAIL

Mark Boss
Stability AI

Varun Jampani
Stability AI

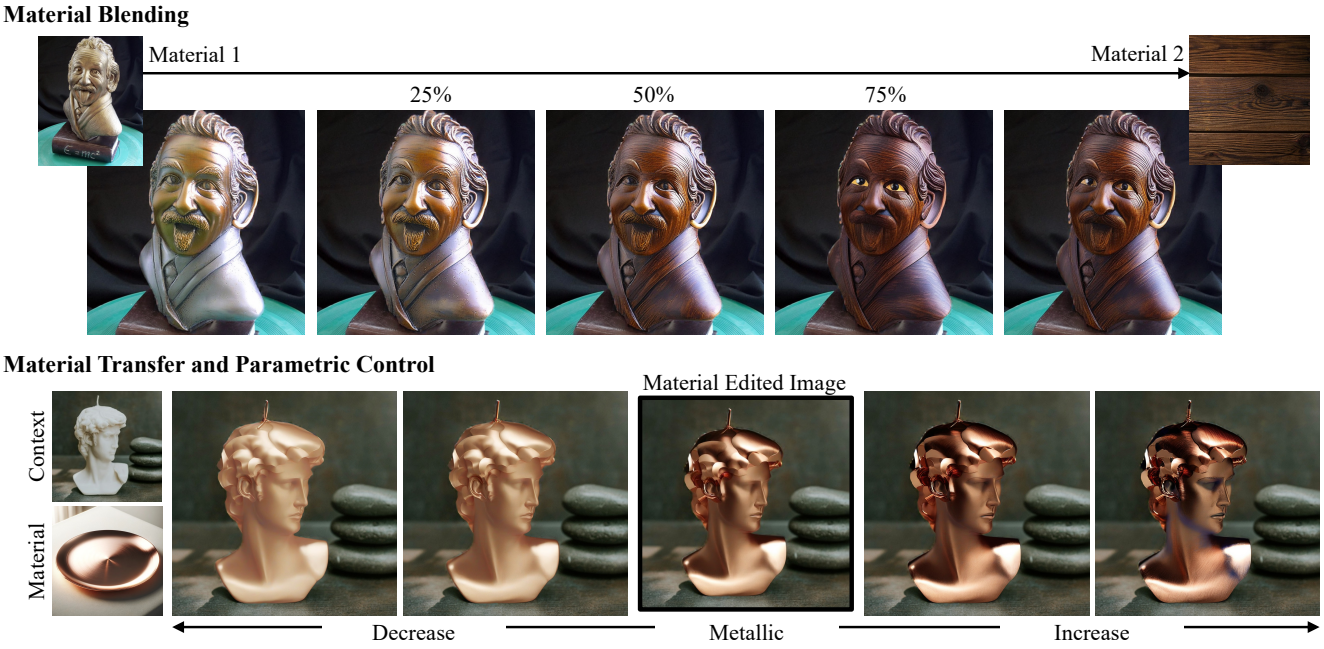

Figure 1. **Overview.** We present MARBLE, a method for performing various material editing in images such as material blending (top row) and parametric control of material properties (bottom row) leveraging CLIP-space and pre-trained generative models. Given two material exemplar images, we can achieve a controllable blend of materials on the object by blending the material representation in CLIP-space. For parametric material attribute control, we learn a shallow network using synthetic data to predict the direction in CLIP-space for changing specific material properties such as metallic.

Abstract

Editing materials of objects in images based on exemplar images is an active area of research in computer vision and graphics. We propose MARBLE, a method for performing material blending and recomposing fine-grained material properties by finding material embeddings in CLIP-space and using that to control pre-trained text-to-image models. We improve exemplar-based material editing by finding a block in the denoising UNet responsible for material attribution. Given two material exemplar-images, we find directions in the CLIP-space for blending the materials. Further, we can achieve parametric control over fine-grained material attributes such as roughness, metallic, transparency, and glow using a shallow network to predict the direction for the desired material attribute change. We perform qualitative and quantitative analysis to demonstrate the efficacy of our proposed method. We also present the ability of our method to perform multiple edits in a single forward pass and applicability to painting.
Project Page: https://marblecontrol.github.io/

1. Introduction

Editing object materials such as diffuse albedo, roughness, etc. in images is instrumental for graphics and vision applications such as game design, advertising, and visual content creation. Performing material editing in a single image using traditional graphics techniques requires under-

*Work was done during internship at Stability AI.

standing of several object and environment properties such as object geometry, its material properties as well as environment illumination, making it a highly challenging task. Previous approaches for material editing use crude approximations of object geometry and environment maps [31], resulting in non-photorealistic results limited to finite material editing options.

In this work, we tackle the problem of material transfer and recompose the material properties given a single image by directly leveraging the implicit knowledge of object and environment properties present in the pre-trained image diffusion models [46]. This circumvents the need for explicit estimation of these properties, which is challenging given a single image. Recent works such as Alchemist [51] and ZeST [15] demonstrate the use of diffusion models for material editing in images. ZeST [15] proposes a zero-shot technique for exemplar-based material transfer, where the object material from an exemplar image is transferred to the target object in the input image. However, this approach is limited to high-level material changes and does not perform fine-grained control of material properties. On the other hand, Alchemist [51] proposes a supervised fine-tuning of Stable Diffusion [46] for fine-grained material control such as roughness, transparency etc. in images. However, such a fine-tuning of the diffusion model has the potential to overfit to the synthetic data used for training, thereby destroying the valuable object prior knowledge in these models.

