

An Edit Friendly DDPM Noise Space: Inversion and Manipulations

Inbar Huberman-Spiegelglas Vladimir Kulikov Tomer Michaeli
Technion – Israel Institute of Technology

inbarhub@gmail.com, {vladimir.k@campus, tomer.m@ee}.technion.ac.il

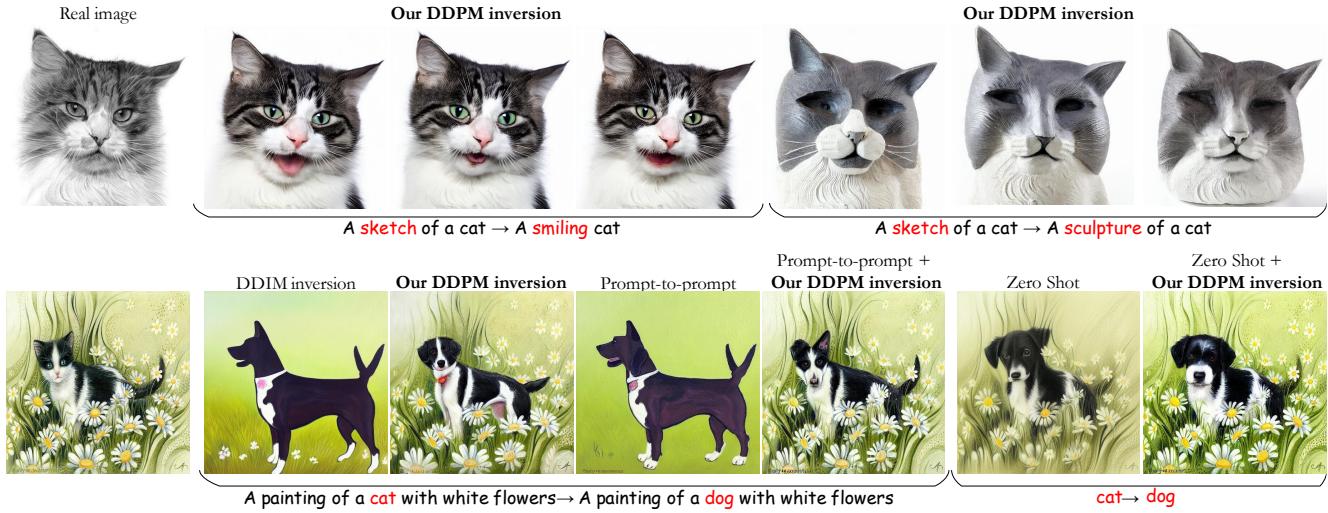


Figure 1: **Edit friendly DDPM inversion.** We present a method for extracting a sequence of DDPM noise maps that perfectly reconstruct a given image. These noise maps are distributed differently from those used in regular sampling, and are more edit-friendly. Our method allows diverse editing of real images without fine-tuning the model or modifying its attention maps, and it can also be easily integrated into other algorithms (illustrated here with Prompt-to-Prompt [7] and Zero-Shot I2I [19]).

Abstract

Denoising diffusion probabilistic models (DDPMs) employ a sequence of white Gaussian noise samples to generate an image. In analogy with GANs, those noise maps could be considered as the latent code associated with the generated image. However, this native noise space does not possess a convenient structure, and is thus challenging to work with in editing tasks. Here, we propose an alternative latent noise space for DDPM that enables a wide range of editing operations via simple means, and present an inversion method for extracting these edit-friendly noise maps for any given image (real or synthetically generated). As opposed to the native DDPM noise space, the edit-friendly noise maps do not have a standard normal distribution and are not statistically independent across timesteps. However, they allow perfect reconstruction of any desired image, and simple transformations on them translate into meaningful manipulations of the output image (e.g. shifting, color edits). Moreover, in text-

conditional models, fixing those noise maps while changing the text prompt, modifies semantics while retaining structure. We illustrate how this property enables text-based editing of real images via the diverse DDPM sampling scheme (in contrast to the popular non-diverse DDIM inversion). We also show how it can be used within existing diffusion-based editing methods to improve their quality and diversity. Code and examples are available on the project’s [webpage](#).

1. Introduction

Diffusion models have emerged as a powerful generative framework, achieving state-of-the-art quality on image synthesis [8, 2, 21, 18, 23, 20]. Recent works harness diffusion models for various image editing and manipulation tasks, including text-guided editing [7, 4, 25, 1, 12], inpainting [15], and image-to-image translation [22, 16, 28]. A key challenge in these methods is to leverage them for editing of real content (as opposed to model-generated images). This

requires inverting the generation process, namely extracting a sequence of noise vectors that would reconstruct the given image if used to drive the reverse diffusion process.

Despite significant advancements in diffusion-based editing, inversion is still considered a major challenge, particularly in the denoising diffusion probabilistic model (DDPM) sampling scheme [8]. Many recent methods (*e.g.* [7, 17, 25, 4, 27, 19]) rely on an approximate inversion method for the denoising diffusion implicit model (DDIM) scheme [24], which is a deterministic sampling process that maps a single initial noise vector into a generated image. However this DDIM inversion method becomes accurate only when using a large number of diffusion timesteps (*e.g.* 1000), and even in this regime it often leads to sub-optimal results in text-guided editing [7, 17]. To battle this effect, some methods fine-tune the diffusion model based on the given image and text prompt [11, 12, 26]. Other methods interfere in the generative process in various ways, *e.g.* by injecting the attention maps derived from the DDIM inversion process into the text-guided generative process [7, 19, 25].

Here we address the problem of inverting the DDPM scheme. As opposed to DDIM, in DDPM, $T + 1$ noise maps are involved in the generation process, each of which has the same dimension as the generated output. Therefore, the total dimension of the noise space is larger than that of the output and there exist infinitely many noise sequences that perfectly reconstruct the image. While this property may provide flexibility in the inversion process, not every consistent inversion (*i.e.* one that leads to perfect reconstruction) is also edit friendly. For example, one property we want from an inversion in the context of text-conditional models, is that fixing the noise maps and changing the text-prompt would lead to an artifact-free image, where the semantics correspond to the new text but the structure remains similar to that of the input image. What consistent inversions satisfy this property? A tempting answer is that the noise maps should be statistically independent and have a standard normal distribution, like in regular sampling. Such an approach was pursued in [28]. However, as we illustrate in Fig. 2, this native DDPM noise space is in fact not edit friendly.

