# Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

Narek Tumanyan* Michal Geyer* Shai Bagon Tali Dekel
Weizmann Institute of Science

Project webpage: https://pnp-diffusion.github.io

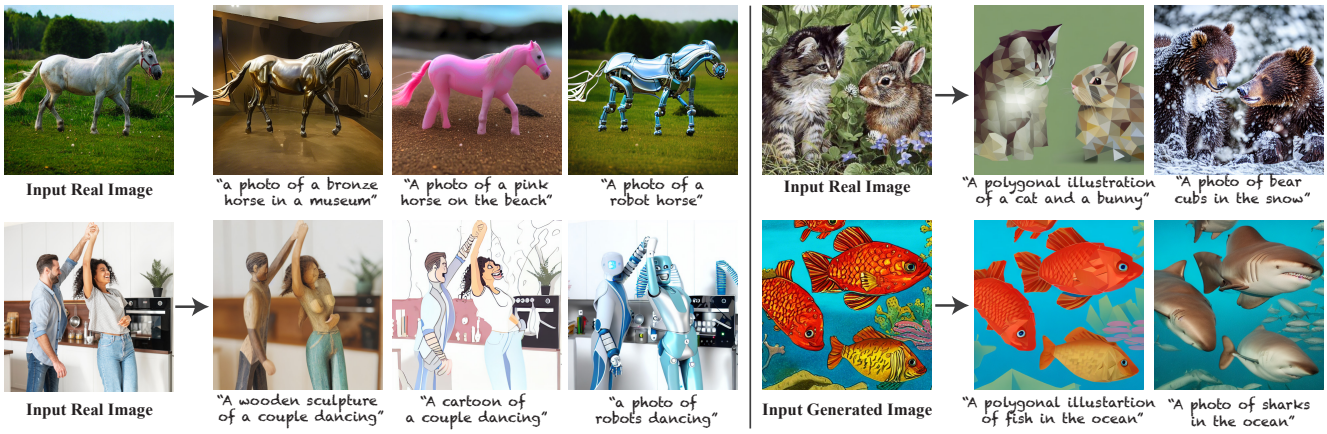


Figure 1. Given a single real-world image as input, our framework enables versatile text-guided translations of the original content. Our results exhibit high fidelity to the input structure and scene layout, while significantly changing the perceived semantic meaning of objects and their appearance. Our method does not require any training, but rather harnesses the power of a pre-trained text-to-image diffusion model through its internal representation. We present new insights about deep features encoded in such models, and an effective framework to control the generation process through simple modification of these features.

## Abstract

*Large-scale text-to-image generative models have been a revolutionary breakthrough in the evolution of generative AI, synthesizing diverse images with highly complex visual concepts. However, a pivotal challenge in leveraging such models for real-world content creation is providing users with control over the generated content. In this paper, we present a new framework that takes text-to- image synthesis to the realm of image-to-image translation – given a guidance image and a target text prompt as input, our method harnesses the power of a pre-trained text-to-image diffusion model to generate a new image that complies with the target text, while preserving the semantic layout of the guidance image. Specifically, we observe and empirically demonstrate that fine-grained control over the generated structure can be achieved by manipulating spatial features and their self-attention inside the model. This results in a simple and effective approach, where features extracted from the guidance image are directly injected into the generation process of the translated image, requiring no training or fine-tuning. We demonstrate high-quality results on versatile text-guided image translation tasks, including translating sketches, rough drawings and animations into realistic images, changing the class and appearance of objects in a given image, and modifying global qualities such as lighting and color.*

## 1. Introduction

With the rise of text-to-image foundation models – billion-parameter models trained on a massive amount of text-image data, it seems that we can translate our imagination into high-quality images through text [13, 35, 37, 41]. While such foundation models unlock a new world of creative processes in content creation, their power and expressivity come at the expense of user controllability, which is largely restricted to guiding the generation solely through an input text. In this paper, we focus on attaining control over the generated structure and semantic layout of the scene – an imperative component in various real-world content creation tasks, ranging from visual branding and marketing to digital art. That is, our goal is to take text-to-image generation to the realm of text-guided Image-to-Image (I2I) translation, where an input image guides the layout (e.g., the structure of the horse in Fig. 1), and the text guides the perceived semantics and appearance of the scene (e.g., "robot horse" in Fig. 1).

A possible approach for achieving control of the generated layout is to design text-to-image foundation models that explicitly incorporate additional guiding signals, such as user-provided masks [13, 29, 35]. For example, recently

* Equal contribution.

Make-A-Scene [13] trained a text-to-image model that is also conditioned on a label segmentation mask, defining the layout and the categories of objects in the scene. However, such an approach requires an extensive compute as well as large-scale text-guidance-image training tuples, and can be applied at test-time to these specific types of inputs. In this paper, we are interested in a unified framework that can be applied to versatile I2I translation tasks, where the structure guidance signal ranges from artistic drawings to photo-realistic images (see Fig. 1). Our method does not require any training or fine-tuning, but rather leverages a pre-trained and *fixed* text-to-image diffusion model [37].

We pose the fundamental question of how structure information is internally encoded in such a model. We dive into the intermediate spatial features that are formed during the generation process, empirically analyze them, and devise a new framework that enables fine-grained control over the generated structure by applying simple manipulations to spatial features inside the model. Specifically, spatial features and their self-attentions are extracted from the guidance image, and are directly injected into the text-guided generation process of the target image. We demonstrate that our approach is not only applicable in cases where the guidance image is generated from text, but also for real-world images that are inverted into the model.

To summarize, we make the following key contributions:
(i) We provide new empirical insights about internal spatial features formed during the diffusion process.

(ii) We introduce an effective framework that leverages the power of pre-trained and fixed guided diffusion, allowing to perform high-quality text-guided I2I translation without any training or fine-tuning.