In contrast, we propose a technique that can perform versatile material editing with material transfer as well as fine-grained control, while also retaining the base diffusion model priors. In particular, we propose to keep the image diffusion model intact and perform material editing via CLIP [43] image features that are injected into the diffusion model. A key contribution of this work is to demonstrate that a surprising amount of material editing is possible with the manipulation of the CLIP-Space features. Our technique called MARBLE (Material Recomposition and Blending in CLIP-space) enables versatile material editing tasks ranging from performing coarse material transfer using an example image or blending materials from multiple objects (Figure 1 top row) to fine-grained material control of properties such as metallic, transparency, etc. (Figure 1 bottom row).

It is far from trivial to achieve such diverse material editing tasks using only CLIP image features as CLIP captures all the object properties such as semantics, geometry etc., not just the material properties. We build our method using ZeST architecture [15] for exemplar-based material transfer with some modifications. ZeST uses IP-Adapter [60] that injects CLIP features into the diffusion model, along with a color-agnostic inpainting technique for material transfer from an exemplar image to the target object image. With systematic experiments, we find a U-Net block in the Stable Diffusion that responds to the object materials. Following this insight, we propose to inject CLIP features into this specific U-Net block resulting in better material transfer. This modified architecture acts as the base for two variants of material editing. First, we show that this technique can also be used for material blending between two or more exemplar images. For the fine-grained control of the material properties such as increasing or decreasing transparency, we propose to learn lightweight MLPs, using a small synthetic dataset, that predicts material editing directions for each of the individual material properties in the CLIP-space. As a result, we can achieve fine-grained control of the material properties by moving the CLIP features along these editing directions.

We provide extensive experimental analysis and results on a wide range of applications, combining a series of coarse and fine-grained material edits. Results on both synthetic and real-world images demonstrate highly plausible material editing using MARBLE for material transfer as well as fine-grained control. We compare MARBLE against other image/material editing approaches when a baseline is available, of which both our quantitative and qualitative analysis shows superiority in performance. As we keep the based diffusion model intact, we find that the learned editing directions using the shader-based synthetic dataset can generalize to various image styles, including anime and paintings. Overall, MARBLE has several favorable properties for material editing in images:

- **Wide Range of Novel Editing Controls.** To the best of our knowledge, MARBLE is the first approach to offer parametric control, exemplar-based guidance, and blending of materials all within one general framework.
- **Operates only in CLIP-Space.** The minimal tuning nature of our approach brings maximal flexibility in model selection and performing multiple edits in one go.
- **Robustness in Various Styles.** MARBLE can not only generate and edit materials of realistic images but can also be incorporated with various painting and artwork styles, bringing much flexibility to graphic designers.

2. Related Work

Controlled Image Editing with Diffusion Models. Recent advances in text and class-conditional image generation using diffusion models enable photorealistic image generation [19, 26–30, 39, 44, 46, 49, 53]. These models act as a base model for performing 3D-aware inpainting [41, 46], text-based editing [7, 22], and controlled generation [13, 33, 48]. High-level semantic and stylistic edits are based on inputs such as text-based instructions [8, 22, 25, 56], semantic segmentation [3], bounding box [12, 35, 58, 59], and images [14, 47, 48, 50, 57, 60]. InstructPix2Pix allows for instruction-based editing of images, allowing for stylistic and high-level semantic changes

Figure 2. **Comparison of material block injection vs. all Blocks injection.** We present examples of using the same input and material exemplar. Given the same depth condition, injecting only into the material block allows much better geometry preservation compared to injecting to all blocks in the UNet.

in the images [7]. These methods are trained and hence limited to domains of high-level semantic changes, failing to edit low-level details such as object geometry and materials. Beyond high-level semantic control, edits can be performed based on mid-level features such as depth maps [6, 64] and edge-maps [38, 62].

Recent work has enabled fine-grained continuous control enhancing the level on control in image editing [4, 21, 42]. Continuous control over concepts such as weather, age, and styles can be achieved in diffusion model-based generative models from a small set of text or images [21]. Baumann et al. identifies directions within token-level CLIP text embeddings allowing for fine-grained over high-level attributes such as age and aesthetics in text-to-image models [4]. These methods demonstrate control over high-level semantics using embeddings of pre-trained models.

Material Editing. Material editing is a challenging task, requiring understanding of object and scene properties such as geometry, illumination, and material attributes. Material acquisition methods extract material properties under known illumination and camera configurations [1, 2, 18]. Recent methods explore material recognition and segmentation with a data driven approach requiring little to no prior knowledge about the environment [5, 32, 36, 52, 55].

Khan et al. proposed in-image material editing with normal estimates as approximations of the scene geometry [31]. Advances in generative models have facilitated more robust and photorealistic material editing techniques in images and 3D models [9–11, 15, 17, 23, 37, 45, 51, 54, 61]. Coarse material editing in a zero-shot manner leveraging generative priors of pre-trained text-to-image models along with geometric and illumination information [15]. Fine-grained material properties can be edited by finetuning a generative model on physically rendered data [51].

In our work, we propose a method for using using CLIP-space for material editing, specifically blending materials and recomposing fine-grained material attributes.

3. Method

Our method, MARBLE, uses CLIP embeddings and a pre-trained diffusion model to perform efficient material transfer, material blending, and parametric tuning of fine-grained material attributes. Specifically, we extend the architecture from ZeST, a zero-shot approach on performing exemplar-based material transfer by Cheng et al. [15].

