# TexSliders: Diffusion-Based Texture Editing in CLIP Space

JULIA GUERRERO-VIU, Universidad de Zaragoza - I3A, Spain and Adobe Research, USA
MILOŠ HAŠAN, Adobe Research, USA
ARTHUR ROULLIER, Adobe Research, France
MIDHUN HARIKUMAR, Adobe, USA
YIWEI HU, Adobe Research, USA
PAUL GUERRERO, Adobe Research, United Kingdom
DIEGO GUTIERREZ, Universidad de Zaragoza - I3A, Spain
BELEN MASIA, Universidad de Zaragoza - I3A, Spain
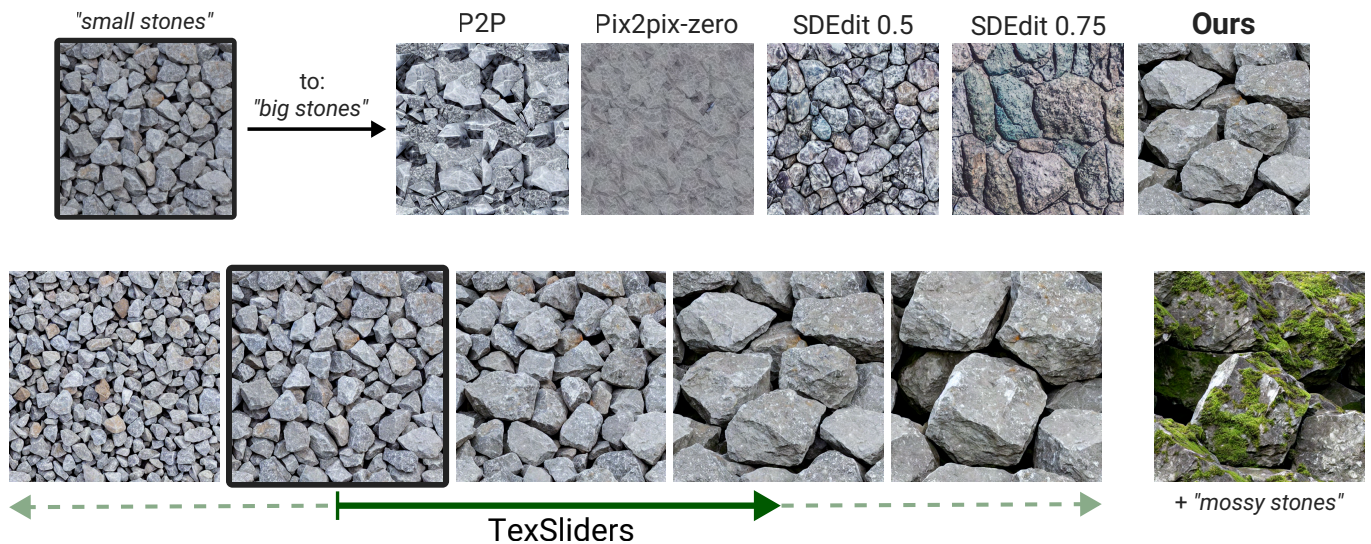VALENTIN DESCHAINTRE, Adobe Research, United Kingdom

Fig. 1. We present TexSliders, a diffusion-based method for texture editing. *Top row:* Given an input texture (highlighted in black) and a pair of natural language prompts describing an arbitrary edit (e.g., "small stones" to "big stones"), existing methods fail to either perform the edit or preserve the texture identity. *Bottom row:* Our approach finds a suitable editing direction in CLIP space [Radford et al. 2021], defining a *slider* that allows to manipulate the texture along such direction (positive and negative) while preserving its original identity. Moreover, our method allows to combine multiple editing directions, as the rightmost image shows ("small stones" to "big, mossy stones").

Generative models have enabled intuitive image creation and manipulation using natural language. In particular, diffusion models have recently shown remarkable results for natural image editing. In this work, we propose to apply diffusion techniques to edit *textures*, a specific class of images that are an essential part of 3D content creation pipelines. We analyze existing editing methods and show that they are not directly applicable to textures, since their common underlying approach, manipulating attention maps, is unsuitable for the texture domain. To address this, we propose a novel approach that instead manipulates CLIP image embeddings to condition the diffusion generation. We define editing directions using simple text prompts (e.g., "aged wood" to "new wood") and map these to CLIP image embedding space using a texture prior, with a sampling-based approach that gives us identity-preserving directions in CLIP space. To further improve identity preservation, we project these directions to a CLIP subspace that minimizes identity variations resulting from entangled texture attributes. Our editing pipeline facilitates the creation of arbitrary sliders using natural language prompts only, with no ground-truth annotated data necessary.

## 1 INTRODUCTION

Generative models have given us unprecedented ability to create and edit visual content using natural language alone. Text-to-image diffusion models are trained on hundreds of millions of images collected online, and usually offer great results for the kind of images found at scale on the Internet [Ramesh et al. 2022; Rombach et al. 2022]. However, using these models for domain-specific applications, such as texture generation, can pose new challenges.

In this work, we target the creation of textures using diffusion models, focusing specifically on texture editing. Textures are ubiquitous in image manipulation, graphic design, illustrations, rendering, and 3D modeling. While diffusion models can occasionally produce a surprisingly good texture out of the box, more precise texture manipulation remains a challenge. Recent work [Hertz et al. 2023; Parmar et al. 2023] proposed to leverage *attention maps* for general diffusion-based image editing, providing "weights" to focus the

editing on the desired region. While this produces impressive photograph editing results, attention maps are not nearly as informative in the context of texture. We demonstrate in the evaluation that this limits their value for texture editing.

We explore a different solution: texture manipulations in the CLIP [Radford et al. 2021] embedding space, akin to latent manipulations in GANs [Patashnik et al. 2021]. We demonstrate our method on various *editing directions* for textures such as weathering, scale, roughness, and more. Our approach allows to define new sliders for custom concepts with simple text prompts in a matter of minutes.

Our goal is to design a method with the following desirable properties: (1) the editing directions should be easy to define, (2) editing should preserve texture identity, (3) no ground-truth annotated data should be necessary, and (4) no modifications to the original diffusion model weights should be required.