Here we present an alternative inversion method, which constitutes a better fit for editing applications, from text-guidance manipulations, to editing via hand-drawn colored strokes. The key idea is to “imprint” the image more strongly onto the noise maps, so that they lead to better preservation of structure when fixing them and changing the condition of the model. Particularly, our noise maps have higher variances than the native ones, and are strongly correlated across timesteps. Importantly, our inversion requires no optimization and is extremely fast. Yet, it allows achieving state-of-the-art results on text-guided editing tasks with a relatively small number of diffusion steps, simply by fixing the noise maps and changing the text condition (*i.e.* without requir-

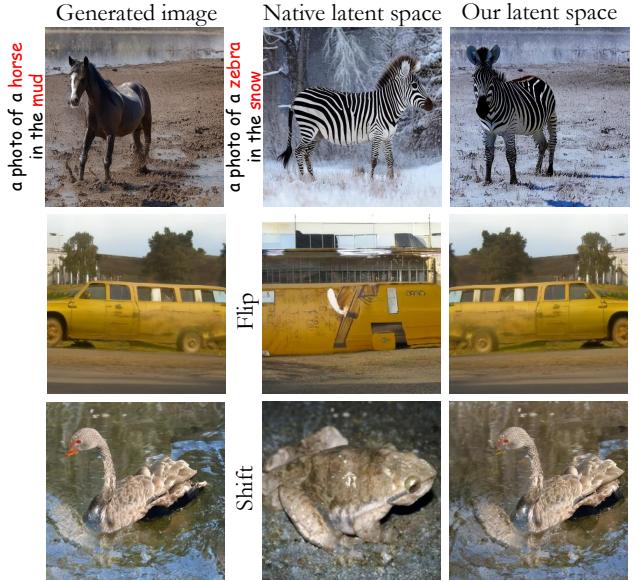


Figure 2: The native and edit friendly noise spaces. When sampling an image using DDPM (left), there is access to the “ground truth” noise maps that generated it. This native noise space, however, is not edit friendly (2nd column). For example, fixing those noise maps and changing the text prompt, changes the image structure (top). Similarly, flipping (middle) or shifting (bottom) the noise maps completely modifies the image. By contrast, our edit friendly noise maps enable editing while preserving structure (right).

ing model fine-tuning or interference in the attention maps). Our DDPM inversion can also be readily integrated with existing diffusion based editing methods that currently rely on approximate DDIM inversion. As we illustrate in Fig. 1, this improves their ability to preserve fidelity to the original image. Furthermore, since we find the noise vectors in a stochastic manner, we can provide a diverse set of edited images that all conform to the text prompt, a property not naturally available with DDIM inversion (see Figs. 1, 16, 17).

2. Related work

2.1. Inversion of diffusion models

Editing a real image using diffusion models requires extracting the noise vectors that would generate that image when used within the generative process. The vast majority of diffusion-based editing works use the DDIM scheme, which is a deterministic mapping from a single noise map to a generated image [7, 17, 25, 4, 27, 19, 6]. The original DDIM paper [24] suggested an efficient approximate inversion for that scheme. This method incurs a small error at every diffusion timestep, and these errors often accumulate into meaningful deviations when using classifier-free guid-

ance [10]. Mokady *et al.* [17] improve the reconstruction quality by fixing each timestep drifting. Their two-step approach first uses DDIM inversion to compute a sequence of noise vectors, and then uses this sequence to optimize the input null-text embedding at every timestep. EDICT [27] enables mathematically exact DDIM-inversion of real images by maintaining two coupled noise vectors which are used to invert each other in an alternating fashion. This method doubles the computation time of the diffusion process. Recently, Wu *et al.* [28] presented a DDPM-inversion method. Their approach recovers a sequence of noise vectors that perfectly reconstruct the image within the DDPM sampling process, however as opposed to our method, their extracted noise maps are distributed like the native noise space of DDPM, which results in limited editing capabilities (see Figs. 2,4).

2.2. Image editing using diffusion models

The DDPM sampling method is not popular for editing of real images. When used, it is typically done without exact inversion. Two examples are Ho *et al.* [8], who interpolate between real images, and Meng *et al.* [16] who edit real images via user sketches or strokes (SDEdit). Both construct a noisy version of the real image and apply a backward diffusion after editing. They suffer from an inherent tradeoff between the realism of the generated image and its faithfulness to the original contents. DiffuseIT [13] performs image translation guided by a reference image or by text, also without explicit inversion. They guide the generation process by losses that measure similarity to the original image [5].

A series of papers apply text-driven image-to-image translation using DDIM inversion. Narek *et al.* [25] achieve this by manipulating spatial features and their self-attention inside the model during the diffusion process. Hertz *et al.* [7] change the attention maps of the original image according to the target text prompt and inject them into the diffusion process. DiffEdit [4] automatically generates a mask for the regions of the image that need to be edited, based on source and target text prompts. This is used to enforce the faithfulness of the unedited regions to the original image, in order to battle the poor reconstruction quality obtained from the inversion. This method fails to predict accurate masks for complex prompts.

Some methods do not use DDIM inversion but utilize model optimization based on the target text prompt. DiffusionCLIP [12] uses model fine-tuning based on a CLIP loss with a target text. Imagic [11] first optimizes the target text embedding, and then optimizes the model to reconstruct the image with the optimized text embedding. UniTune [26] also uses fine-tuning and shows great success in making global stylistic changes and complex local edits while maintaining image structure. Other recent works like Palette [22] and InstructPix2Pix [3], learn conditional diffusion models tailored for specific image-to-image translation tasks.

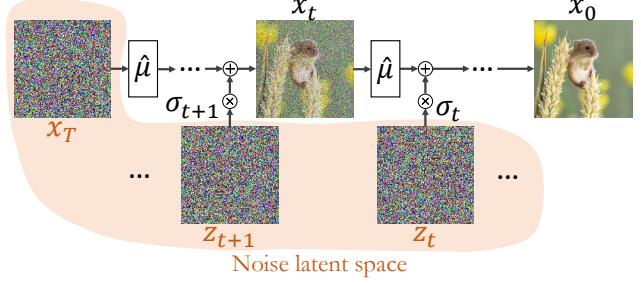


Figure 3: **The DDPM latent noise space.** In DDPM, the generative (reverse) diffusion process synthesizes an image x_0 in T steps, by utilizing $T + 1$ noise maps, $\{x_T, z_T, \dots, z_1\}$. We regard those noise maps as the latent code associated with the generated image.

3. The DDPM noise space

Here we focus on the DDPM sampling scheme, which is applicable in both pixel space [8] and latent space [21]. DDPM draws samples by attempting to reverse a diffusion process that gradually turns a clean image x_0 into white Gaussian noise,

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} n_t, \quad t = 1, \dots, T, \quad (1)$$

where $\{n_t\}$ are iid standard normal vectors and $\{\beta_t\}$ is some variance schedule. This diffusion process can be equivalently expressed as

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \quad (2)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, and $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. It is important to note that in (2), the vectors $\{\epsilon_t\}$ are *not independent*. In fact, ϵ_t and ϵ_{t-1} are highly correlated for all t . This is irrelevant for the training process, which is not affected by the joint distribution of ϵ_t 's across different timesteps, but it will be important for our discussion below.

The generative (reverse) diffusion process starts from a random noise vector $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it using the recursion

$$x_{t-1} = \hat{\mu}_t(x_t) + \sigma_t z_t, \quad t = T, \dots, 1, \quad (3)$$

where $\{z_t\}$ are iid standard normal vectors, and

$$\hat{\mu}_t(x_t) = \sqrt{\bar{\alpha}_{t-1}} P(f_t(x_t)) + D(f_t(x_t)). \quad (4)$$

Here, $f_t(x_t)$ is a neural network that is trained to predict ϵ_t from x_t , $P(f_t(x_t)) = (x_t - \sqrt{1 - \bar{\alpha}_t} f_t(x_t)) / \sqrt{\bar{\alpha}_t}$ is the predicted x_0 , and $D(f_t(x_t)) = \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} f_t(x_t)$ is a direction pointing to x_t . The variance schedule is taken to be $\sigma_t = \eta \beta_t (1 - \bar{\alpha}_{t-1}) / (1 - \bar{\alpha}_t)$, where $\eta \in [0, 1]$. The case $\eta = 1$ corresponds to the original DDPM work, and $\eta = 0$ corresponds to the deterministic DDIM scheme.