Exemplar-based material transfer methods aim to transfer the material M from a given exemplar image I_m to an object in an input image I. ZeST performs the material transfer in a zero-shot manner using a pre-trained inpainting model (e.g., Stable Diffusion XL [46]) $\mathcal{S}$. It guides the inpainting model using the foreground mask of the object F_I, depth map D_I as geometric cue, and a foreground grayscale initial image I_{init} as illumination cue, aiming to utilize only the material features $f(z_m)$ from I_m during the generation process.

$$I_{gen} = \mathcal{S}(I_{init}, F_I, D_I, f(z_m)) \tag{1}$$

Note that $f(\cdot)$ is the cross attention injection of feature z_m originally computed using the CLS token from CLIP encoder with a fine-tuned head provided by IP-Adapter. While this method results in images with plausible material transfer, this approach is not robust due to the convoluted nature of the CLIP embeddings – some information besides materials is still passed to the denoising process, leading to cases of shifts in object geometry and shading.

3.1. Targeted Material Block Injection

To mitigate these limitations, we find and inject the material embedding only to attention blocks in the denoising UNet of the inpainting model, the one responsible for attributing materials on the objects. Instead of injecting the material embedding z_m at each of the attention layers in the denoising UNet, we find specific block responsible for material attribution following the process inspired by InstantStyle [57]. InstantStyle identifies a specific block responsible for injecting style information to the objects. We perform a similar study of exhaustively visualizing the generated results when injecting the information across each block of the denoising UNet to identify which layer contributes specifically to material transfer. Our results illustrate that both material and style attribution on objects is performed by the same layer close to the bottleneck of the UNet. This is an expected outcome as material transfer can be seen as a specialized form of style transfer.

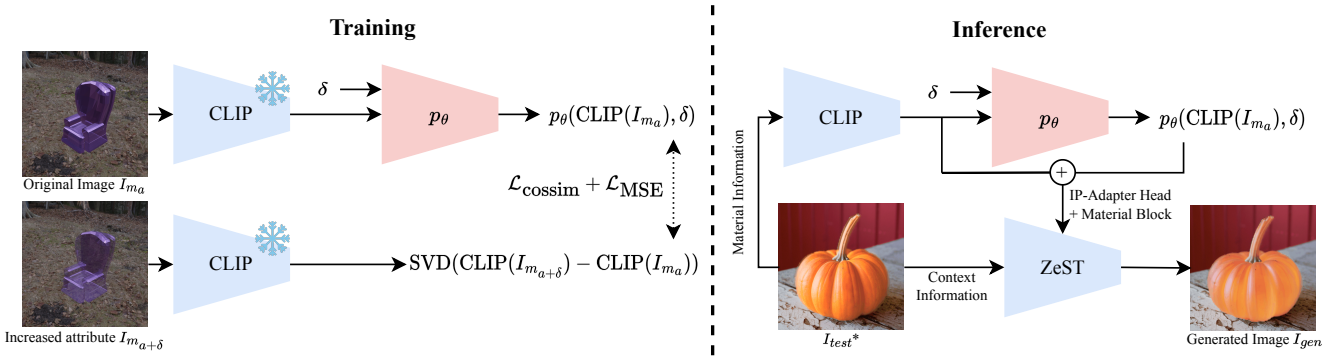

Figure 3. **Method overview for parametric material attribute control.** During training, we aim to learn p_θ, a shallow MLP that predicts the editing direction in CLIP space given an image I_{m_a}. During inference, we can use p_θ to predict the offset that can be added to the CLIP embedding for parametric control. Note that I_{test}* during test time can be separated into two images, one for the context information (background, shading, geometry) and another for material.

Figure 4. **Examples of Dataset.** We show samples from the rendered dataset for varying roughness, metallic, transparency, and glow.

To this end, we propose to alter $f(\cdot)$ to inject the material embedding z_m only in that specific block of denoising UNet. Figure 2 presents examples comparing material block injection against all blocks injection (proposed by ZeST). In all examples, the geometry of the initial input condition is better preserved when the features are injected only into the material block. Note in Row 1 that the material transfer preserves the details of the material exemplar, while the original result of ZeST hallucinates hands on the toy figure – a result primarily caused by the entanglement of the identity of the jacket and material. This modification helps in preserving the geometry and lighting of the object and thus acts as the base architecture for MARBLE.

3.2. Material Blending

Using this improved architecture, we aim to edit the context image with a material interpolated between two material exemplars. We observe that interpolating features from two material exemplars is also interpretable within the CLIP embeddings, similar to many results on finding interpretable directions in pre-trained models for pose and appearance [24, 40]. This enables blended materials for image editing given two material features z_{m_1} and z_{m_2} extracted from two images using the CLIP encoder:

$$I_{gen} = \mathcal{S}(I_{init}, F_I, D_I, f(\alpha z_{m_1} + (1-\alpha) z_{m_2})), \quad (2)$$

where $\alpha > 0$ is the interpolation weights.

Material blending can be performed with three different configurations of the exemplar images: (1) different objects and materials, (2) different objects made of the same materials with a single attribute (e.g. roughness) varied, and (3) same object, same materials with a single attribute varied.

3.3. Parametric Control from a Single Image

In addition to blending between two materials from two exemplar images, we explore the use of CLIP-space embeddings for achieving parametric control over fine-grained material attributes. Specifically, we demonstrate parametric control over roughness, metallic, transparency, and glow.

Given a material exemplar image I_{m_a} with a specific material attribute a, and an editing strength δ, we train a attribute editing network p_θ to predict the corresponding CLIP feature of $I_{m_{a+\delta}}$. We train the attribute editing network for each attribute individually using a synthetically rendered dataset. Next, we describe the dataset preparation, training, and inference setup of our method.