(iii) We show, both quantitatively and qualitatively that our method outperforms existing state-of-the-art baselines, achieving significantly better balance between preserving the guidance layout and deviating from its appearance.

## 2. Related Work

**Image-to-image translation.** Image-to-Image (I2I) translation is aimed at estimating a mapping of an image from a source domain to a target domain, while preserving the domain-invariant characteristics of the input image, e.g., objects' structure or scene layout. From classical to modern data-driven methods, numerous visual problems have been formulated and tackled as an I2I task (e.g., [8,11,18,33,43]). Seminal deep-learning-based methods have proposed various GAN-based frameworks to encourage the output image to comply with the distribution of the target domain [23, 30, 31, 51]. Nevertheless, these methods require datasets of example images from both source and target domains, and often require training from scratch for each translation task (e.g., horse-to-zebra, day-to-night, summer-to-winter). Other works utilize pre-trained GANs by performing the translation in its latent space [1, 36, 46]. Several methods have also considered the task of zero-shot I2I by training a generator on a single source-target image pair example [47, 49]. With the advent of unconditional image diffusion models, several methods have been proposed to adopt or extend them for various I2I tasks [40, 50]. In this paper, we consider the task of *text-guided image-to-image translation* where the target domain is not specified through a dataset of images but rather via a target text prompt. Our method is zero-shot, does not require training and is applicable to versatile I2I tasks.

**Text-guided image manipulation.** With the tremendous progress in language-vision models, a surge of methods have been proposed to perform various types of text-driven image edits. Various methods have proposed to combine CLIP [34], which provides a rich and powerful joint image-text embedding space, with a pre-trained unconditional image generator, e.g., a GAN [7, 15, 27, 32] or a diffusion model [2, 3, 22, 25]. For example, DiffusionCLIP [22] uses CLIP to fine-tune a diffusion model to perform text guided manipulations. Concurrent to our work, [25] uses CLIP and semantic losses of [47] to guide a diffusion process to perform I2I translation. Aiming to edit the appearance of objects in real-world images, Text2LIVE [5] trains a generator on a single image-text pair, without additional training data;

Recently, text-to-image generative models [13,29,35,37, 41] have demonstrated unprecedented capabilities in generating high-quality and diverse images from text, capturing complex visual concepts (e.g., object interactions, geometry, or composition). Nevertheless, such models offer little control over the generated content. This creates a great interest in developing methods to adopt such unconstrained text-to-image models for controlled content creation.

For example, SDEdit [28] edits user-provided images using free text prompts by noising the guidance image to an intermediate diffusion step, and then denoising it conditioned on the input prompt. This simple approach leads to impressive results, yet exhibit a tradeoff between preserving the guidance layout and fulfilling the target text. Several *concurrent* methods have taken first steps in controlling different properties of the generated content [14,21,39,48,50]. DreamBooth [39] and Textual Inversion [14] personalize a pre-trained text-to-image diffusion model given a few user-provided images. Our method also leverages a pre-trained text-to-image diffusion model to achieve our goal, yet does not involve any training or fine-tuning. Instead, we devise a simple framework that intervenes in the generation process by directly manipulating the spatial features.

Our approach of operating in the diffusion feature space is related to Prompt-to-Prompt (P2P) [17], which recently observed that by manipulating the cross-attention layers, it is possible to control the relation between the spatial layout of the image to each word in the text. Intuitively, since the cross attention is formed by the association of spatial features to *words*, it allows to capture rough regions at the *object level*, yet localized spatial information that is not expressed in the source text prompt (e.g., ob-

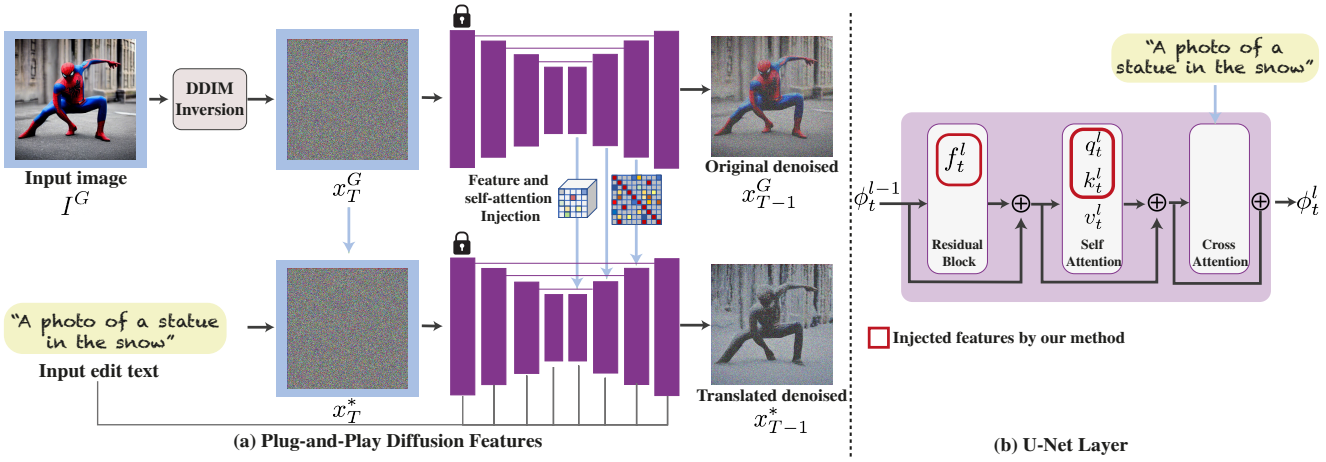


Figure 2. *Plug-and-play Diffusion Features.* (a) Our framework takes as input a guidance image and a text prompt describing the desired translation; the guidance image is inverted to initial noise $\boldsymbol{x}_T^G$, which is then progressively denoised using DDIM sampling. During this process, we extract $(\boldsymbol{f}_t^l, \boldsymbol{q}_t^l, \boldsymbol{k}_t^l)$ – spatial features from the decoder layers and their self-attention, as illustrated in (b). To generate our text-guided translated image, we fix $\boldsymbol{x}_T^* = \boldsymbol{x}_T^G$ and inject the guidance features $(\boldsymbol{f}_t^l, \boldsymbol{q}_t^l, \boldsymbol{k}_t^l)$ at certain layers, as discussed in Sec. 4.