This generative process can be conditioned on text [21] or class [10] by using a neural network f that has been trained conditioned on those inputs. Alternatively, conditioning can be achieved through guided diffusion [5, 1], which requires utilising a pre-trained classifier or CLIP model during the generative process.

The vectors $\{x_T, z_T, \dots, z_1\}$ uniquely determine the image x_0 generated by the process (3) (but not vice versa). We therefore regard them as the latent code of the model (see Fig. 3). Here, we are interested in the inverse direction. Namely, given a real image x_0 , we would like to extract noise vectors that, if used in (3), would generate x_0 . We refer to such noise vectors as consistent with x_0 . Our method, explained next, works with any $\eta \in (0, 1]$.

4. Edit friendly inversion

It is instructive to note that *any* sequence of $T + 1$ images x_0, \dots, x_T , that starts with the real image x_0 , can be used to extract consistent noise maps by isolating z_t from (3) as¹

$$z_t = \frac{x_{t-1} - \hat{\mu}_t(x_t)}{\sigma_t}, \quad t = T, \dots, 1. \quad (5)$$

However, unless such an auxiliary sequence of images is carefully constructed, they are likely to be very far from the distribution of inputs on which the network $f_t(\cdot)$ was trained. In that case, fixing the so-extracted noise maps and changing the text condition, may lead to poor results.

What is a good way of constructing auxiliary images x_1, \dots, x_T for (5) then? A naive approach is to draw them from a distribution that is similar to that underlying the generative process. Such an approach was pursued by Wu *et al.* [28]. Specifically, they start by sampling $x_T \sim \mathcal{N}(0, \mathbf{I})$. Then, for each $t = T, \dots, 1$ they isolate ϵ_t from (2) using x_t and the real x_0 , substitute this ϵ_t for $f_t(x_t)$ in (4) to compute $\hat{\mu}_t(x_t)$, and use this $\hat{\mu}_t(x_t)$ in (3) to obtain x_{t-1} .

The noise maps extracted by this method are distributed similarly to those of the generative process. However, unfortunately, they are not well suited for editing of global structures. This is illustrated in Fig. 4 in the context of text guidance and in Fig. 7 in the context of shifts. The reason for this is that DDPM’s native noise space is not edit-friendly in the first place. Namely, even if we take a model-generated image, for which we have the “ground-truth” noise maps, fixing them while changing the text prompt does not preserve the structure of the image (see Fig. 2).

Here we propose to construct the auxiliary sequence x_1, \dots, x_T such that structures within the image x_0 would be more strongly “imprinted” into the noise maps extracted from (5). Specifically, we propose to construct them as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \tilde{\epsilon}_t, \quad 1, \dots, T, \quad (6)$$

¹Commonly, $z_1 = 0$ in DDPM, so that we run only over $t = T, \dots, 2$.

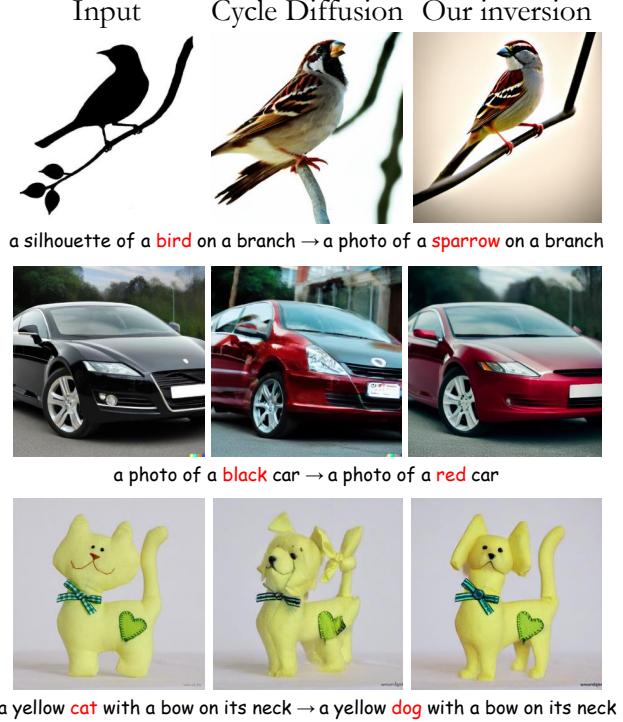


Figure 4: **DDPM inversion via CycleDiffusion vs. our method.** CycleDiffusion’s inversion [28] extracts a sequence of noise maps $\{x_T, z_T, \dots, z_1\}$ whose joint distribution is close to that used in regular sampling. However, fixing this latent code and replacing the text prompt fails to preserve the image structure. Our inversion deviates from the regular sampling distribution, but better encodes the image structure.

where $\tilde{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ are *statistically independent*. Note that despite the superficial resemblance between (6) and (2), these equations describe fundamentally different stochastic processes. In (2) every pair of consecutive ϵ_t ’s are highly correlated, while in (6) the $\tilde{\epsilon}_t$ ’s are independent. This implies that in our construction, x_t and x_{t-1} are typically going to be farther away from one another than in (2), so that every z_t extracted from (5) is expected to have a higher variance than in the regular generative process. A pseudo-code of our method is summarized in Alg. 1.

A few comments are in place regarding this inversion method. First, it reconstructs the input image up to machine precision, given that we compensate for accumulation of numerical errors (last row in Alg. 1). Second, it is straightforward to use with any kind of diffusion process (*e.g.* a conditional model [9, 8, 21], guided-diffusion [5], classifier-free [10]) by using the appropriate form for $\hat{\mu}_t(\cdot)$. Lastly, due to the randomness in (6), we can obtain many different inversions. While each of them leads to perfect reconstruction, when used for editing they will lead to different variants of the edited image. This allows generating diversity in *e.g.*

Algorithm 1 Edit-friendly DDPM inversion

```

Input: real image  $x_0$ 
Output:  $\{x_T, z_T, \dots, z_1\}$ 
for  $t = 1$  to  $T$  do
     $\tilde{\epsilon} \sim \mathcal{N}(0, 1)$ 
     $x_t \leftarrow \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ 
end for
for  $t = T$  to 1 do
     $z_t \leftarrow (x_{t-1} - \hat{\mu}_t(x_t)) / \sigma_t$ 
     $x_{t-1} \leftarrow \hat{\mu}_t(x_t) + \sigma_t z_t$  // to avoid error accumulation
end for
Return:  $\{x_T, z_T, \dots, z_1\}$ 

```

text-based editing tasks, a feature not naturally available with DDIM inversion methods (see Fig. 1).