In our work, we consider texture "identity" to include all attributes of an input texture except the ones being edited. To make the editing direction as easy as possible to define, we use two unrestricted textual prompts. We have experimented mostly with antithetic ones (e.g., "aged wood" to "new wood") and more specific ones (e.g., "metal" to "rusty metal"), but any prompts that define an editing direction could be used. We compute dozens of possible CLIP image embeddings for each of the two text prompts using a texture prior network [Aggarwal et al. 2023; Ramesh et al. 2022]. This gives us two clusters of embeddings, one for each prompt. We then compute the direction between the two cluster centers. Averaging over multiple CLIP image embeddings factors out identity from the editing attribute to a large extent, but we still observe some identity variation when taking steps along the editing direction. We empirically find that many dimensions do not contribute to the attribute that is being edited, but rather contain noise that results in these identity variations. Thus, we propose to automatically select a subset of the dimensions from the average direction based on per-dimension intra- and inter-cluster statistical distances, effectively improving identity preservation even for large step sizes.

A key finding during our work is that identity preservation is significantly easier in diffusion models trained on image embeddings (as opposed to trained on text embeddings). Intuitively, this is due to the loss of information that occurs when describing an image (or a texture) with text, compared to the actual information contained in the image itself. We therefore leverage a diffusion model trained with image embeddings for our editing method. This diffusion model is moreover guaranteed to always produce tileable textures.

In summary, we propose an editing approach for textures by manipulating CLIP image embeddings for a pre-trained image-conditioned diffusion model using a texture diffusion prior. This is enabled by the following contributions:

- Definition of CLIP-space editing directions from text prompts.
- Pruning of dimensions for improved identity preservation.
- Analysis of our editing directions' generalization, combinations, and use for generated images and real photographs.

Our method allows text-based texture editing through sliders constructed from simple natural language prompts, while preserving texture identity, high image quality, and tileability.

## 2 RELATED WORK

*Diffusion-Based Generation.* Diffusion models [Ramesh et al. 2022; Rombach et al. 2022] have recently revolutionized the field of image generation by creating high-quality image content from natural language prompts. They are based on the concept of diffusion probabilistic models [Ho et al. 2020; Nichol and Dhariwal 2021; Sohl-Dickstein et al. 2015; Song et al. 2021] and typically employ a UNet architecture that performs iterative denoising from random per-pixel initial noise. For efficiency, this can be done in a latent space of reduced dimensionality [Rombach et al. 2022]. Importantly, they also use a conditioning mechanism to guide the denoising from an input embedding [Dhariwal and Nichol 2021; Nichol et al. 2022]. This embedding can be a class [Dhariwal and Nichol 2021], a text prompt [Nichol et al. 2022], or an image [Ramesh et al. 2022]. Text and image prompts are typically encoded into vector embeddings by large vision-language models like CLIP [Radford et al. 2021]. UnCLIP [Ramesh et al. 2022] (the base architecture of DALL-E 2) uses image embeddings previously generated by a prior network from text embeddings. Aggarwal et al. [2023] show how to specialize this prior for domain-specific contexts, such as textures. In our work we address the problem of text-based, continuous texture *editing*, leveraging a latent diffusion model with a domain prior. Other works focus on texture generation on 3D shapes [Chen et al. 2023; Richardson et al. 2023].

*Diffusion-Based Image Editing.* Diffusion models are also widely applied to solve image editing problems. SDEdit [Meng et al. 2022] allows text-based image editing by leveraging a pretrained diffusion model, re-inserting noise and continuing diffusion towards a new condition. However, for larger edits, this approach cannot preserve the identity of the original image well, due to its stochastic nature. Recently, methods like Prompt-to-Prompt [Hertz et al. 2023], InstructPix2Pix [Brooks et al. 2023], Pix2Pix-Zero [Parmar et al. 2023], or Diffusion Self-Guidance [Epstein et al. 2023] have shown how manipulating cross-attention maps can help retain the structure of the original image. These maps could also be used in combination with self-attention and certain intermediate layers of the diffusion UNet to better preserve global consistency [Hertz et al. 2023; Tumanyan et al. 2023]. These methods can be applied to diffusion-generated images or to real photographs with the help of inversion [Mokady et al. 2023]. However, cross-attention maps rely on semantic information, which is usually missing in textures (see Figure 2), and we find that the performance of these methods on texture editing degrades. Imagic [Kawar et al. 2023] better preserves identity by fine-tuning both the text embedding and the weights of the diffusion model. While this gives compelling editing results, it requires a lengthy fine-tuning process for each target edit and each input image. Other approaches have been proposed to guide the editing by example images instead of text [Šubrtová et al. 2023]. Concurrently, Concept Sliders [Gandikota et al. 2023] proposes to find edit directions in the parameter space of a diffusion model through composable low-rank parameter updates. While we share a similar goal, we use the orthogonal approach of finding edit directions in the CLIP embedding space, rather than in the parameter space of the diffusion model, and our method does not require any fine-tuning or other changes to the diffusion model itself.

*Finding Directions in Generative Models.* Our work finds semantically meaningful directions in the latent space created by CLIP embeddings, which we leverage for texture editing. Exploring latent spaces of generative models aiming to disentangle semantic directions has been widely studied in the case of GANs [Härkönen et al. 2020; Wu et al. 2021]. This is more challenging for diffusion models, which typically lack such manageable space. Recent work has shown that a more convenient space can be found in pre-trained diffusion models through optimization [Kwon et al. 2022]. In contrast, our approach does not require optimization, simply requiring two text prompts to define a direction.

*Texture Generation and Editing.* Texture synthesis is a long standing challenge in computer graphics. In the last decade various deep learning approaches were proposed, targeting the synthesis of larger versions of input textures (e.g., [Henzler et al. 2020; Isola et al. 2017; Li et al. 2017; Mardani et al. 2020; Rodriguez-Pardo et al. 2019; Shaham et al. 2019; Ulyanov et al. 2016; Zhou et al. 2018]). A popular approach is procedural modeling where the textures are defined as a combination of noise, patterns, and filters functions [Adobe 2023; Guehl et al. 2020; Hu et al. 2022]. One key benefit of such representation is controllability: as each function is defined by a set of parameters, this representation lets artists expose semantic control as sliders. Procedural content is however challenging to author, requiring significant artistic and technical skills, since the parameters do not always correspond to intuitive concepts, and their interactions may be exceedingly complex to understand [Serrano et al. 2016]. Our approach targets the creation of similar sliders, enabling disentangled control over existing textures with simple text prompts, without the need for complex procedural representations.

## 3 DIFFUSION-BASED TEXTURE EDITING

In this section, we first show how practices that have led to great success in diffusion-based image editing are not well-suited to the particular case of texture manipulation. These observations motivate our choices when proposing a new method for texture editing, which we build upon an existing latent diffusion model without the need to re-train or fine-tune for editing directions.

### 3.1 Preliminary Observations

We review design decisions made when editing with diffusion models, and analyze their effectiveness in the particular case of textures.