Dataset Creation. Since collecting real world dataset with controlled material attribute changes is impractical, we render a small dataset using Blender with controlled shader properties. Contrary to the approach of Alchemist [51], our model only predicts directions in the CLIP-space and thus requires much less data. We show in Section 4.3 an ablation on the quality of generated images against the number of objects used, where our attribute changing network could be learned with even as few as 8 objects.

We used 300 synthetic objects [16] (250 for training and 50 for validation) and pair each with a random HDR map from a collection of 50 maps. To create a dataset for attribute a, we create a default material per object with attributes other than a randomly assigned. Then, we render the object at a random viewpoint with traversing the value of a from uniform steps. Note that for transparency, we

not only increase the transmission value of the material but also decrease the roughness effect to create a more glass-like transparent appearance. We present some examples in Figure 4.

Training and Inference Setup. While this rendered dataset proved to be useful for our task, we identify two key limitations. First, the dataset is fairly small resulting in potential inductive biases in the image features. Second, note that CLIP features are fairly noisy [34]. These observations suggest that there may be a small set of features within the CLIP features we should not learn from. To mitigate this, we stack the editing directions of a given attribute (computed as the difference of two CLIP features given an image pair) and perform singular value decomposition to obtain a low-rank approximation of the stacked matrix. The rank for each attribute is decided by the elbow method when plotting out the singular values. The variance explained for all four attributes fall within the range of $67\% - 80\%$.

Figure 3 provides an overview of the training and inference setup. During training, we take an input image I_{m_a} and an editing strength δ. We then train our attribute editing network $p_\theta(I_{m_a}, \delta)$ (a 2-layer MLP) with the criterion:

$$\arg\min_\theta [\text{cossim}(s_{m_{a+\delta}}, p_\theta(I_{m_a}, \delta)) + \text{MSE}(s_{m_{a+\delta}}, p_\theta(I_{m_a}, \delta))], \quad (3)$$

where $s_{m_{a+\delta}}$ is the low-rank material attribute from a to $a+\delta$ approximated by SVD, [cossim, MSE] are the cosine similarity loss and mean-squared loss, respectively. With this objective, p_θ learns to predict the low-rank approximated CLIP feature of the same original image with one attribute α increased by δ.

With a learned attribute editing network p_θ at inference time, we can obtain fine-grained material features after tuning attribute a to $a+\delta$ altering z_{m_a} into $z_{m_{a+\delta}}$:

$$z_{m_{a+\delta}} = \text{CLIP}(I^i_m) + p_\theta(I^i_m, \delta). \quad (4)$$

This feature allows us to build on top of our exemplar-based transfer pipeline, where we can regenerate an image with recomposed material attributes.

Since we did not finetune the pre-trained diffusion model, each attribute network can be trained separately and used jointly to find the designated CLIP feature. We provide examples of this in Section 4.3.

4. Experiments

We present qualitative and quantitative evaluations to validate MARBLE. We analyze the effectiveness of our method at material blending and fine-grained parametric control over material attributes. Further analysis on the robustness beyond natural images and dataset efficiency for achieving parametric control demonstrates the practical impact for material editing.

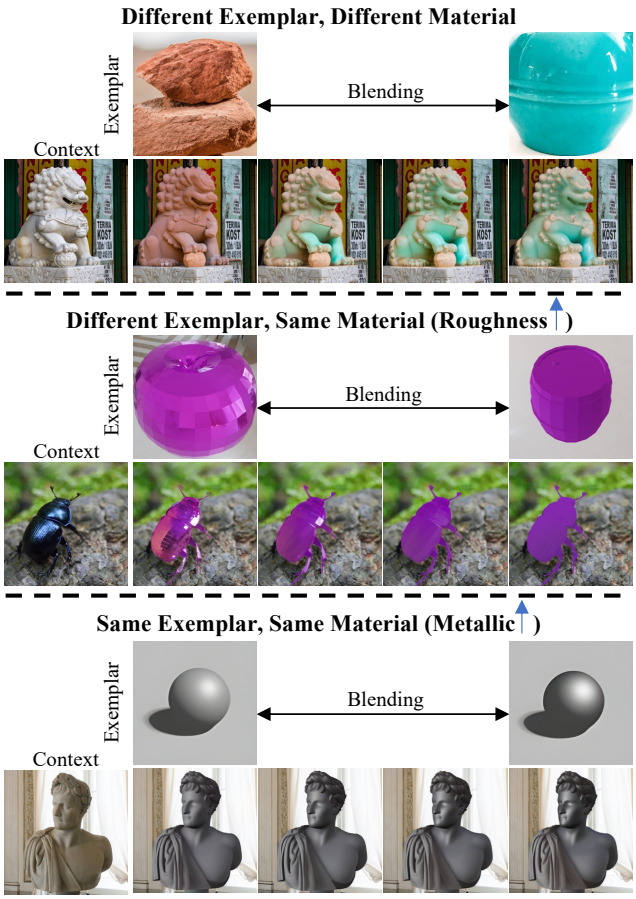

Figure 5. **Material blending results.** By interpolating the CLIP features of the material exemplars, MARBLE can transfer the intermediate blended features to the input image, creating material blending effect. Blending can work with exemplar images with following configurations: (1) Different objects with different materials, (2) Different objects made of the same base material except one varying attribute (metallic), and (3) Same object and same material with one attribute (metallic) varying.