5. Properties of the edit-friendly noise space

We now explore the properties of our edit-friendly noise space and compare it to the native DDPM noise space. We start with a 2D illustration, depicted in Fig. 5. Here, we use a diffusion model designed to sample from $\mathcal{N}(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \mathbf{I})$. The top-left pane shows a regular DDPM process with 40 inference steps. It starts from $x_T \sim \mathcal{N}(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{I})$ (black dot at the bottom left), and generates a sequence $\{x_t\}$ (green dots) that ends in x_0 (black dot at the top right). Each step is broken down to the deterministic drift $\hat{\mu}_t(x_t)$ (blue arrow) and the noise vector z_t (red arrow). On the top-right pane, we show a similar visualization, but for our latent space. Specifically, here we compute the sequences $\{x_t\}$ and $\{z_t\}$ using Alg. 1 for some given $x_0 \sim \mathcal{N}(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \mathbf{I})$. As can be seen, in our case, the noise perturbations $\{z_t\}$ are larger. Furthermore, close inspection reveals that the angles between consecutive noise vectors tend to be obtuse. In other words, our noise vectors are (negatively) correlated across consecutive times. This can also be seen from the two plots in the bottom row, which depict the histograms of angles between consecutive noise vectors for the regular sampling process and for ours. In the former case, the angle distribution is uniform, and in the latter it has a peak at 180° .

The same qualitative behavior occurs in diffusion models for image generation. Figure 6 shows the per-pixel variance of z_t and correlation between z_t and z_{t-1} for sampling with 100 steps from an unconditional diffusion model trained on Imagenet. Here the statistics were calculated over 10 images drawn from the model. As in the 2D case, our noise vectors have higher variances and they exhibit negative correlations between consecutive steps. As we illustrate next, these properties make our noise vectors more edit friendly.

Image shifting Intuitively, shifting an image should be possible by shifting all $T + 1$ maps of the latent code. Figure 7 shows the result of shifting the latent code of a model-

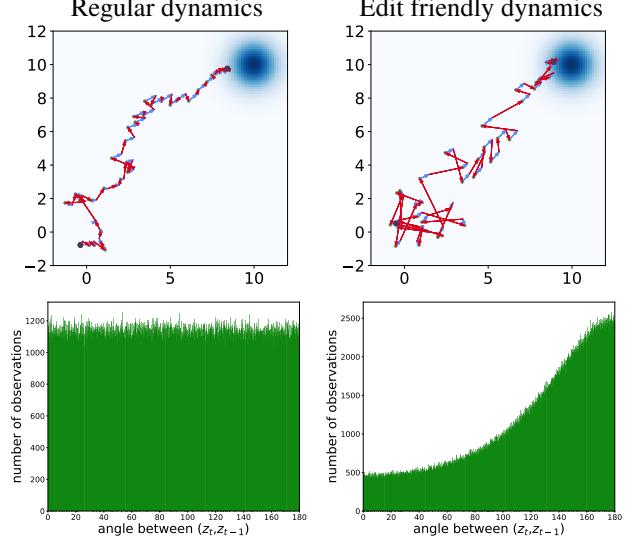


Figure 5: Regular vs. edit-friendly diffusion. In the regular generative process (top left), the noise vectors (red) are statistically independent across timesteps and thus the angle between consecutive vectors is uniformly distributed in $[0, 180^\circ]$ (bottom left). In our dynamics (top right) the noise vectors have higher variances and are negatively correlated across consecutive times (bottom right).

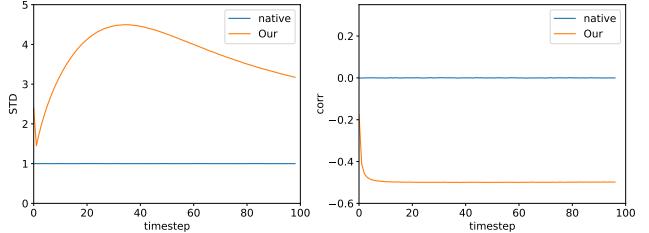


Figure 6: Native vs. edit friendly noise statistics. Here we show the per-pixel standard deviations of $\{z_t\}$ and the per-pixel correlation between them for model-generated images.

generated image by various amounts. As can be seen, shifting the native latent code (the one used to generate the image) leads to a complete loss of image structure. In contrast, shifting our edit-friendly code, results in minor degradation. Quantitative evaluation is provided in App. A.

Color manipulations Our latent space also enables convenient manipulation of colors. Specifically, suppose we are given an input image x_0 , a binary mask B , and a corresponding colored mask M . We start by constructing $\{x_1, \dots, x_T\}$ and extracting $\{z_1, \dots, z_T\}$ using (6) and (5), as before. Then, we modify the noise maps as

$$z_t^{\text{edited}} = z_t + sB \odot (M - P(f_t(x_t))), \quad (7)$$

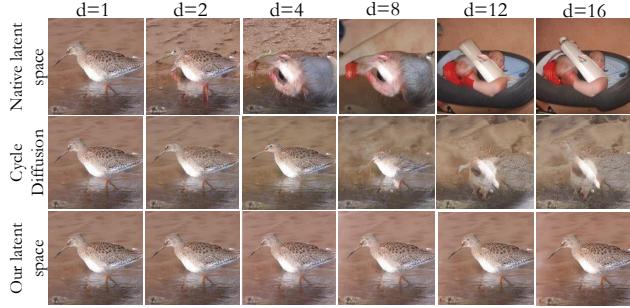


Figure 7: Image shifting. We shift to the right a 256×256 image generated using 100 inference steps from an unconditional model trained on ImageNet, by $d = 1, 2, 4, 8, 12, 16$ pixels. When using the native noise maps (top) or the ones extracted by CycleDiffusion [28] (middle) the structure is lost. With our latent space, the structure is preserved.

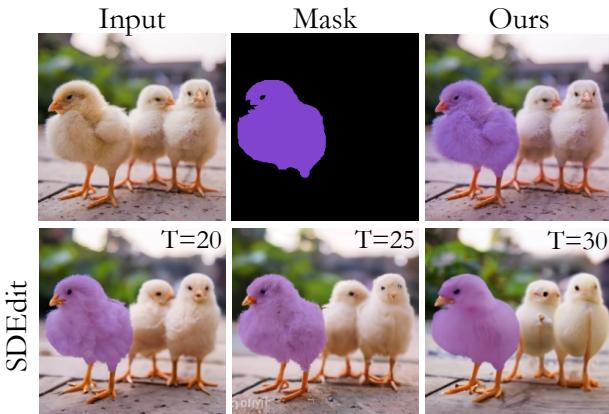


Figure 8: Color manipulation on a real image. Our method (applied here from $T_2 = 70$ to $T_1 = 20$ with $s = 0.05$) leads to a strong editing effect without modifying textures and structures. SDEdit, on the other hand, either does not integrate the mask well when using a small noise (left) or does not preserve structures when the noise is large (right). In both methods, we use an unconditional model trained on ImageNet with 100 inference steps.

with $P(f_t(x_t))$ from (4), where s is a parameter controlling the editing strength. We perform this modification over a range of timesteps $[T_1, T_2]$. Note that the term in parenthesis encodes the difference between the desired colors and the predicted clean image in each timestep. Figure 8 illustrates the effect of this process in comparison to SDEdit, which suffers from an inherent tradeoff between fidelity to the input image and conformation to the desired edit. Our approach can achieve a strong editing effect without modifying textures (neither inside nor outside the mask).