*Cross-Attention Maps.* Previous work on image editing with diffusion models (Prompt-to-Prompt [Hertz et al. 2023] and subsequent follow-up works [Brooks et al. 2023; Mokady et al. 2023]) made the key observation that the spatial layout of the generated image depends on the *cross-attention maps*, which are spatial attention maps computed for each token of the text prompt (see Figure 2, top row). These maps help with identity preservation during the editing process, avoiding the need to mask a part of the image as often done in previous methods [Avrahami et al. 2023, 2022; Nichol et al. 2022], and which is impractical for textures. Cross-attention maps yield impressive results for non-texture images that typically have a clear structure with individual objects that correspond to phrases of the prompt. Attention maps encode this structure and
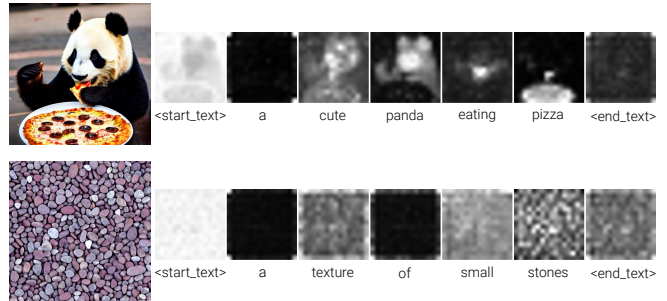


Fig. 2. **Visualization of cross-attention maps**. We show maps at the last diffusion step of SD 1.4 [Rombach et al. 2022], given two different input prompts. *Top:* "a cute panda eating pizza" (non-texture). *Bottom:* "a texture of small stones" (texture). The attention maps contain interesting semantic information for the panda image, but fail to capture the texture structure.
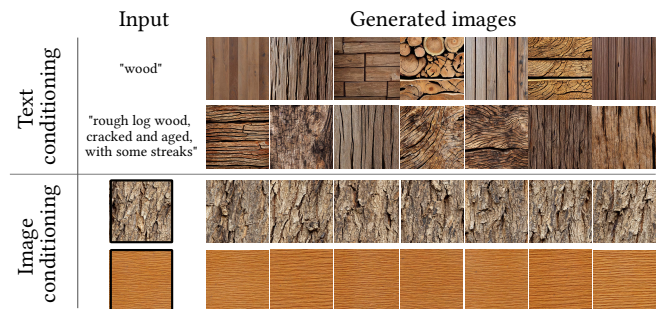


Fig. 3. **Text vs. image conditioning in diffusion models.** Images generated by a diffusion model conditioned with two text prompts (top rows) and images (bottom rows). In both cases, each column represents a different seed when sampling the diffusion model. Text conditioning, even with a specific prompt (second row), maps to a larger region in appearance space and can thus result in many different visual identities, while image conditioning maps the result to a more specific appearance. Text conditioning is done with SD 1.4 [Rombach et al. 2022], and image conditioning is done with the Latent Diffusion Model of Aggarwal et al. [2023].

can be used to both manipulate or preserve it. However, textures usually lack such a clear separation into individual objects, and thus cross-attention maps are unable to capture structure and map it to the textual prompt, failing to properly represent identity (Figure 2, bottom row). Consequently, our approach does not employ cross-attention maps; instead, it relies on finding a direction in CLIP embedding space that preserves identity, as explained in Section 3.3.

*Domain-Specific Generation.* Large pre-trained models have yielded remarkable performance on specific subspaces when fine-tuned on smaller datasets [Deschaintre et al. 2023; Gal et al. 2022a; Ruiz et al. 2023]. An alternative to fine-tuning an entire diffusion model is to attain domain-specific generation by means of a prior model trained on the specific data of interest [Aggarwal et al. 2023; Ramesh et al. 2022; Zhou et al. 2023a]. This prior model, given a textual input, outputs an embedding (an image representation) that can then be used as conditioning to a diffusion decoder to produce
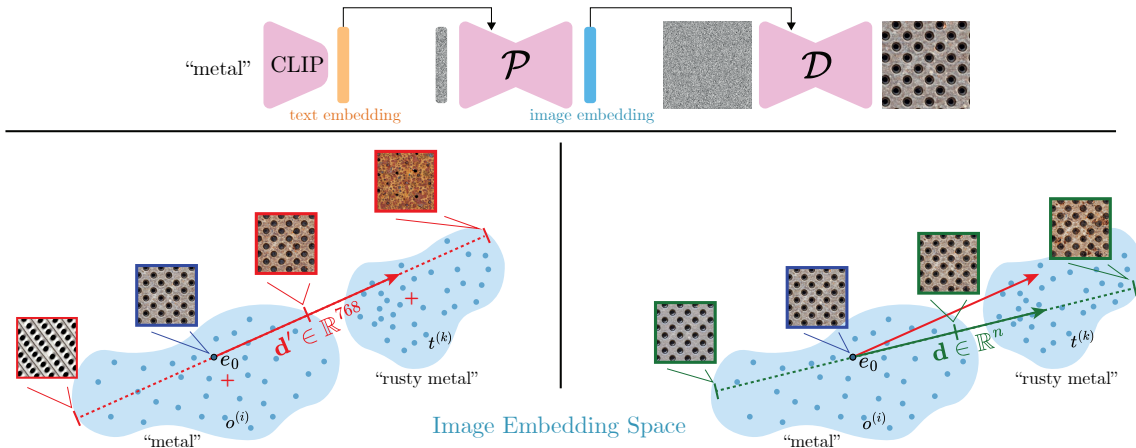
Fig. 4. **Overview of our diffusion-based texture editing approach.** *Top row:* Our approach leverages a diffusion prior model $\mathcal{P}$ to convert text embeddings to image embeddings, enabling the use of an image-conditioned pre-trained diffusion model $\mathcal{D}$. *Bottom row:* To perform the desired edits, we first compute direction $\mathbf{d}' \in \mathbb{R}^{768}$ (red arrow) as the difference between the centroids of the clusters (two small red crosses) formed by the image embeddings of the two prompts that define the edit (e.g., "metal" to "rusty metal"). Naively applying this direction to a specific texture $e_0$ (highlighted in blue), leads to significant identity variations as we march along such direction (left). Instead, we select a subset of $n$ relevant dimensions ($n < 768$) that do contribute to the desired edit, leading to our final editing direction $\mathbf{d}$ (green arrow), which preserves the identity of the input texture (right). We represent the high-dimensional CLIP image embedding space in 2D for visualization purposes.

the final image, constrained to the subdomain the prior was trained with.

Textures are a specific subdomain within the larger distribution of images typically learnt by diffusion models. To help preserve identity and to constrain the generation to textures, our approach employs a diffusion prior model trained on textures, described in Section 3.2, instead of fine-tuning the entire model.