ject parts) is not guaranteed to be preserved by P2P. Instead, our method focuses only on *spatial features* and their self-affinities – we show that such features exhibit high granularity of spatial information, allowing us to control the generated structure, while not restricting the interaction with the text. Thus, our method offers several key advantages: (i) enables fine-grained control over the generated shape and layout, (ii) allows to use arbitrary text-prompts to express the target translation; in contrast to P2P that requires word-to-word alignment between a source and target text prompts, (iii) demonstrates superior performance of real-world guidance images.

## 3. Preliminary

Diffusion models [12, 19, 37, 44] are probabilistic generative models in which an image is generated by progressively removing noise from an initial Gaussian noise image, $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$. These models are founded on two complementary random processes. the *forward* process, in which Gaussian noise is progressively added to a clean image, $\boldsymbol{x}_0$:

$$\boldsymbol{x}_t = \sqrt{\alpha_t} \cdot \boldsymbol{x}_0 + \sqrt{1-\alpha_t} \cdot \boldsymbol{z} \tag{1}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$ and $\{\alpha_t\}$ are the noise schedule.

The *backward* process is aimed at gradually denoising $\boldsymbol{x}_T$, where at each step a cleaner image is obtained. This process is achieved by a neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ that predicts the added noise $\boldsymbol{z}$. Once trained, each step of the backward process consists of applying $\boldsymbol{\epsilon}_\theta$ to the current $\boldsymbol{x}_t$, and adding a Gaussian noise perturbation to obtain a cleaner $\boldsymbol{x}_{t-1}$.

Diffusion models are rapidly evolving and have been extended and trained to progressively generate images *conditioned* on a guiding signal $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{y}, t)$, e.g., conditioning the generation on another image [40], class label [20], or text [22, 29, 35, 37].

In this work, we leverage a pre-trained text-conditioned Latent Diffusion Model (LDM), a.k.a Stable Diffusion [37], in which the diffusion process is applied in the latent space of a pre-trained image autoencoder. The model is based on a U-Net architecture [38] conditioned on the guiding prompt $P$. Layers of the U-Net comprise a residual block, a self-attention block, and a cross-attention block, as illustrated in Fig. 2 (b). The residual block convolve image features $\boldsymbol{\phi}_t^{l-1}$ from the previous layer $l$–1 to produce intermediate features $\boldsymbol{f}_t^l$. In the self-attention block, features are projected into queries, $\boldsymbol{q}_t^l$, keys, $\boldsymbol{k}_t^l$, and values, $\boldsymbol{v}_t^l$, and the output of the block is given by:

$$\hat{\boldsymbol{f}}_t^l = \boldsymbol{A}_t^l \boldsymbol{v}_t^l \quad \text{where} \quad \boldsymbol{A}_t^l = \text{Softmax}\left(\boldsymbol{q}_t^l \boldsymbol{k}_t^{l\,T}\right) \tag{2}$$

This operation allows for long-range interactions between image features. Finally, cross-attention is computed between the spatial image features and the token embedding of the text prompt $P$.

## 4. Method

Given an input guidance image $I^G$ and a target prompt $P$, our goal is to generate a new image $I^*$ that complies with $P$ and preserves the structure and semantic layout of $I^G$. We consider StableDiffusion [37], a state-of-the-art pre-trained and fixed text-to-image LDM model, denoted by $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, P, t)$. This model is based on a U-Net architecture, as illustrated in Fig. 2 and discussed in Sec. 3.

Our key finding is that fine-grained control over the generated structure can be achieved by manipulating spatial features inside the model during the generation process. Specifically, we observe and empirically demonstrate that: (i) spatial features extracted from intermediate decoder layers encode localized semantic information and are less affected by appearance information, and (ii) the self-attention,