Text-Guided Image Editing Our latent space can be further utilized for text-driven image editing. Suppose we are

given a real image x_0 , a text prompt describing it p_{src} , and a target text prompt p_{tar} . To modify the image according to these prompts, we extract the edit-friendly noise maps $\{x_T, z_T, \dots, z_1\}$, while injecting p_{src} to the denoiser. We then fix those noise maps and generate an image while injecting p_{tar} to the denoiser. We run the generation process starting from timestep $T - T_{\text{skip}}$, where T_{skip} is a parameter controlling the adherence to the input image. Figures 1, 2, 4, and 9 show several text driven editing examples using this approach. As can be seen, this method nicely modifies semantics while preserving the structure of the image. In addition, it allows generating diverse outputs for any given edit (Figs. 1, 16 and 17). Figures 1, 9 and 10 further illustrate the effect of using our inversion in combination with methods that currently rely on DDIM inversion. As can be seen, these methods often do not preserve fine textures, like fur, flowers, or leaves of a tree, and oftentimes also do not preserve the global structure of the objects. By integrating our inversion, structures and textures are better preserved.

6. Experiments

We evaluate our method both quantitatively and qualitatively on text-guided editing of real images. We analyze the usage of our extracted latent code by itself (as explained in Sec. 5), and in combination with existing methods that currently use DDIM inversion. In the latter case, we extract the noise maps using our DDPM-inversion and inject them in the reverse process, in addition to any manipulation they perform on *e.g.* attention maps. All experiments use a real input image as well as source and target text prompts.

Implementation details We use Stable Diffusion [21], in which the diffusion process is applied in the latent space of a pre-trained image autoencoder. The image size is $512 \times 512 \times 3$, and the latent space is $64 \times 64 \times 4$. Our method is also applicable in unconditional pixel space models using CLIP guidance, however we found Stable Diffusion to lead to better results. We use $\eta = 1$ and 100 inference steps, unless noted otherwise. Two hyper-parameters control the balance between faithfulness to the input image and adherence to the target prompt: the strength of the classifier-free guidance [10], and T_{skip} explained in Sec. 5. In all our numerical analyses and all results in Fig. 9, we fixed those parameters to strength = 15 and $T_{\text{skip}} = 36$. In App. C, we provide a thorough analysis of their effects.

Datasets We use two datasets of real images: (i) “modified ImageNet-R-TI2I” from [25] with additional examples collected from the internet and from other datasets, (ii) “modified Zero-Shot I2IT”, which contains images of 4 classes (Cat, Dog, Horse, Zebra) from [19] and from the internet. The first dataset comprises 48 images, with 3-5 different target prompts for each image. This results in a total of 212 image-text pairs. The second dataset has 15 images in

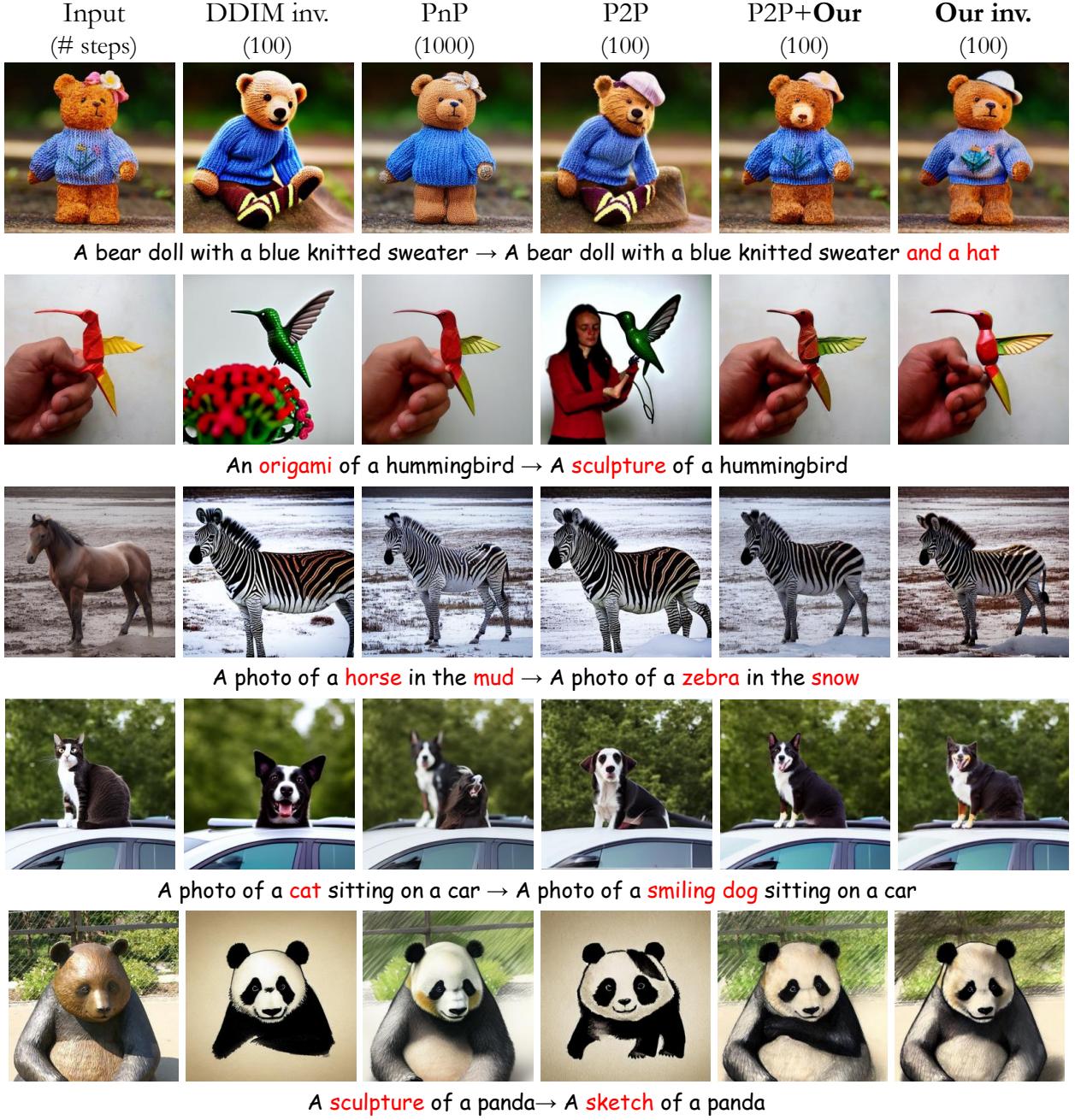


Figure 9: **Comparisons.** We show results for editing of real images using our inversion by itself and in combination with other methods. Our approach maintains high fidelity to the input while conforming to the text prompt.

each category with one target prompt for each, making 60 image-text pairs in total. Please see App. D for full details.

Metrics We numerically evaluate the results using two complementary metrics: LPIPS [29] to quantify the extent of structure preservation (lower is better) and a CLIP-based score to quantify how well the generated images comply with the text prompt (higher is better). We also measure

running time in seconds, and diversity among generated outputs (higher is better). Specifically, for each image and source text p_{src} , we generate 8 outputs with target text p_{tar} and calculate the average LPIPS distance over all $\binom{8}{2}$ pairs.