*Manipulations in Image Embedding Space.* We seek to control the editing process by means of sliders with semantic meaning, which can be defined with simple text prompts. While the editing directions could thus be defined in text embedding space [Deckers et al. 2023; Ramesh et al. 2022], we observe that the notion of texture identity is more easily preserved in *image* embedding space. Intuitively, it is easier to define the appearance of a texture with images than with textual descriptions, since these typically cannot describe all the details that constitute its identity. An intuition of this is given in Figure 3, where we compare text- and image-conditioned generation, and show that even a detailed textual prompt leads to a much higher variance in the generated images than using image conditioning.

## 3.2 Diffusion Model

We use a Latent Diffusion Model alongside a *domain diffusion prior* [Aggarwal et al. 2023; Ramesh et al. 2022] trained for the texture domain, as described in Section 3.1.

The domain diffusion prior is a diffusion model trained to generate CLIP *image* embeddings matching a given CLIP *text* embedding, as shown in Figure 4 (top). Note that this is a generative process, as there are generally multiple image embeddings matching a text embedding. More specifically, the diffusion prior is trained to generate image embeddings in the "texture" part of the CLIP L/14 embedding space, encouraging the generation of texture-like images from the

main, pre-trained latent diffusion model. The diffusion model is trained on a dataset of 77M images with no humans or text, and the prior is trained separately on a 10M subset of that dataset, classified as textures. The prior further allows the use of text prompts to interact with a network trained on image embeddings, while retaining high generation quality and prompt alignment.

## 3.3 Computing Sliders for Texture Editing

As described above, our method employs a diffusion model $\mathcal{D}$ coupled with a texture diffusion prior $\mathcal{P}$, and performs the editing in image embedding space. We thus seek to define a *direction* $\mathbf{d}$ in such space, specified by a pair of simple text prompts that describe the original and target appearance (e.g., from "metal" to "rusty metal"), and which will act as a slider: marching along such direction (positive and negative) will progressively increase or decrease the intensity of the desired edit.

Figure 4 (bottom) illustrates this. We first compute the CLIP text embeddings of the original and target prompts, and feed them to the prior $\mathcal{P}$, yielding image embeddings within the texture domain that fit the textual descriptions. In order to obtain a robust representation of the editing prompts, we compute a set of $n_e$ image embeddings for the *original* and *target* prompts by sampling the diffusion prior. We term these image embeddings $\mathbf{o}^{(i)}$ and $\mathbf{t}^{(k)}$, respectively, with both $i$ and $k \in \{1..n_e\}$. The number $n_e$ of image embeddings is an adjustable parameter that we set to 150 in our experiments for both the original and target embeddings. We then define an initial editing direction $\mathbf{d}'$ in image embedding space as the difference between the centroids of the clusters formed by the original and target embeddings. Note that $\mathbf{d}' \in \mathbb{R}^{768}$, as it corresponds to the dimensionality of the image embedding obtained from the diffusion prior model.

Each component $d'_j$ is given by:

$$d'_j = \frac{1}{n_e}\left(\sum_k t_j^{(k)} - \sum_i o_j^{(i)}\right). \tag{1}$$

Computing multiple image embeddings to obtain this initial direction aids in disentangling the relevant attribute(s) from the rest but does not suffice, leading to poor results (see Figure 4, bottom left). To better preserve the identity of the input texture while progressively changing only the desired attribute, we need to select the subset of relevant dimensions, avoiding those that do not contribute to the desired edit, or lead to identity variations.

We select the relevant dimensions based on their intra-cluster variability, as given by the standard deviation $std$, compared to their inter-cluster variability, as given by the distance between cluster centroids. Dimensions with high inter-cluster variability typically contribute more to the desired edit, while dimensions with high intra-cluster variability typically encode the identity of each individual texture. We therefore select those dimensions whose inter-cluster variability is larger than the intra-cluster one, as those are more likely to be representative of the edited attribute only. We set the remaining dimensions to zero. The components of our resulting direction vector $\mathbf{d}$ (shown in Figure 4, bottom right) are thus:

$$d_j = \begin{cases} d'_j, & \text{if } |\tilde{d}'_j| > \tau \cdot std_k(\tilde{t}_j^{(k)}) \text{ and } |\tilde{d}'_j| > \tau \cdot std_i(\tilde{o}_j^{(i)}) \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The relationship is modulated by a threshold $\tau$ (typically 0.8 in our experiments), and applied over normalized vectors $\tilde{\mathbf{d}}'$, $\tilde{\mathbf{t}}$ and $\tilde{\mathbf{o}}$, so that the comparison is meaningful.

Given $\mathbf{d}$, we can march along that direction to obtain different degrees of the desired edit, for instance by means of a slider. In particular, given the image embedding $\mathbf{e}_0$ of a texture we wish to edit, the final image embedding $\mathbf{e}_\alpha$ becomes:

$$\mathbf{e}_\alpha = \mathbf{e}_0 + \alpha \cdot \mathbf{d}, \tag{3}$$

where $\alpha$ modulates the intensity of the edit, and can take positive or negative values. The resulting $\mathbf{e}_\alpha$ is then used as conditioning in the diffusion model $\mathcal{D}$ to generate the final, edited texture image.

*Definition of New Sliders.* Besides using pre-computed sliders for editing, users can easily create new ones adapted to their needs by providing two text prompts; defining a new slider takes about 2 minutes on a single A10G GPU (in our unoptimized implementation).

## 4 RESULTS

We evaluate our diffusion-based editing approach on textures representing different kinds of materials, with various editing directions. We first explain the details of our test dataset containing both diffusion-based generated and real photographs of textures (Section 4.1). We qualitatively evaluate our results on *generated* images in Section 4.2, including an ablation study of the main components of our method, and additional results on real *photographs* of textures in Section 4.3. In Section 4.4, we compare to alternative diffusion-based image editing approaches, including a quantitative comparison to previous works. Finally, Section 4.5 includes further

evaluation about the generality of our method for multiple practical applications. Additional results and implementation details are available in the Supplemental Material.

### 4.1 Test Dataset

To exhaustively evaluate our approach on a diverse set of textures, and given the lack of standard benchmarks for this purpose, we create a test dataset of generated and real photographs of textures, paired with editing attributes. We first define a set of seven material types, typically found in 3D creation tools: *bricks*, *fabric*, *leather*, *metal*, *paper*, *stones*, and *wood*. Then, we define between one and four prompts for nine of the main editing attributes that people use to describe materials [Deschaintre et al. 2023], such as "rough" and "smooth" for the attribute *touch*, or "aged" and "rusty" for *weathering*. We filter the {material, attribute} pairs that are semantically valid, and generate two different textures for each pair, using the original prompt formed by material and attribute (e.g., "smooth stones", or "aged wood") and two different random seeds for the diffusion prior. Finally, we add two real photographs of textures for each material type (from Adobe Stock [2023]), making a total of 80 images (66 generated and 14 real) associated to a pair of original and target prompts. Our full dataset is included in the Supplemental Material.