4.1. Qualitative Results

Material Blending. We present material blending results with different selection of material pairs in Figure 5. Note that exemplar materials m_1 and m_2 can be in different configurations. They can be two completely different material exemplars (example 1) or the same exemplar and material with only one attribute varying (example 3). Surprisingly, even when the two exemplars are of different objects with the same base material with a single varying attribute (second example), the CLIP embeddings are sufficiently able to identify the underlying attribute and perform parametric control through material blending.

Parametric Control of Material Attributes. We present slider results for roughness, metallic, transparency, and glow in Figure 6. For each of the attributes, we show two sliding examples, one using the reference image for both

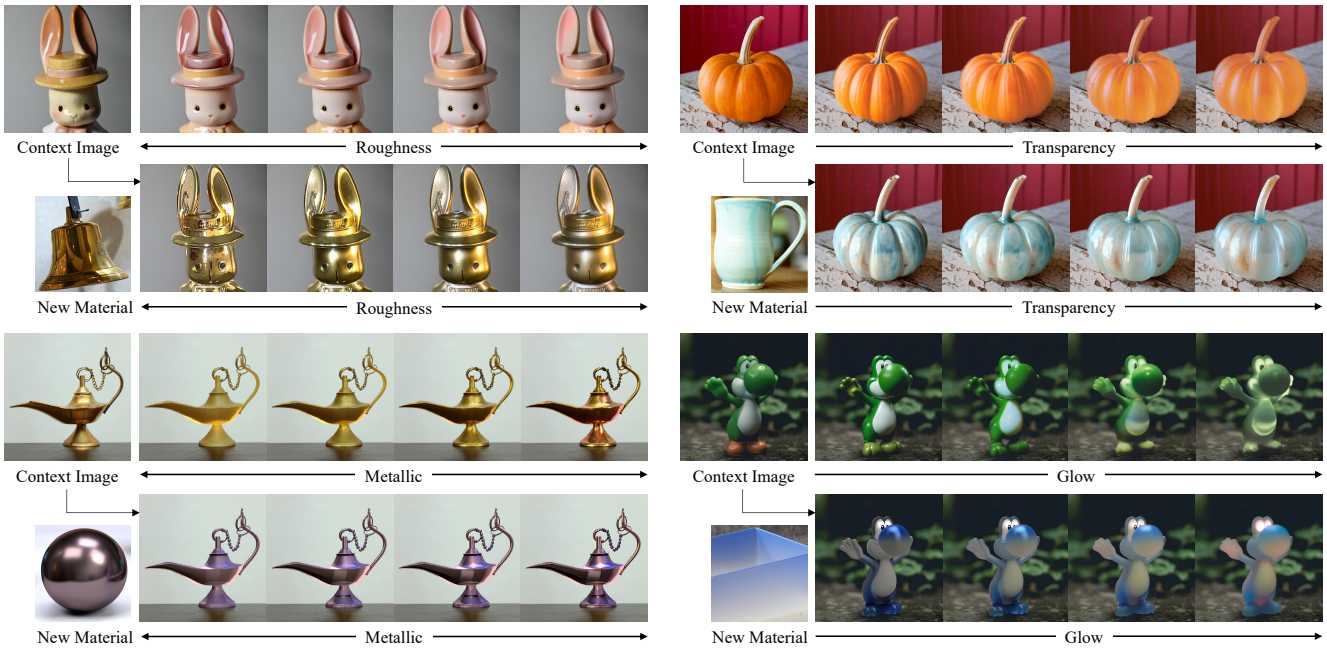

Figure 6. **Parametric control results.** We present four sets of results controlling roughness, transparency, metallic, and glow. For each set of results, we present one example directly using the reference image for context and material, and another set where we change to a new material exemplar. MARBLE disentangles the reflections from the albedo to provide perceptually convincing results.

context and material, and the other with a new material applied demonstrating the combination of material editing and fine-grained control in a single forward pass.

Note that the attribute we intend to control is disentangled from the other attributes. For the roughness examples, we observe reduction in the specularity on the surface as roughness increases. For the transparency and metallic examples, the reflection is disentangled with the albedo/base color of the object, lighting up the colors in some regions and darkening the others. For the glow example, the color of the glow follows the original albedo of the object.

Parametric Control Qualitative Comparisons. We compare our method to three baselines, namely InstructPix2Pix [7], Concept Slider (Text), and Concept Slider (Image) [21].

For InstructPix2Pix, we use prompts to guide attribute changes. Specifically, given an image, we use the prompt "`Make the *object more/less *attribute`", where `*object` is the object in the image and `*attribute` is the intended attribute change (e.g., `Make the chair transparent`). Note that this does not allow for parametric control.

We also implement two types of Concept Sliders using text and image with the SDXL backbone. For text concept sliders, we find opposite words describing the attribute (e.g. `transparent` and `opaque`, `rough` and `smooth`) and train a slider for each set of attributes. For image sliders, we use pairs from the two ends of the spectrum of our dataset to train for each material attribute. During inference, we perform DDIM inversion on the image and increase the slider value to the maximum before noticeable artifacts occur.

Figure 7 presents the qualitative comparisons against the baselines. Due to the ambiguity of text descriptions, InstructPix2Pix often leads to unintended changes on other attributes such as the geometry of the pot and chair, the albedo of the car for metallic, or the background for the toy figure for glow. On the other hand, editing with Concept Sliders (trained with either text or image pairs) requires DDIM inversion in the first place, which leads to inaccurate reconstructions even without parametric changes. While the sliders occasionally produces reasonable outputs for metallic and roughness (rougher pot surfaces and reflection on the car), the concept of transparency and glow were not captured by this approach. Our method produces high-fidelity results, showing disentangled and accurate edits for all attributes.