Comparisons on the modified ImageNet-R-TI2I dataset
We perform comparisons with Prompt-to-Prompt (P2P) [7], which uses DDIM inversion. We also evaluate the integra-

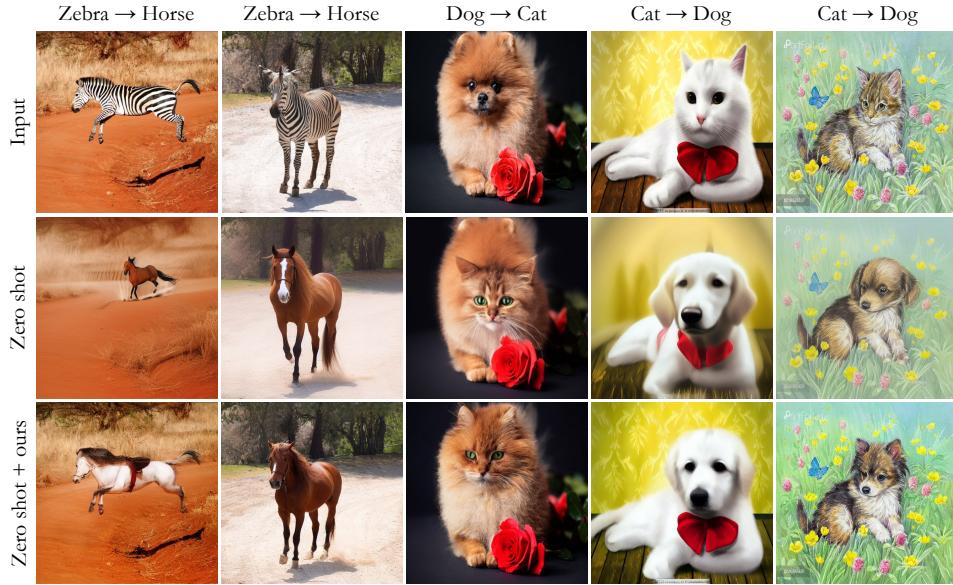


Figure 10: **Improving Zero-Shot I2I Translation.** Images generated by Zero Shot I2I suffer from loss of detail. With our inversion, fine textures, like fur and flowers, are retained. All results use the default cross-attention guidance weight (0.1).

tion of P2P with our inversion. In that case, we decrease the hyper-parameter controlling the cross-attention from 0.8 to 0.6 (as our latent space already strongly encodes structure). Furthermore, we compare to plain DDIM inversion and to Plug-and-Play (PnP) [25]. We note that P2P has different modes for different tasks (swap word, prompt refinement), and we chose its best mode for each image-prompt pair. As seen in Fig. 9, our method successfully modifies real images according to the target prompts. In all cases, our results exhibit both high fidelity to the input image and adherence to the target prompt. DDIM inversion and P2P do not preserve structure well. Yet, P2P does produce appealing results when used with our inversion. PnP often preserves structure, but requires 1000 steps and more than 3 minutes to edit an image. As seen in Tab. 1, our inversion achieves a good balance between LPIPS and CLIP, while requiring short edit times and supporting diversity among generated outputs. Integrating our inversion into P2P improves their performance in both metrics. See more analyses in App. D.1.

Comparisons on the modified Zero-Shot I2IT dataset
 Next, we compare our method to Zero-Shot Image-to-Image Translation (Zero-Shot) [19], which uses DDIM inversion. This method only translates between several predefined classes. We follow their setting and use 50 diffusion steps. When using our inversion, we decrease the hyper-parameter controlling the cross-attention from 0.1 to 0.03. As can be seen in Fig. 10, while Zero-Shot’s results comply with the target text, they are typically blurry and miss detail. Integrating our inversion adds back the details from the input image. As seen in Tab. 2, integrating our inversion improves

Method	CLIP sim. \uparrow	LPIPS \downarrow	Diversity \uparrow	Time
DDIM inv.	0.31	0.62	0.00	39
PnP	0.31	0.36	0.00	206
P2P	0.30	0.61	0.00	40
P2P+Our	0.31	0.25	0.11	48
Our inv.	0.32	0.29	0.18	36

Table 1: Evaluation on modified ImageNet-R-TI2I dataset.

Method	CLIP Acc. \uparrow	LPIPS \downarrow	Diversity \uparrow	Time
Zero-Shot	0.88	0.35	0.07	45
Zero-Shot+Our	0.88	0.27	0.16	46

Table 2: Evaluation on the modified Zero-Shot I2IT dataset.

the similarity to the input image while keeping the CLIP accuracy high and exhibiting non-negligible diversity among the generated outputs. See more details in App. D.2.

7. Conclusion

We presented an inversion method for DDPM. Our noise maps behave differently than the ones used in regular sampling: they are correlated across timesteps and have a higher variance. However, they encode the image structure more strongly, and are thus better suited for image editing. Particularly, we showed that they can be used to obtain state-of-the-art results on text-based editing tasks, either by themselves or in combination with other editing methods. And in contrast to methods that rely on DDIM inversion, our approach can generate diverse results for any given image and text.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. [1](#), [4](#)
- [2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. [1](#)
- [3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. [3](#)
- [4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. [1](#), [2](#), [3](#)
- [5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. [3](#), [4](#)
- [6] Adham Elarabawy, Harish Kamath, and Samuel Denton. Direct inversion: Optimization-free text-driven real image editing with diffusion models. *arXiv preprint arXiv:2211.07825*, 2022. [2](#)
- [7] 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. [1](#), [2](#), [3](#), [7](#), [15](#)
- [8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020. [1](#), [2](#), [3](#), [4](#)
- [9] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23:47:1–47:33, 2021. [4](#)
- [10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. [3](#), [4](#), [6](#), [14](#)
- [11] 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](#), [3](#)
- [12] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2416–2425, 2022. [1](#), [2](#), [3](#)
- [13] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. [3](#)
- [14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022. [15](#)
- [15] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022. [1](#)
- [16] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022. [1](#), [3](#)
- [17] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. [2](#), [3](#)
- [18] Alexander Quinn 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. In *Proceedings of the 39th International Conference on Machine Learning (ICLR)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. [1](#)
- [19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. [1](#), [2](#), [6](#), [8](#), [15](#)
- [20] 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](#)
- [21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. [1](#), [3](#), [4](#), [6](#)
- [22] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 2021. [1](#), [3](#)
- [23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. [1](#)
- [24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. [2](#)
- [25] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. [1](#), [2](#), [3](#), [6](#), [8](#), [15](#)
- [26] 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. *arXiv preprint arXiv:2210.09477*, 2022. [2](#), [3](#)

- [27] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact diffusion inversion via coupled transformations. *arXiv preprint arXiv:2211.12446*, 2022. [2](#), [3](#)
- [28] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to CycleDiffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. [1](#), [2](#), [3](#), [4](#), [6](#), [12](#)
- [29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. [7](#)

A. Shifting the latent code

As described in Sec. 3, we can shift an input image by shifting its extracted latent code. This requires inserting new columns/rows at the boundary of the noise maps. To guarantee that the inserted columns/rows are drawn from the same distribution as the rest of the noise map, we simply copy a contiguous chunk of columns/rows from a different part of the noise map. In all our experiments, we copied into the boundary the columns/rows indexed $\{50, \dots, 50 + d - 1\}$ for a shift of d pixels. We found this strategy to work better than randomly drawing the missing columns/rows from a white normal distribution having the same mean and variance as the rest of the noise map. Figure 11 depicts the MSE over the valid pixels that is incurred when shifting the noise maps. This analysis was done using 25 generated-images. As can be seen, shifting our edit-friendly code results in minor degradation while shifting the native latent code leads to a complete loss of the image structure.