### 4.2 Qualitative Evaluation

Our editing method is conceptually simple and zero-shot, since it does not require ground-truth data, nor training or fine-tuning of the diffusion model. We evaluate its performance on our test dataset, and show some representative results in Figure 5. Since the output of our pipeline is a *direction* in the embedding space that conditions the diffusion model, we can obtain different degrees of the desired edit by marching along the direction (varying $\alpha$ in Eq. 3), even extrapolating in the forward ($\alpha > 1.0$, rightmost column) and backward ($\alpha < 0.0$, first and second columns) directions. Our method successfully performs challenging edits, such as those related to weathering ("aged", "rusty", or "mossy"), metallicness, or touch ("rough" to "smooth"), on various material types. Importantly, most of these edits require globally changing all regions of the texture, which we accomplish while preserving the identity of the input texture.

*Ablation Study.* We ablate our key design choices and evaluate how they impact performance, in terms of disentanglement of the editing attributes and identity preservation. Figure 6 shows results for the "metal" to "rusty metal" example. In the top row, we use a *single* image embedding given the original and target prompts ($n_e = 1$); the results fail to preserve the identity of the input texture, due to the entanglement of attributes along the resulting direction. The middle row shows results using several image embeddings, but using the direction $\mathbf{d}'$, i.e., including *all* dimensions in CLIP space ($\mathbf{d}' \in \mathbb{R}^{768}$). Although the editing improves, the identity of the input texture is still not properly maintained. In contrast, our approach (bottom row) selecting a *subset* of relevant dimensions yields good edits while preserving the identity. Further ablations, including different methods to select the subset of relevant dimensions, can be found in the Supplemental Material.

Fig. 5. **Qualitative results.** We show our method on different kinds of materials for various editing directions. We can see that our method applies convincing editing, including when extrapolating the directions (two leftmost columns and rightmost column), and preserves well the texture identity. Here we use generated input textures; Figure 7 and the Supplemental Material provide results also on photographed inputs.
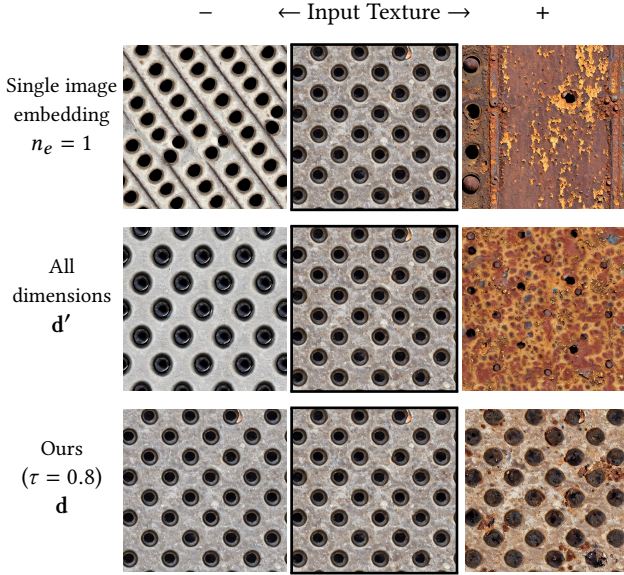
Fig. 6. **Ablation study**. We show qualitative ablations for the direction "metal" to "rusty metal" (positive and negative). We compare to using a single image embedding for original and target prompts ($n_e = 1$, top row) and to using direction $\mathbf{d}'$, computed for multiple image embeddings ($n_e = 150$), but including all the 768 dimensions of the direction (middle row). Both options show a higher identity shift than our approach (bottom row), which successfully edits the texture maintaining its original structure.

## 4.3 Real Photographs

Our results are not limited to diffusion-based generated textures. Since the diffusion model is conditioned on CLIP image embeddings, and these are a faithful representation of the texture appearance (see Figure 3), we can use the CLIP embedding of any input image as conditioning to reconstruct it. From this embedding and a pair of prompts, we can use our method to compute the editing direction $\mathbf{d}$ and generate textures with different degrees of editing. We show examples of these results using real photographs in Figure 7, where our method successfully edits different material types and attributes, such as wetness and smoothness. While out of the scope of this work, we could further improve the accuracy of the reconstruction, and hence our editing results, by inverting the diffusion model (e.g., using null-text inversion [Mokady et al. 2023]).

## 4.4 Comparison to the State of the Art

Last, we compare our approach to previous state-of-the-art diffusion-based image editing methods. To the best of our knowledge, ours is the first diffusion-based editing framework targeting the domain of textures, so we compare to general methods for natural images. Since these methods use Stable Diffusion [Rombach et al. 2022] as generative model, and do not include a textures domain prior, for fair comparison, we add the preamble "a texture of…" to all prompts in our test dataset. With that, we restrict the generated images of Stable Diffusion to the textures subspace. We use the same number of diffusion steps, 50, and the same classifier-free guidance (guidance scale = 7.5) for all methods.

Table 1. **Quantitative metrics** (mean values on our full test dataset, including generated images and photographs). Our approach's results better match the required edit direction (CLIP-Direction) than previous work, while also better preserving the identity of the input texture (CLIP-Im2Im).

| Method | ↑ CLIP-Direction | ↑ CLIP-Im2Im |
|---|---|---|
| P2P | 0.0880 | 0.9033 |
| Pix2pix-zero | 0.0811 | 0.8667 |
| SDEdit 0.5 | 0.0706 | 0.8753 |
| SDEdit 0.75 | 0.0690 | 0.8707 |
| Ours | **0.1063** | **0.9303** |

We include results of SDEdit [Meng et al. 2022][1] at two different noise levels (0.5 and 0.75), Prompt-to-Prompt (P2P) [Hertz et al. 2023][2], and Pix2pix-zero [Parmar et al. 2023][3], the last two using cross-attention maps to keep structural information.

As shown in Figure 8, our results exhibit both strong identity preservation, as well as good adherence to the target prompt. Contrarily, SDEdit shows an undesirable trade-off between both, where if using a low noise level (0.5) the identity is preserved but the appearance is hardly edited (e.g., "mossy bricks"), or if the noise level is increased to 0.75 the edit is performed but the identity changes noticeably (e.g., "rusty metal"). Pix2pix-zero and P2P both demonstrate clear loss of identity in several cases, or even some noticeable visual artifacts that push the result out of the texture domain (e.g., "big stones"), due to their reliance on cross-attention maps that are unsuitable for texture editing. We also outperform previous methods on real photographs (last row), despite the fact that our approach does not use inversion for reconstruction, unlike P2P and Pix2pixzero. Remarkably, our approach is the only one able to successfully perform scale changes in the pattern (e.g., "small stones" to "big stones"), as in that case texture identity is not defined by the spatial structure.