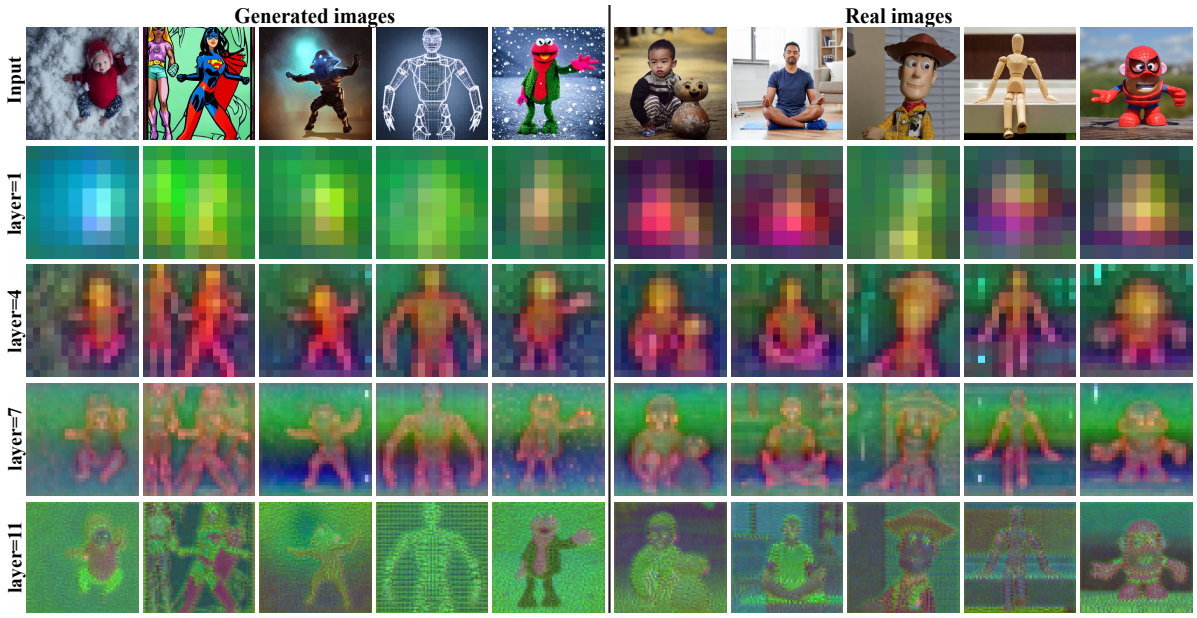


Figure 3. *Visualising diffusion features.* We used a set of 20 humanoid images (real and generated), and extracted spatial features from different decoder layers, at roughly 50% of the sampling process ($t = 540$). For each block, we applied PCA on the extracted features across *all* images and visualized the top three leading components. Intermediate features (layer 4) reveal semantic regions (e.g., legs or torso) that are shared across all images, under large variations in object appearance and image domain. Deeper features capture more high-frequency information, which eventually forms the noise predicted by the model. See SM for additional visualizations.

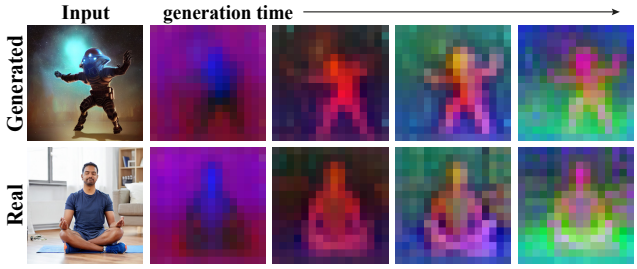


Figure 4. *Diffusion features over generation time-steps.* Visualizing PCA of spatial features of layer $l=4$ for the humanoid images (Fig. 3). Semantic parts are shared (have similar colors) across images at each time step.

representing the affinities between the spatial features, allows to retain fine layout and shape details.

Based on our findings, we devise a simple framework that extracts features from the generation process of the guidance image $I^G$ and directly injects them along with $P$ into the generation process of $I^*$, requiring no training or fine-tuning (Fig. 2). Our approach is applicable for both text-generated and real-world guidance images, for which we apply DDIM inversion [45] to get the initial $\boldsymbol{x}_T^G$.

**Spatial features.** In text-to-image generation, one can use descriptive text prompts to specify various scene and object proprieties, including those related to their shape, pose and scene layout, e.g., "*a photo of a horse galloping in the forest*". However, the exact scene layout, the shape of the object and its fine-grained pose often significantly vary across generated images from the same prompt under different initial noise $\boldsymbol{x}_T$. This suggests that the diffusion process itself and the resulting *spatial* features have a role in forming such fine-grained spatial information. This hypothesis is strengthened by [6], which demonstrated that semantic part segments can be estimated from spatial features in an unconditional diffusion model.

We opt to gain a better understanding of how such semantic spatial information is internally encoded in $\boldsymbol{\epsilon}_\theta$. To this end, we perform a simple PCA analysis which allows us to reason about the visual properties dominating the high-dimensional features in $\boldsymbol{\epsilon}_\theta$. Specifically, we generated a diverse set of images containing various humanoids in different styles, including both real and text-generated images; sample images are shown in Fig. 3. For each image, we extract features $\boldsymbol{f}_t^l$ from each layer of the decoder at each time step $t$, as illustrated in Fig. 2(b). We then apply PCA on $\boldsymbol{f}_t^l$ across all images.

Fig. 3 shows the first three principal components for a representative subset of the images across different layers and a single time step. As seen, the coarsest and shallowest layer is mostly dominated by foreground-background separation, depicting only a crude blob in the location of the foreground object. Interestingly, we can observe that the intermediate features (layer 4) encode localized semantic information shared across objects from different domains and under significant appearance variations – similar object parts (e.g., legs, torso, head) are depicted in similar colors *across* all images (*layer=4* row in Fig. 3). These proper-

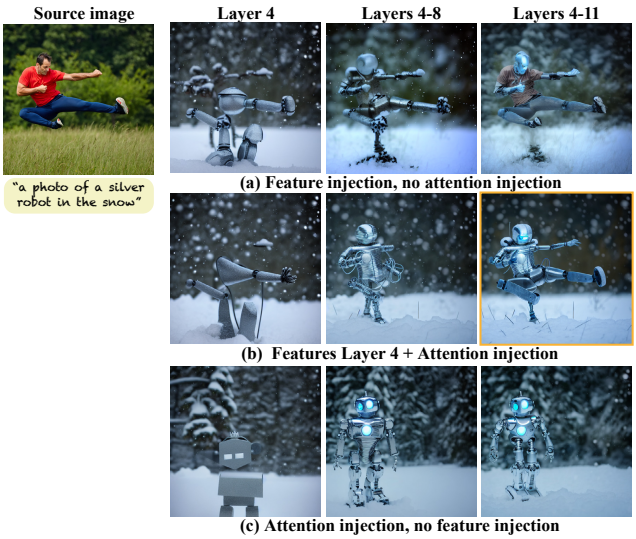


Figure 5. *Ablating features and attention injection.* (a) Features extracted from the guidance image (left) are injected into the generation process of the translated image (guided by a given text prompt). While features at intermediate layers (*Layer 4*) exhibit localized semantic information (Fig. 3), solely injecting these features is insufficient for retaining the guidance structure. Incorporating deeper (and higher resolution) features leads to better structure preservation, but results in appearance leakage from the guidance image to the generated one (*Layers 4-11*). (b) Injecting features only at layer 4 and self-attention maps at higher-resolution layers alleviates this issue. (c) Injecting only self-attention maps restricts the affinities between the features, yet there is no semantic association between the guidance features and the generated ones, resulting in misaligned structure. *The result of our final configuration is highlighted in orange.*