4.2. Quantitative Results

Out of the three baselines, only the image-trained concept slider allows us to compute quantitative metrics. InstructPix2Pix does not allow continuous control and neither does the editing strength of text-trained concept slider correspond well with the actual shader value changes in Blender.

Using a rendered validation set comprising of 50 objects, each with changing material attributes, we compare against image-trained concept sliders in terms of PSNR, LPIPS [63], CLIP Score [43] and DreamSim [20]. Table 1 shows that MARBLE performs better than baselines across

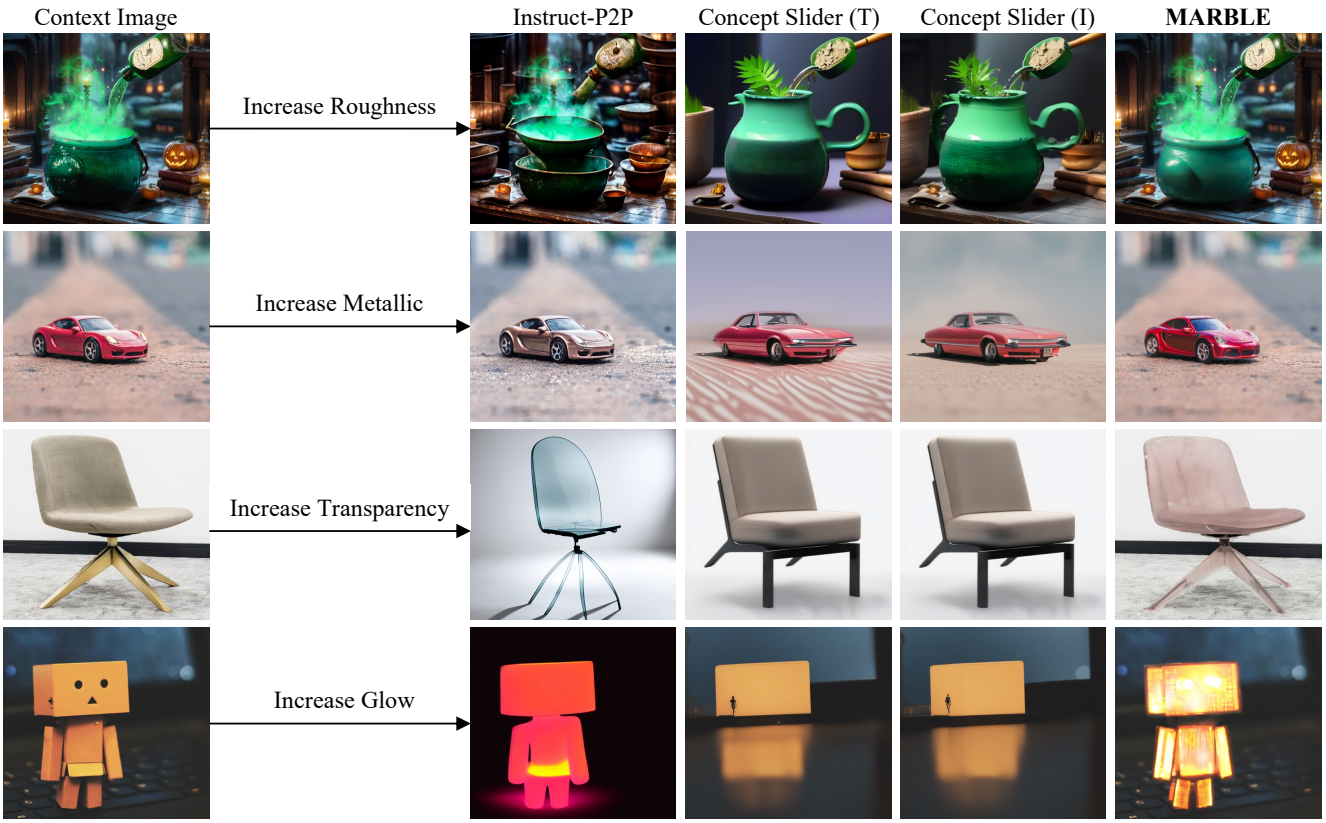

Figure 7. **Qualitative comparisons.** We compare against InstructPix2Pix and 2 versions of Concept Sliders. (T) and (I) denote text and image trained versions, respectively. All baselines either fail to capture the parametric control (Concept Sliders/chair/transparency), or result in undesired changes in object geometry (toy figure/glow, pot/roughness) or albedo (InstructPix2Pix/car/metallic).

Table 1. **Quantitative comparisons.** We present quantitative comparisons for all attribute controls compared to Concept Sliders trained using our dataset.

	PSNR↑	LPIPS↓	CLIP↑	DreamSim↓
Roughness				
Concept Slider (Images)	18.87	0.356	0.597	0.567
MARBLE	**26.56**	**0.056**	**0.931**	**0.129**
Metallic				
Concept Slider (Images)	19.45	0.317	0.655	0.479
MARBLE	**26.82**	**0.053**	**0.928**	**0.121**
Transparency				
Concept Slider (Images)	19.85	0.346	0.639	0.525
MARBLE	**26.99**	**0.070**	**0.905**	**0.163**
Glow				
Concept Slider (Images)	16.92	0.301	0.661	0.509
MARBLE	**19.73**	**0.111**	**0.890**	**0.213**

all metrics for all attributes.

User Study. To further validate the effectiveness of MARBLE on real-world images, we conduct a user study with 16 participants. We generate results on 20 real-world images with edits controlling a random material attribute using our method and image-based Concept Slider. Each user was provided 3 image sets to compare based on the intended control. As a result, 87.5% participants chose images generated by our method, MARBLE.