*Quantitative Analysis.* As a perceptual task, it is not trivial how to quantitatively measure the quality of an edit (both in terms of adherence to the edit and identity preservation). In order to give an objective comparison, we follow previous works and use metrics in CLIP embedding space, as follows:

- CLIP-Direction: Measures edit adherence as cosine similarity between the directions in CLIP space of the pre/post edit text prompts and images [Brooks et al. 2023; Gal et al. 2022b].
- CLIP-Im2Im: Measures identity preservation as cosine similarity in CLIP space between input and edited image [Brooks et al. 2023].

We show in Table 1 that our approach significantly outperforms previous work in both metrics, supporting our conclusions from the qualitative comparison.
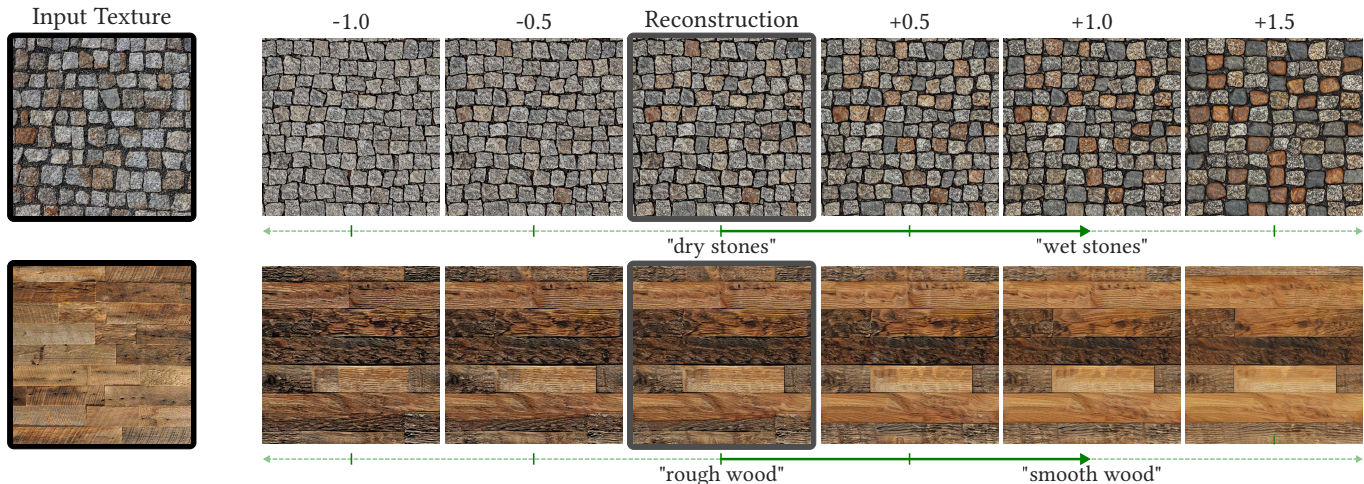
---

Fig. 7. **Photographed texture edits.** We use the CLIP embedding of the input texture to reconstruct the photograph using the diffusion model, and apply our editing approach to the reconstructed texture (highlighted in grey), showing that our approach can be used for non-synthetic textures as well.

## 4.5 Generality

We further evaluate our approach's generality along three axes: (a) generalization of edit directions across material types, (b) combination of multiple edit directions, and (c) using images instead of text to describe the editing direction.

To evaluate (a), we apply directions computed for a certain material to a different one, showing reasonable results in exchange for a small identity loss (see Figure 9, left, and the Supplemental Material). We demonstrate (b) through full matrices of edit combinations in the Supplemental Material. The diagonal of the matrix that combines scale and mossiness for stones (from (-1,-1) to (+1.5,+1.5)) is shown in Figure 9, right. Last, regarding (c), when an editing direction cannot be captured well by two textual prompts but we have instead a set of image pairs that illustrate the change, we can first obtain the CLIP embeddings of such set of images and then apply our method to compute the editing direction. This is possible because our approach operates in CLIP image space, and generalizes well across multiple material types. We show successful results using 20 images of textures cropped and resized to define both original ("zoomed in") and target ("zoomed out") clusters in the Supplemental Material.

## 5 DISCUSSION AND LIMITATIONS

We propose TexSliders, a diffusion-based texture editing approach that enables slider-like disentangled editing for customized directions. To do this, we leverage image-conditioned diffusion models together with a texture diffusion prior and explore the CLIP image space to derive identity-preserving editing directions. We show that our simple yet effective solution is superior to other image editing methods when applied to the texture domain.

Our method is not free of limitations. Since we leverage CLIP space for diffusion-based texture editing, our approach inherits some limitations from the diffusion model and the CLIP representation. For some cases, the intended editing may fail to change the texture appearance significantly, and a key observation in this regard is that CLIP and the diffusion model can be more sensitive to some concepts compared to others. For example, we found that CLIP represents the concept "vertical" better than "horizontal"; as a result, the CLIP-conditioned generative model biases the generated textures toward vertically-oriented patterns, and textures with vertical stripes can struggle to turn horizontal despite following the corresponding editing direction (Figure 10, top row). We also note that in some cases the identity of the input texture is not perfectly preserved during the editing; an example is shown in Figure 10 (middle row), where the edit is performed, but the pattern of the cobblestones is not completely preserved. This may be due to noise or spurious dimensions remaining present in the editing direction, and other methods for dimension selection could be explored. Identity preservation can also be hampered when extrapolating in the editing direction too far from the input texture, leading to embeddings outside the sampled distribution (e.g., last column in Figure 10, bottom row).