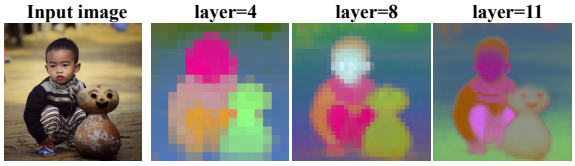


Figure 6. *Self-attention visualization.* Showing 3 leading components of the self-attention matrix $\boldsymbol{A}_t^l$ computed for the input image for three different layers. The principal components are aligned with the layout of the image: similar regions share similar colors

ties are consistent across the generation process as shown in Fig. 4. As we go deeper into the network, the features gradually capture more high-frequency low-level information which eventually forms the noise predicted by the network. Extended feature visualizations can be found in the Supplementary Materials (SM) on our website.

**Feature injection.** Based on these observations, we now turn to the translation task. Let $\boldsymbol{x}_T^G$ be the initial noise, obtained by inverting $I^G$ using DDIM [45].

Given the target prompt $P$, the generation of the translated image $I^*$ is carried with the same initial noise, i.e., $\boldsymbol{x}_T^* = \boldsymbol{x}_T^G$; we refer the reader to SM for an analysis and justification of this design choice.

At each step $t$ of the backward process, we extract the guidance features $\{\boldsymbol{f}_t^l\}$ from the denoising step: $\boldsymbol{z}_{t-1}^G = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^G, \varnothing, t)$.[1] These features are then injected into the generation of $I^*$, i.e., in the denoising step of $\boldsymbol{x}_t^*$, we override the resulting features $\{\boldsymbol{f}_t^{*l}\}$ with $\{\boldsymbol{f}_t^l\}$. This operation is expressed by:

$$\boldsymbol{z}_{t-1}^* = \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^*, P, t\ ;\ \{\boldsymbol{f}_t^l\}) \tag{3}$$

where we use $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\cdot\ ;\ \{\boldsymbol{f}_t^l\})$ to denote the modified denoising step with the injected features $\{\boldsymbol{f}_t^l\}$. In case of no injection, $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, P, t\ ;\ \varnothing) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, P, t)$.

Fig. 5(a) shows the effect of injecting spatial features $\boldsymbol{f}_t^l$ at increasing layers $l$. As seen, injecting features only at layer $l=4$ is insufficient for preserving the structure of the guidance image. As we inject features in deeper layers, the structure is better preserved, yet appearance information is leaked into the generated image (e.g., shades of the red t-shirt and blue jeans are apparent in *Layer 4-11*). To achieve a better balance between preserving the structure of $I^G$ and deviating from its appearance, we do not modify spatial features at deep layers, but rather leverage the self-attention layers as discussed below.

**Self-attention.** Self-attention modules compute the *affinities* $\boldsymbol{A}_t^l$ between the spatial features after linearly projecting them into queries and keys. These affinities have a tight connection to the established concept of self-similarly, which has been used to design structure descriptors by both classical and modern works [4,24,42,47]. This motivates us to consider the attention matrices $\boldsymbol{A}_t^l$ to achieve fine-grained control over the generated content.

Fig. 6, shows the leading principal components of a matrix $\boldsymbol{A}_t^l$ for a given image. As seen, in early layers, the attention is aligned with the semantic layout of the image, grouping regions according to semantic parts. Gradually, higher-frequency information is captured.

Practically, injecting the self-attention matrix is done by replacing the matrix $\boldsymbol{A}_t^l$ in Eq. 2. Intuitively, this operation pulls features close together, according to the affinities encoded in $\boldsymbol{A}_t^l$. We denote this additional operation by modifying Eq. (3) as follows:

$$\boldsymbol{z}_{t-1}^* = \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, P, t; \boldsymbol{f}_t^4, \{\boldsymbol{A}_t^l\}) \tag{4}$$

Fig. 5(b) shows the effect of Eq. (4) for increasing injection layers; the maximal injection layer of $\boldsymbol{A}_t^l$ controls the fidelity to the original structure, while mitigating the issue of appearance leakage. Fig. 5(c) demonstrates the pivotal role of the features $\boldsymbol{f}_t^4$. As seen, with only self-attention, i.e., $\boldsymbol{z}_{t-1}^* = \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, P, t; \{\boldsymbol{A}_t^l\})$, there is no semantic association between the original and translated contents, resulting in large structural deviations.

[1] In the case of a *generated* guidance image, $\boldsymbol{z}_{t-1}^G = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^G, P_G, t)$, where $P_G$ is the text used to generate $I^G$.