4.3. Discussion

Multi-Concept Control-Grid. One of the main merits of CLIP-based control is the ability to control multiple attributes in a single forward pass. Figure 8 presents an example of controlling roughness and metallic components of the toy car's material. MARBLE allows us to transfer a metal material onto the toy, while simultaneously enabling fine-grained control over metallic and roughness of material. Each image is generated in a single forward pass. While trained separately, we can see that the two attributes are disentangled from one another even when applied together.

Robustness on Real-World Images. In addition to the eight slider examples, we also present a variety of results of increasing the value one attribute. MARBLE was able to perform realistic edits across a variety of object from different backgrounds. As In-the-wild editing by Subias et al. also support metallic, we show the qualitative comparisons for the three examples given.

Parametric Control with Different Styles. By solely

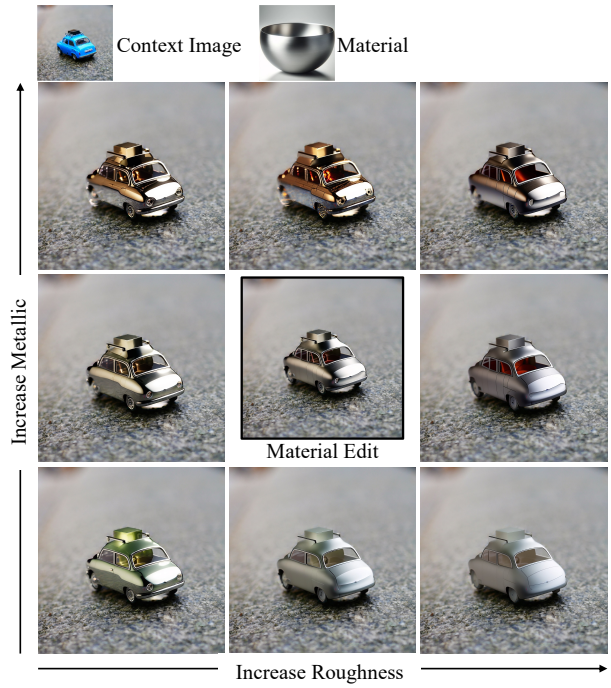

Figure 8. **Multiple controls at once**. With minimal tuning on the pre-trained components, MARBLE can perform material transfer and incorporate multiple attribute controls all in a single pass on real-world images. We present a grid of results of increasing roughness and metallic of a toy car, where we can see that the two attributes are properly disentangled from one another.

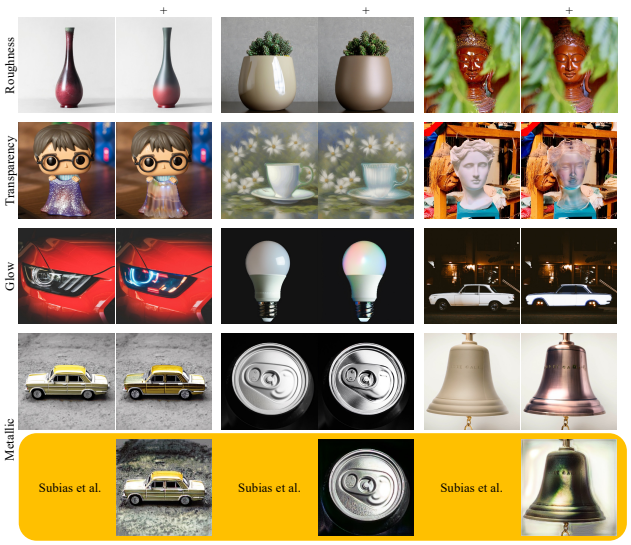

Figure 9. **Additional Results for Attribute Control.** We present 12 pairs of results on increasing attribute value (From left to right). As In-the-wild editing by Subias et al. also support metallic, we show the qualitative comparisons for the three examples. *Zoom in for details.*

operating in the CLIP space and not changing the pre-trained weights of the base diffusion model, MARBLE also shows capabilities to perform parametric control on various styles understood within the CLIP text features. Figure 10 presents two examples of paintings with different styles generated by SDXL. MARBLE changes the transparency and roughness of the foreground object while preserving the original style of the image. This is particularly evident on the wiggly brush strokes on the transparent pot mimicking the style of Van Gogh.

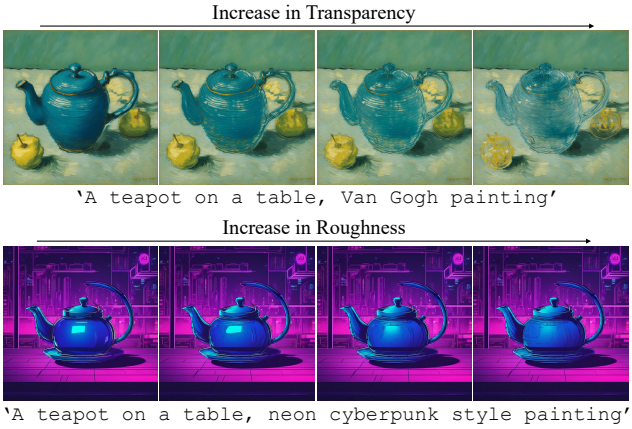

Figure 10. **Parametric control with different styles.** By leveraging the generalization capability of CLIP, our parametric controls can be also be adopted for images with various styles. We present parametric control over two styles of paintings generated by SDXL. Despite being trained on rendered images, the parametric controlled editing preserves the given style when presenting attribute changes.