One must note that the notion of identity in textures is not a formally defined one, even if it is arguably easy to judge qualitatively. For our purposes, identity depends on the input texture and the attributes being edited, and it includes all other attributes of the input texture except the ones being edited. We leave an in-depth study of texture identity in diffusion-based models for future work. We focus on color textures in this work, from which material maps could be approximated [Sartor and Peers 2023; Vecchio et al. 2023]. Extending our editing approach to such material maps should be possible, although current material diffusion models rely on a relatively limited amount of data (orders of magnitude smaller than for images), which would limit their expressivity. Better understanding the CLIP embedding space for images [Zhou et al. 2023b], and for the particular case of textures [Deschaintre et al. 2023; Wu and Maji 2022], is also an interesting line of research, both for synthesis and editing. Finally, the integration of our approach within a complete user interface would allow for artists to provide feedback about its usability and desirable improvements.
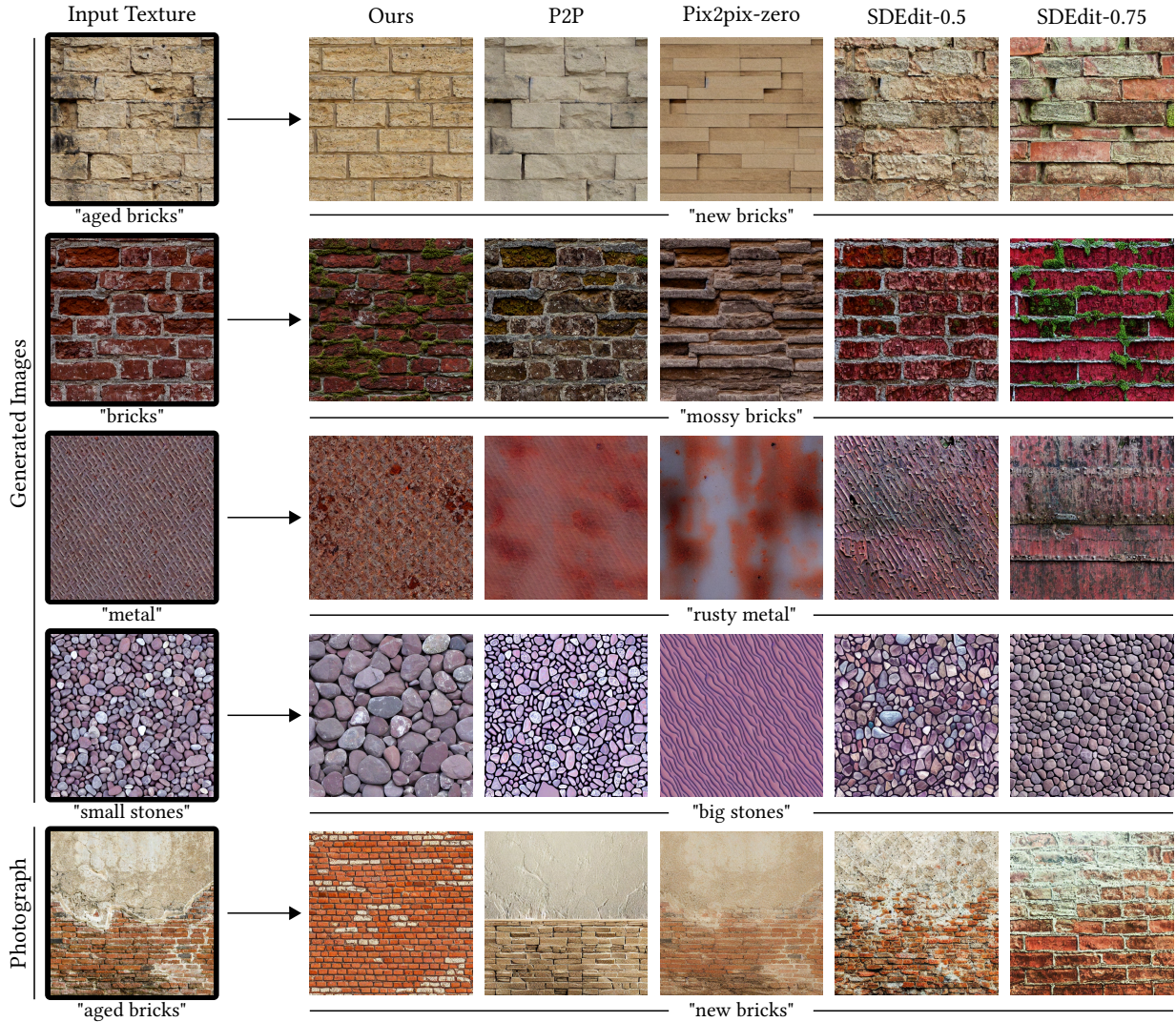
**Fig. 8. Comparisons to state-of-the-art methods.** For every input image, and pair of prompts that define an editing direction, we show the output of the following diffusion-based editing methods (from left to right): Ours ($\tau = 0.8$, $\alpha = 1.5$), Prompt-to-Prompt (P2P) [Hertz et al. 2023] (cross_replace_steps = 0.8, self_replace_steps = 0.3), Pix2pix-zero [Parmar et al. 2023] (xa_guidance = 0.5), and SDEdit [Meng et al. 2022] with noise levels 0.5 and 0.75, respectively. We can see that our approach leads to much better identity preservation and realistic editing, following the desired direction.
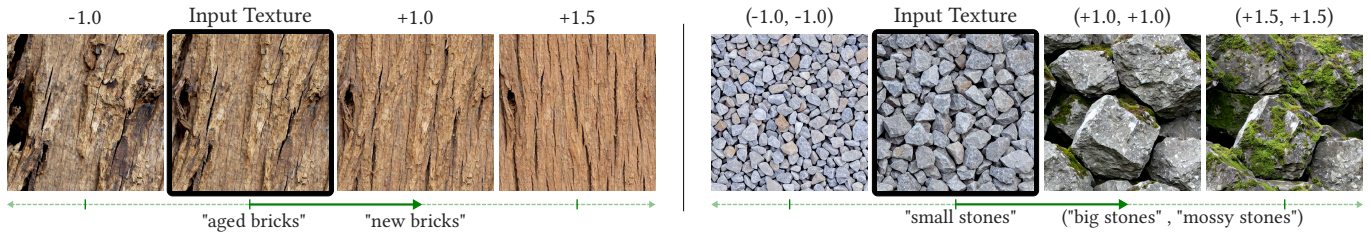