**Algorithm 1** Plug-and-Play Diffusion Features

**Inputs:**
$I^G$ ▷ real guidance image
$P$ ▷ target text prompt
$\tau_f, \tau_A$ ▷ injection thresholds

$\boldsymbol{x}_T^G \leftarrow \text{DDIM-inv}(I^G)$
$\boldsymbol{x}_T^* \leftarrow \boldsymbol{x}_T^G$ ▷ Starting from same seed
**for** $t \leftarrow T \dots 1$ **do**
  $\boldsymbol{z}_{t-1}^G, \boldsymbol{f}_t^4, \left\{\boldsymbol{A}_t^l\right\} \leftarrow \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t^G, \varnothing, t\right)$
  $\boldsymbol{x}_{t-1}^G \leftarrow \text{DDIM-samp}\left(\boldsymbol{x}_t^G, \boldsymbol{z}_{t-1}^G\right)$
  **if** $t > \tau_f$ **then** $\boldsymbol{f}_t^{*4} \leftarrow \boldsymbol{f}_t^4$ **else** $\boldsymbol{f}_t^{*4} \leftarrow \varnothing$
  **if** $t > \tau_A$ **then** $\boldsymbol{A}_t^{*l} \leftarrow \boldsymbol{A}_t^l$ **else** $\boldsymbol{A}_t^{*l} \leftarrow \varnothing$
  $\boldsymbol{z}_{t-1}^* \leftarrow \hat{\boldsymbol{\epsilon}}_\theta\left(\boldsymbol{x}_t^*, P, t\,;\ \boldsymbol{f}_t^{*4}, \left\{\boldsymbol{A}_t^{*l}\right\}\right)$
  $\boldsymbol{x}_{t-1}^* \leftarrow \text{DDIM-samp}\left(\boldsymbol{x}_t^*, \boldsymbol{z}_{t-1}^*\right)$
**end for**
**Output:** $I^* \leftarrow \boldsymbol{x}_0^*$

Our *plug-and-play diffusion features* framework is summarized in Alg. 1, and is controlled by two parameters: (i) $\tau_f$ is the sampling step until which $\boldsymbol{f}_t^4$ are injected. (ii) $\tau_A$ is the sampling step until which $\boldsymbol{A}_t^l$ are injected. In all our results, self-attention is injected in all decoder layers. The exact parameters settings are discussed in Sec. 5.

**Negative-prompting.** To increase the deviation from guidance image content, we use negative prompting [26] with a prompt $P_n$ that describes the guidance image. Additionally, we use a parameter $\alpha \in [0, 1]$ to balance between neutral and negative prompting. That is, at each sampling step, our final noise prediction becomes $\epsilon = w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, P, t) + (1-w)\tilde{\boldsymbol{\epsilon}}$, where $\tilde{\boldsymbol{\epsilon}}$ is given by

$$\tilde{\boldsymbol{\epsilon}} = \alpha\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \varnothing, t) + (1-\alpha)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, P_n, t) \tag{5}$$

where $w$ is the classifier-free guidance scale. We find negative-prompting to be beneficial for translating textureless "primitives" images (e.g., silhouettes). For natural-looking guidance images, it plays a minor role. See SM for more details.

## 5. Results

We thoroughly evaluate our method both quantitatively and qualitatively on diverse guidance image domains, both real and generated ones, as discussed below. Please see SM for full implementation details of our method.

**Datasets.** Our method supports versatile text-guided image-to-image translation tasks and can be applied to arbitrary image domains. Since there is no existing benchmark for such diverse settings, we created two new datasets: (i) *Wild-TI2I*, comprises of 148 diverse text-image pairs, 53% of which consists of real guidance images that we gathered from the Web; (ii) *ImageNet-R-TI2I*, a benchmark we derived from the ImageNet-R dataset [16], which comprises of various renditions (e.g., paintings, embroidery, etc.) of ImageNet object classes. To adopt this dataset for our purpose, we manually selected 3 high-quality images from 10 different classes. To generate our image-text examples, we created a list of text templates by defining for each source class target categories and styles, and automatically sampled their combinations. This results in total of 150 image-text pairs. See SM for full details.


Figure 7. Sample results of our method on image-text pairs from *Wild-TI2I* and *ImageNet-R-TI2I* benchmarks.

Figs. 1 and 7 show a sample of our results on both real and generated guidance images. Our results show both adherence to the guidance shape and compliance with different target prompts. Our method successfully handles both naturally looking as well as artistic and textureless guidance images.

### 5.1. Comparison to Prior/Concurrent Work

We focus our comparisons on state-of-the-art baselines that can be applied to diverse text-guided I2I tasks, including: (i) SDEdit [28] under three different noising levels, (ii) P2P [17], (iii) DiffuseIT [25], and (iv) VQGAN-CLIP [10]. We further provide qualitative comparisons to Text2LIVE [5], FlexIT [9] and DiffusionCLIP [22].

We note that P2P requires a source prompt that is word-aligned to the target prompt. Thus, we include a qualita-