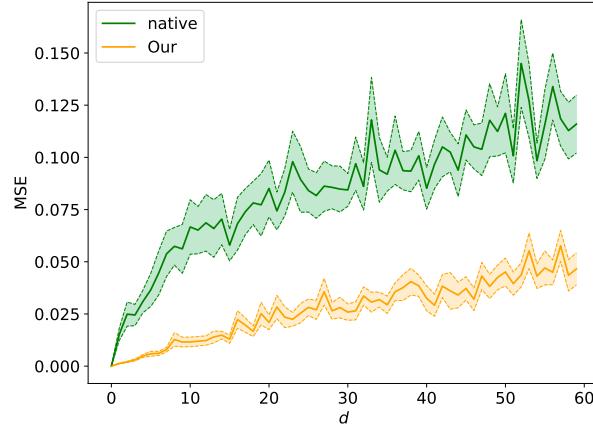


Figure 11: **Shifting the latent code.** We plot the MSE over the valid pixels after shifting the latent code and generating the image. The colored regions represent one standard error of the mean (SEM) in each direction.

B. CycleDiffusion

As mentioned in Sec. 5, CycleDiffusion [28] extracts a sequence of noise maps $\{x_T, z_T, \dots, z_1\}$ for the DDPM scheme. However, in contrast to our method, their noise maps have statistical properties that resemble those of regular sampling. This is illustrated in Fig. 12, which depicts the per-pixel standard deviations of $\{z_t\}$ and the correlation between z_t and z_{t-1} for CycleDiffusion, for regular sampling, and for our approach. These statistics were calculated over 10 images using an unconditional diffusion model trained on Imagenet. As can be seen, the CycleDiffusion curves are almost identical to those of regular sampling, and are different from ours.

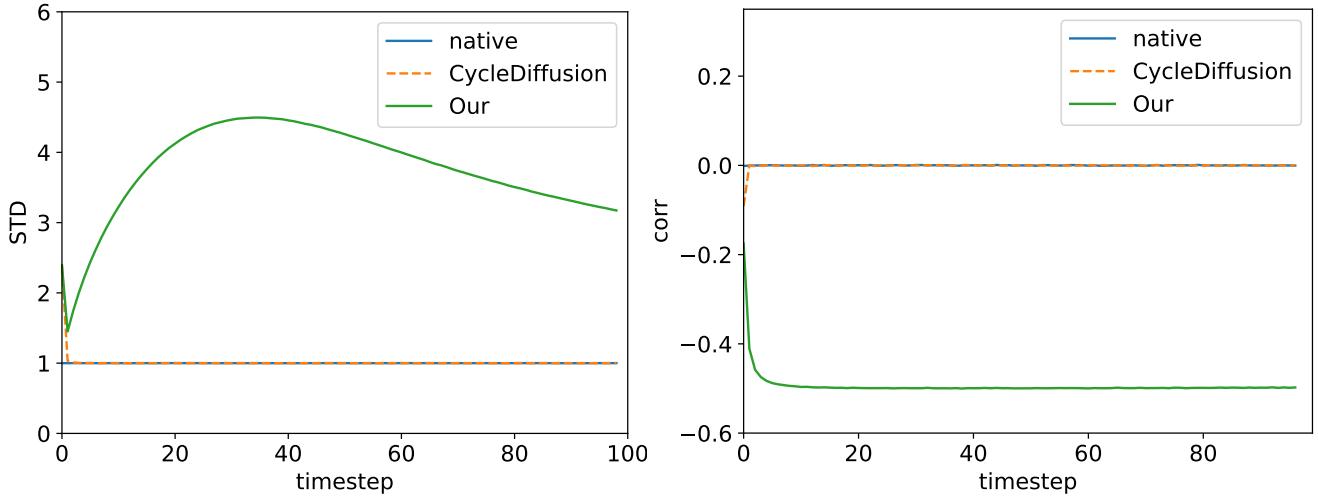


Figure 12: **CycleDiffusion noise statistics.** Here we show the per-pixel standard deviations of $\{z_t\}$ and the per-pixel correlation between them for mssodel-generated images.

The implication of this is that similarly to the native latent space, simple manipulations on CycleDiffusion’s noise maps cannot be used to obtain artifact-free effects in pixel space. This is illustrated in Fig. 13 in the context of horizontal flip and horizontal shift by 30 pixels to the right. As opposed to Cycle diffusion, applying those transformations on our latent code, leads to the desired effects, while better preserving structure.

This behavior also affects the text based editing capabilities of CycleDiffusion. Particularly, the CLIP similarity and LPIPS distance achieved by CycleDiffusion on the modified ImageNet-R-TI2I dataset are plotted in Fig. 15. As can be seen, when tuned to achieve a high CLIP-similarity (*i.e.* to better conform with the text), CycleDiffusion’s LPIPS loss increases significantly, indicating that the output images become less similar to the input images. For the same level of CLIP similarity, our approach achieves a substantially lower LPIPS distance.

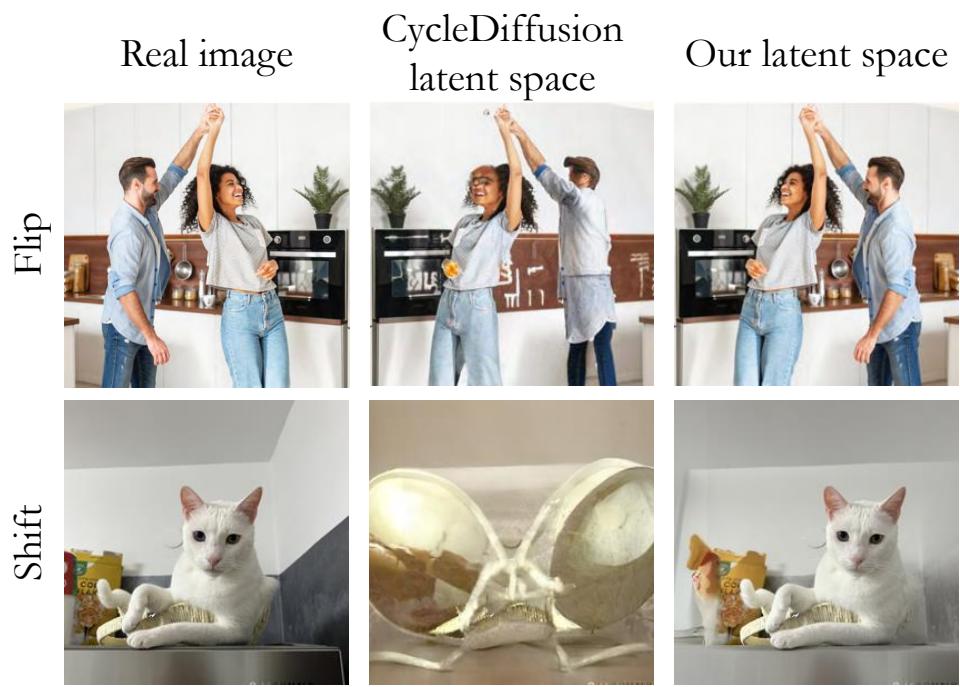


Figure 13: **Flip and shift with CycleDiffusion and with our inversion.**

C. The effects of the skip and the strength parameters

Recall from Sec. 3 that to perform text-guided image editing using our inversion, we start by extracting the latent noise maps while injecting the source text into the model, and then generate an image by fixing the noise maps and injecting a target text prompt. Two important parameters in this process are T_{skip} , which controls the timestep ($T - T_{\text{skip}}$) from which we start the generation process, and the strength parameter of the classifier-free scale [10]. Figure 14 shows the effects of these parameters. When T_{skip} is large, we start the process with a less noisy image and thus the output image remains close to the input image. On the other hand, the strength parameter controls the compliance of the output image with the target prompt.