**Fig. 9. Generalization and multi-dimensional edits.** *Left:* We show that using directions computed for bricks ("aged" → "new") generalize to other classes such as wood. *Right:* We evaluate the combination of multiple editing directions ("small stones" → "big stones" and "mossy stones"). We show here the diagonal of the matrix of edits, where both directions are changed by the same amount (see complete matrices of edits and additional results in the Supplemental Material).
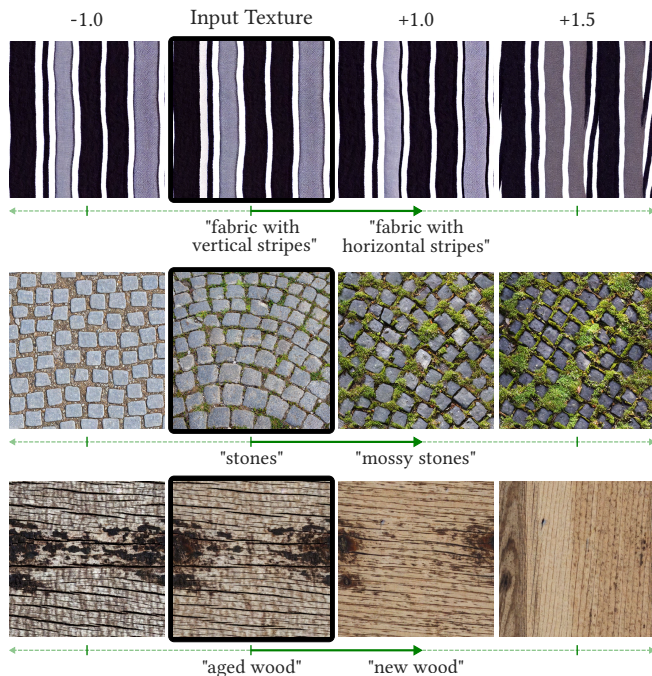
Fig. 10. **Limitations.** *Top*: Biases in CLIP and the diffusion model, such as an existing bias towards vertically-oriented patterns, can lead to failed editing. *Middle*: The identity of an input texture (e.g., the arrangement of the cobblestones) may be altered due to remaining noise in the editing direction. *Bottom*: Textures may not be guaranteed to preserve identity when extrapolated too far from the input image, leading to embeddings that are outside the sampled CLIP distributions.

Our method enables pre-computation of editing directions that can be used as sliders to edit specific attributes of any input texture. Besides, one can compute any new custom direction specified via two unrestricted text prompts—or alternatively with sample images, see Section 4.5—in under two minutes. Our work is not only the first to propose diffusion-based editing for the domain of textures, but also has the potential to be adapted to other domains, such as vector graphics-like images.

# REFERENCES

Adobe. 2023. Adobe Substance Designer.

Adobe Stock. 2023. https://stock.adobe.com/.

Pranav Aggarwal, Hareesh Ravi, Naveen Marri, Sachin Kelkar, Fengbin Chen, Vinh Khuc, Midhun Harikumar, Ritiz Tambi, Sudharshan Reddy Kakumanu, Purvak Lapsiya, Alvin Ghouas, Sarah Saber, Malavika Ramprasad, Baldo Faieta, and Ajinkya Kale. 2023. Controlled and Conditional Text to Image Generation with Diffusion Prior. *arXiv preprint arXiv:2302.11710* (2023).

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–11.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-Driven Editing of Natural Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18208–18218.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18392–18402.

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

Niklas Deckers, Julia Peters, and Martin Potthast. 2023. Manipulating Embeddings of Stable Diffusion Prompts. *arXiv preprint arXiv:2308.12059* (2023).

Valentin Deschaintre, Julia Guerrero-Viu, Diego Gutierrez, Tamy Boubekeur, and Belen Masia. 2023. The Visual Language of Fabrics. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–15.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. *arXiv preprint, arXiv:2105.05233* (2021).

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. 2023. Diffusion Self-Guidance for Controllable Image Generation. *Advances in Neural Information Processing Systems* (2023).

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618* (2022).

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022b. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. 2023. Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models. *arXiv preprint arXiv:2311.12092* (2023).

Pascal Guehl, Remi Allègre, Jean-Michel Dischler, Bedrich Benes, and Eric Galin. 2020. Semi-Procedural Textures Using Point Process Texture Basis Functions. *Computer Graphics Forum* (2020).

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems* 33 (2020), 9841–9850.

Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. 2020. Learning a Neural 3D Texture Space from 2D Exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8356 – 8364.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations (ICLR)*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Yiwei Hu, Chengan He, Valentin Deschaintre, Julie Dorsey, and Holly Rushmeier. 2022. An Inverse Procedural Modeling Pipeline for SVBRDF Maps. *ACM Transactions on Graphics (TOG)* 41, 2 (2022), 1–17.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 5967–5976.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2022. Diffusion Models Already Have A Semantic Latent Space. In *International Conference on Learning Representations*.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems*.

Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, and Bryan Catanzaro. 2020. Neural ffts for universal texture image synthesis. *Advances in Neural Information Processing Systems* 33 (2020).

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *International Conference on Learning Representations*.

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

Alex Nichol and Prafulla Dhariwal. 2021. Improved Denoising Diffusion Probabilistic Models. *International Conference on Machine Learning (ICML)* (2021).

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*. 16784–16804.

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

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)* (2021), 8748–8763.

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

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. Texture: Text-guided texturing of 3d shapes. *ACM SIGGRAPH Conference Proceedings* (2023).

Carlos Rodriguez-Pardo, Sergio Suja, David Pascual, Jorge Lopez-Moreno, and Elena Garces. 2019. Automatic extraction and synthesis of regular repeatable patterns. *Computers & Graphics* 83 (2019), 33–41.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

Sam Sartor and Pieter Peers. 2023. MatFusion: A Generative Diffusion Model for SVBRDF Capture. In *ACM SIGGRAPH Asia Conference Proceedings (SA '23)*. Association for Computing Machinery, New York, NY, USA, Article 86.

Ana Serrano, Diego Gutierrez, Karol Myszkowski, Hans-Peter Seidel, and Belen Masia. 2016. An intuitive control space for material appearance. *ACM Transactions on Graphics (TOG)* 35, 6 (2016).

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. 2019. SinGAN: Learning a Generative Model From a Single Natural Image. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 4569–4579.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*. 2256–2265.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Machine Learning (ICML)*.

Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. 2023. Diffusion image analogies. In *ACM SIGGRAPH Conference Proceedings*. 1–10.

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

Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. 2016. Texture Networks: Feed-Forward Synthesis of Textures and Stylized Images. *International Conference on Machine Learning (ICML)*, 1349–1357.

Giuseppe Vecchio, Rosalie Martin, Arthur Roullier, Adrien Kaiser, Romain Rouffet, Valentin Deschaintre, and Tamy Boubekeur. 2023. ControlMat: Controlled Generative Approach to Material Capture. *arXiv preprint arXiv:2309.01700* (2023).

Chenyun Wu and Subhransu Maji. 2022. How well does CLIP understand texture? *arXiv preprint arXiv:2203.11449* (2022).

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12863–12872.

Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. 2023b. CLIP-PAE: Projection-Augmentation Embedding to Extract Relevant Features for a Disentangled, Interpretable and Controllable Text-Guided Face Manipulation. In *ACM SIGGRAPH Conference Proceedings*. 1–9.

Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023a. Shifted diffusion for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10157–10166.

Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2018. Non-stationary texture synthesis by adversarial expansion. *ACM Transactions on Graphics (TOG)* 37 (2018), 1 – 13.