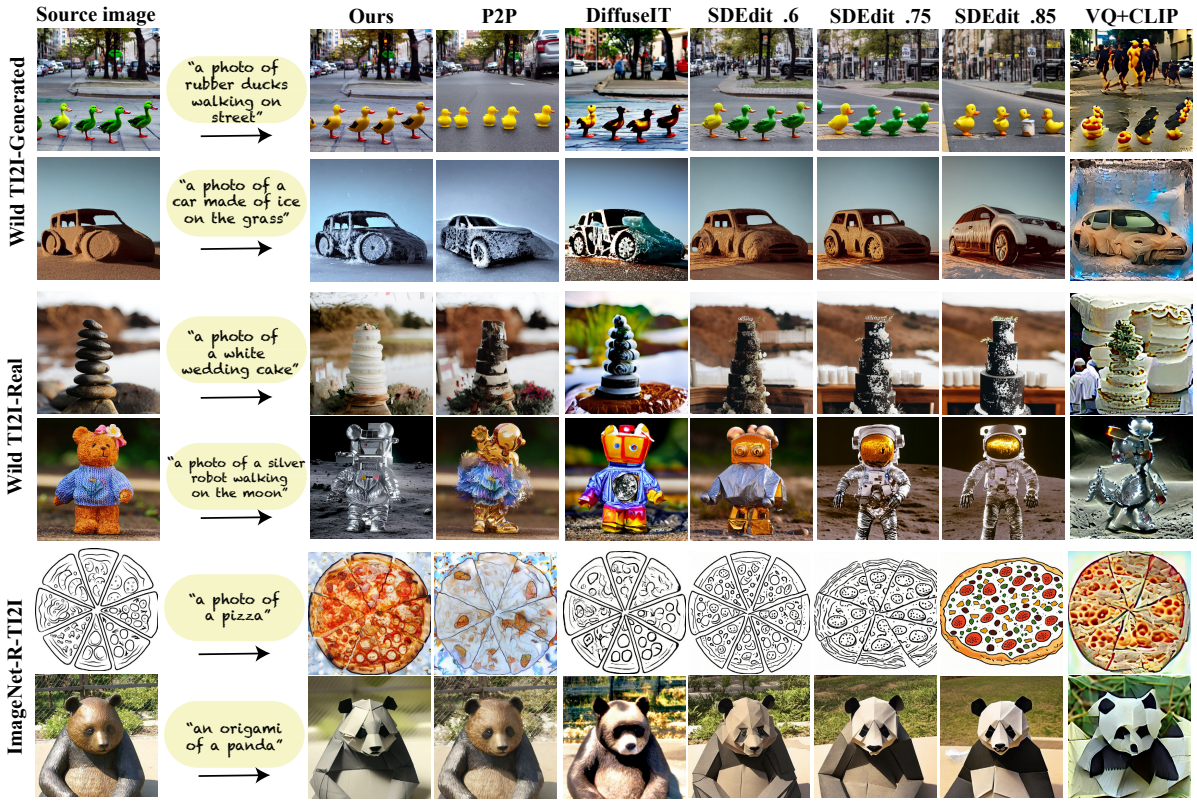


Figure 8. *Comparisons.* Sample results are shown for *ImageNet-R-TI2I* and *Wild-TI2I* benchmarks, including real and generated guidance images. Left to right: guidance image, text prompt, our result, P2P [17], DiffuseIT [25], SDedit [28] w/ 3 noising levels, and VQ+CLIP [10].

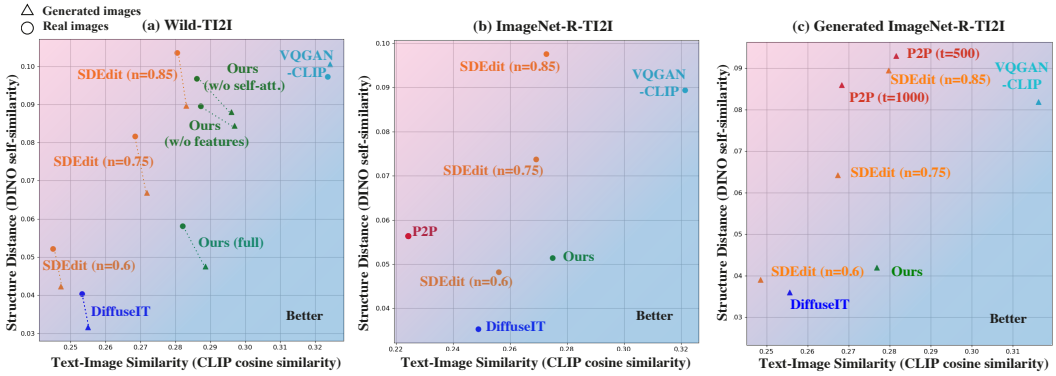


Figure 9. *Quantitative evaluation*. We measure CLIP cosine similarity (higher is better) and DINO-ViT self-similarity distance (lower is better) to quantify the prompt fidelity and structure preservation, respectively. The metrics are reported on 3 benchmarks: (a) *Wild-TI2I*, which includes ablations of our method, (b) *ImageNet-R-TI2I*, and (c) *Generated-ImageNet-R-TI2I*. Note that P2P can be applied only for (b) and (c) due to the prompts restriction. All baselines struggle to achieve both low structure distance and a high CLIP score. Our method exhibits a better balance between these two ends across all benchmarks.

tive and quantitative comparison to P2P on our *ImageNet-R-TI2I* benchmark, for which we created aligned source-target prompts. We further include qualitative comparison to a subset of *Wild-TI2I* for which the source and target prompts are aligned. For evaluating P2P on real guidance images, we applied DDIM inversion with the source text as in [17].

Fig. 8 shows sample results of our method compared with the baselines. As seen, our method successfully translates diverse inputs, both for real and generated guidance images. In all cases, our results exhibit both high preservation of the layout and high fidelity to the target prompt. In contrast, SDEdit suffers from an inherent tradeoff between the two – with low noise level, the structure is well preserved in the expanse of hardly changing the appearance; larger deviation in appearances is achieved with higher noise level, yet the structure is damaged. VQGAN+CLIP

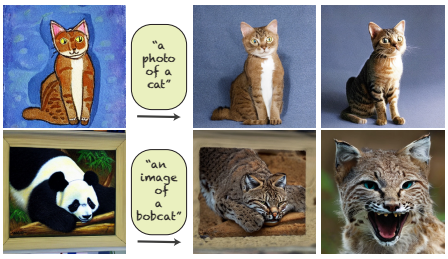


Figure 10. *Comparison to P2P on generated-ImageNet-R-TI2I benchmark.* P2P has noticeable deviation from the structure, especially in multiple word swaps (last row). See more examples in SM.

exhibits the same behavior, with overall lower image quality. Similarly, DiffuseIT shows high fidelity to the structure, with little changes to the appearance.