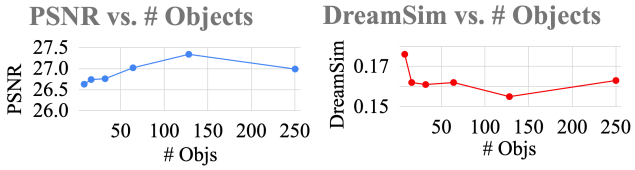

Figure 11. **Data efficiency.** We trained the transparency controller with 8, 16, 32, 64, 128 and 250 objects and present their PSNR and DreamSim scores. Even with as few as 16 objects, we can still obtain decent numbers on the validation dataset.

How small can the training dataset be? Furthermore, we investigate how small the training dataset can be by measuring the PSNR and DreamSim on synthetically rendered validation dataset (Figure 11). To our surprise, training on as few as 16 objects was sufficient to achieve similar results compared to using the full dataset. Qualitative results on real-world dataset are also presented in the Appendix.

Limitations. Our method has two main limitations, as shown in Figure 12. First, parametric control would sometimes change the textural patterns of an object, such as the pattern on the leather backpack of the left example. Second, the effect of the control causes undesired artifacts when the model is not expected to result in no change, as observed in the case of increasing transparency of the glass. These artifacts and loss of the high-frequency details can be caused

Increase Roughness Increase Transparency

Figure 12. **Limitations.** Our method has two primary limitations. (1) Sometimes performing parametric control also changes the texture patterns of the object such as the pattern on side of the leather backpack changes as roughness increases (left). (2) Sometimes the effects of the parametric control leads to artifacts.

due to multiple reasons such as the effect of noise pattern added to the latent of the context image, operations in the noisy CLIP-space, or the information loss in the encoding-decoding process of SDXL.

5. Conclusion

We present MARBLE, a method using CLIP-space for material editing in images. MARBLE builds on top of previous works in parametric and exemplar-based control, while adding a new blending mechanism, to allow flexibility of users to blend and recompose materials in a given image. The controls can be trained without finetuning generative model, and allows for multiple edits in one single forward pass. Overall, MARBLE presents an interesting direction for fine-grained, graphics-based controls of generative models revealing the advantages of CLIP-space representation for low-level controlled editing.

References

[1] Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. Practical svbrdf capture in the frequency domain. *ACM Trans. Graph.*, 32(4):110–1, 2013. 3

[2] Miika Aittala, Tim Weyrich, Jaakko Lehtinen, et al. Two-shot svbrdf capture for stationary materials. *ACM Trans. Graph.*, 34(4):110–1, 2015. 3

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2

[4] Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. *arXiv preprint arXiv:2403.17064*, 2024. 3

[5] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015. 3

[6] Shariq Farooq Bhat, Niloy J Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. *arXiv preprint arXiv:2312.03079*, 2023. 3

[7] 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, 3, 6

[8] 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

[9] Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4169–4181, 2023. 3

[10] Duygu Ceylan, Valentin Deschaintre, Thibault Groueix, Rosalie Martin, Chun-Hao Huang, Romain Rouffet, Vladimir Kim, and Gaëtan Lassagne. Matatlas: Text-driven consistent geometry texturing and material assignment. *arXiv preprint arXiv:2404.02899*, 2024.

[11] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 3

[12] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 2

[13] 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

[14] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[15] Ta-Ying Cheng, Prafull Sharma, Andrew Markham, Niki Trigoni, and Varun Jampani. Zest: Zero-shot material transfer from a single image. In *European Conference on Computer Vision*, pages 370–386. Springer, 2025. 2, 3

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 4

[17] 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*, pages 453–464. Wiley Online Library, 2022. 3

[18] Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. Flexible svbrdf capture with a multi-image deep network. In *Computer graphics forum*, pages 1–13. Wiley Online Library, 2019. 3

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

[20] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 6

[21] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. *arXiv preprint arXiv:2311.12092*, 2023. 3, 6

[22] 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

[23] Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. Texsliders: Diffusion-based texture editing in clip space. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[24] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 4

[25] 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

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

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

[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.

[29] 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.

[30] 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

[31] 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, 3

[32] Peter Kocsis, Vincent Sitzmann, and Matthias Nießner. Intrinsic image diffusion for indoor single-view material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5198–5208, 2024. 3

[33] 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

[34] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 5

[35] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[36] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022. 3

[37] Ivan Lopes, Fabio Pizzati, and Raoul de Charette. Material palette: Extraction of materials from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4379–4388, 2024. 3

[38] 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. 3

[39] 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

[40] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 4

[41] Karran Pandey, Paul Guerrero, Metheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. *CVPR*, 2024. 2

[42] Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pages 469–487. Springer, 2025. 3

[43] 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. 2, 6

[44] 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

[45] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3

[46] 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, 3

[47] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 2

[48] 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

[49] 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

[50] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2025. 2

[51] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. *arXiv preprint arXiv:2312.02970*, 2023. 2, 3, 4

[52] Prafull Sharma, Julien Philip, Michaël Gharbi, Bill Freeman, Fredo Durand, and Valentin Deschaintre. Materialistic: Selecting similar materials in images. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3

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

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

[55] Paul Upchurch and Ransen Niu. A dense material segmentation dataset for indoor and outdoor scene parsing. In *European Conference on Computer Vision*, pages 450–466. Springer, 2022. 3

[56] 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

[57] Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2, 3

[58] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. *arXiv preprint arXiv:2402.03290*, 2024. 2

[59] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2

[60] 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

[61] Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. *arXiv preprint arXiv:2401.09416*, 2024. 3

[62] 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. 3

[63] 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

[64] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3