We find that P2P struggles to deviate from the guidance appearance to satisfy the edit when applied on real images (Fig. 8 rows 3-6). We speculate that since DDIM inversion is applied with a source text, it requires using low guidance scale at sampling, which limits the editability. To factor out the effect of DDIM inversion, we expand our comparison to P2P on *generated* guidance images. Specifically, we created a *generated-ImageNet-R-TI2I* benchmark by using prompts expressing the classes and renditions from [16] (see SM for further details on the benchmark creation). As seen in Fig. 10 and in Fig. 8 (first 2 rows), both our method and P2P comply with the target text. However, P2P often largely deviates from the structure, especially when applying multiple word swaps (second row in Fig. 10), while ours demonstrates fine-grained structure preservation across all examples (more examples in SM).

We numerically evaluate these results using two complementary metrics: text-image CLIP similarity to quantify the fidelity of the generated images to the text prompt (higher is better), and DINO-ViT self-similarity distance [47], to quantify structure preservation (lower is better). As seen in Fig. 9, our method outperforms the baselines by achieving both high structure preservation (in par with SDEdit w/ low noising level), and high prompt compliance (in par with SDEdit w/ high noising level). Note that VQGAN-CLIP and DiffuseIT directly use the evaluation metrics in their objective (CLIP loss in [10] and DINO self-similarity in [25]), which explains their respective scores in these metrics.

**Additional baselines.** Fig. 11 shows qualitative comparisons with: (i) Text2LIVE [5], (ii) DiffusionCLIP [22], and (iii) FlexIT [9]. These methods either fail to deviate from the guidance image or result in noticeable visual artifacts.

### 5.2. Ablation

We ablate our key design choices by evaluating our performance for the following cases: (i) w/o spatial features injection (w/o features), (ii) w/o self-attention injection. The metrics are reported in Fig. 9(a) and a representative example is shown in Fig. 5. The results demonstrate that both


Figure 11. Qualitative comparisons to additional baselines: Text2LIVE [5], DiffusionCLIP [22], FlexIT [9]. These methods fail to deviate from the structure for conveying the edit, or create undesired artifacts. More comparisons are included in the SM.

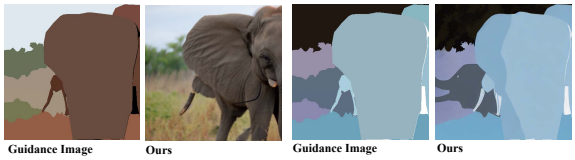


Figure 12. *Limitations.* Our method fails when there is no semantic association between the guidance content and the target text (e.g. arbitrarily colored segmentation masks)

features and self-attention are critical for structure preservation – the features provide a semantic association between the original and translated content, while self-attention is essential for maintaining this association and capturing finer structural information. See SM for further ablations.

## 6. Discussion and Conclusion

We presented a new framework for diverse text-guided image-to-image translation, founded on new insights about the internal representation of a pre-trained text-to-image diffusion model. Our method, based on simple manipulation of features, outperforms existing baselines, achieving a significantly better balance between preserving the guidance layout and deviating from its appearance. As for limitations, our method relies on the semantic association between the original and translated content in the diffusion feature space. Thus, it does not work well on arbitrarily colored segmentation masks (Fig. 12). In addition, we observed that for textureless "minimal" images, DDIM may occasionally result in a latent that encodes dominant low-frequency appearance information, which would result in appearance information leakage into our results. We believe that our work demonstrates the yet unrealized potential of the rich and powerful feature space spanned by pre-trained text-to-image diffusion models. We hope it will motivate future research in this direction.

**Acknowledgments:** We thank Omer Bar-Tal for his insights. This project received funding from the Israeli Science Foundation (grant 2303/20), the Carolito Stiftung, and the NVIDIA Applied Research Accelerator Program. Dr. Bagon is a Robin Chemers Neustein AI Fellow.

## References


[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 2

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[4] Shai Bagon, Ori Brostovski, Meirav Galun, and Michal Irani. Detecting and sketching the common. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 5

[5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*. Springer, 2022. 2, 6, 8

[6] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 4

[7] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. *arXiv preprint arXiv:2103.10951*, 2021. 2

[8] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: internet image montage. *ACM Trans. Graph.*, 2009. 2

[9] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. FlexIT: Towards flexible semantic image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 8

[10] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 6, 7, 8

[11] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021. 3

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

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

[15] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 2022. 2

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 6, 8

[17] 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, 6, 7

[18] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David Salesin. Image analogies. In Lynn Pocock, editor, *ACM Trans. on Graphics (Proceedings of ACM SIGGRAPH)*, 2001. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 3

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 3

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

[22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 6, 8

[23] Kunhee Kim, Sanghun Park, Eunyeong Jeon, Taehun Kim, and Daijin Kim. A style-aware discriminator for controllable image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[24] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5

[25] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2, 6, 7, 8

[26] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *arXiv preprint arXiv:2206.01714*, 2022. 6

[27] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021. 2

[28] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2, 6, 7

[29] 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. 1, 2, 3

[30] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 2

[31] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 2020. 2

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

[33] Lara Raad and Bruno Galerne. Efros and freeman image quilting algorithm for texture synthesis. *Image Process. Line*, 2017. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings the International Conference on Machine Learning (ICML)*, 2021. 2

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 3

[36] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1, 2, 3

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 3

[39] 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

[40] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2, 3

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[42] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 5

[43] Yichang Shih, Sylvain Paris, Frédo Durand, and William T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Trans. Graph.*, 2013. 2

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2015. 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 4, 5

[46] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[47] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 5, 8

[48] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image, 2022. 2

[49] Yael Vinker, Eliahu Horwitz, Nir Zabari, and Yedid Hoshen. Image shape manipulation from a single augmented training sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13769–13778, 2021. 2

[50] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. In *arXiv*, 2022. 2

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2