Figure 14: The effects of the skip and the strength parameters.

D. Additional details on experiments and further numerical evaluation

For all our text-based editing experiments, we used Stable Diffusion as our pre-trained text-to-image model. We specifically used the StableDiffusion-v-1-4 checkpoint. We ran all experiments on an RTX A6000 GPU. We now provide additional details about the evaluations reported in the main text. All datasets and prompts will be published.

D.1. Experiments on the modified ImageNet-R-TI2I

Our modified ImageNet-R-TI2I dataset contains 44 images: 30 taken from PnP [25], and 14 from the Internet and from the code bases of other existing text-based editing methods. We verified that there is a reasonable source and target prompt for each image we added. For P2P [7] (with and without our inversion), we used the first 30 images with the “replace” option, since they were created with rendering and class changes. That is, the text prompts were of the form “a <>rendering<> of a <>class<>” (*e.g.* “a sketch of a cat” to “a sculpture of a cat”). The last 14 images include prompts with additional tokens and different prompt lengths (*e.g.* changing “A photo of an old church” to “A photo of an old church with a rainbow”). Therefore for those images we used the “refine” option in P2P. We configured all methods to use 100 forward and backward steps, except for PnP whose supplied code does not work when changing this parameter.

Table 3 summarizes the hyper-parameters we used for all methods. For our inversion and for P2P with our inversion, we arrived at those parameters by experimenting with various sets of the parameters and choosing the configuration that led to the best CLIP loss under the constraint that the LPIPS distance does not exceed 0.3. For DDIM inversion and for P2P (who did not illustrate their method on real images), such a requirement could not be satisfied. Therefore for those methods we chose the configuration that led to the best CLIP loss under the constraint that the LPIPS distance does not exceed 0.62. For PnP, we used the default parameters supplied by the authors. In Fig. 15 we show the CLIP-LPIPS losses graph for all methods reported in the paper, as well as for CycleDiffusion. In this graph, we show three different parameter configurations for our inversion, for P2P with our inversion, and for CycleDiffusion. As can be seen, our method (by itself or with P2P) achieves the best LPIPS distance for any given level of CLIP similarity.

Method	#inv. steps	#edit steps	strength	T_{skip}	τ_x/τ_a
PnP	1000	50	10	0	40/25
DDIM inv.	100	100	9	0	—
Our inv.	100	100	15	36	—
P2P	100	100	9	0	80/40
P2P + Our inv.	100	100	9	12	60/20

Table 3: **Hyper-parameters used in experiments on the modified ImageNet-R-TI2I dataset.** The parameter ‘strength’ refers to the classifier-free scale of the generation process. As for the strength used in the inversion stage, we set it to 3.5 for all methods except for PnP, which uses 1. The timestep at which we start the generation is $T - T_{\text{skip}}$ and, in case of injecting attentions, we also report the timestep at which the cross- and self-attentions start to be injected, τ_x and τ_a respectively.

D.2. Experiments on the modified zero-shot I2IT dataset

The second dataset we used is the modified Zero-Shot I2IT dataset, which contains 4 categories (cat, dog, horse, zebra). Ten images from each category were taken from Parmar *et al.* [19], and we added 5 more images from the Internet to each category. Zero-Shot I2I [19] does not use source-target pair prompts, but rather pre-defined source-target classes (*e.g.* cat↔dog). For their optimized DDIM-inversion part, they use a source prompt automatically generated with BLIP [14]. When combining our inversion with their generative method, we use $T_{\text{skip}} = 0$ and an empty source prompt. Table 4 summarizes the hyper-parameters used in every method.

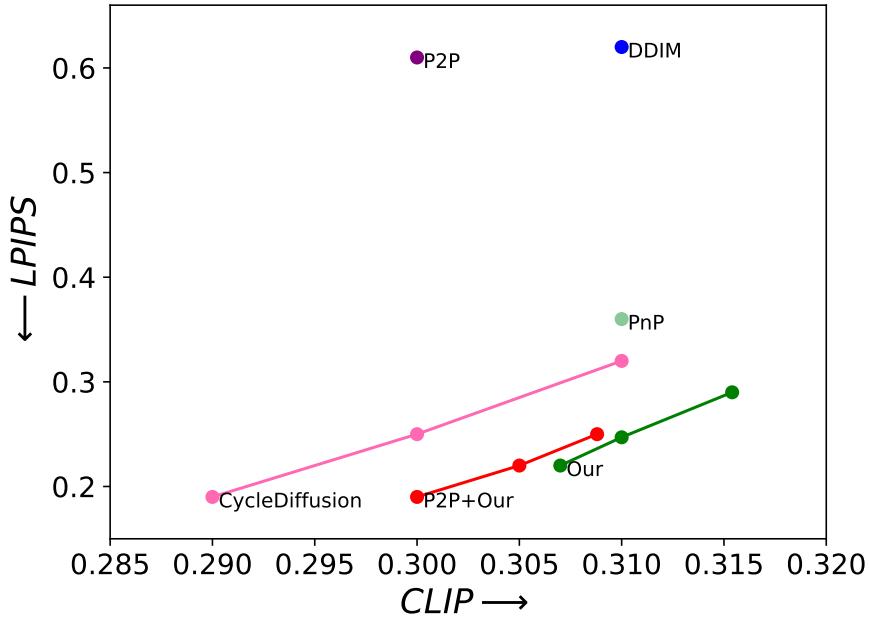


Figure 15: **Fidelity to source image vs. compliance with target text.** We show a comparison of all methods over the modified ImageNet-R-TI2I in terms of the LPIPS and CLIP losses. The parameters used for the evaluation are reported in Tab. 3. In addition, we depict three options for our inversion, for P2P with our inversion, and for CycleDiffusion. These three options correspond to different choices of the parameters (strength, T_{skip}): for our method (15, 36), (12, 36), (9, 36), for P2P+Ours (7.5, 8), (7.5, 12), (9, 20), and for CycleDiffusion (3, 30), (4, 25), (4, 15). For the CLIP loss, higher is better while for the LPIPS loss, lower is better.

Method	#inv. steps	#edit steps	strength	T_{skip}	λ_{xa}
Zero-Shot	50	50	7.5	0	0.1
Zero-Shot+Our	50	50	7.5	0	0.03

Table 4: **Hyper-parameters used in experiments on the modified Zero-Shot I2IT dataset.** In this method, cross-attention guidance weight is the parameter used to control the consistency in the cross-attention maps, denoted here as λ_{xa} . We set the strength (classifier-free scale) in the inversion part to be 1 and 3.5 for “Zero-shot” and “Zero-shot+Our” respectively.

E. Additional results

Due to the stochastic manner of our method, we can generate diverse outputs, a feature that is not naturally available with methods relying on the DDIM inversion. Figures 16 and 17 show several diverse text-based editing results. Figures 18 and 19 provide further qualitative comparisons between all methods tested on the ImageNet-R-TI2I dataset.

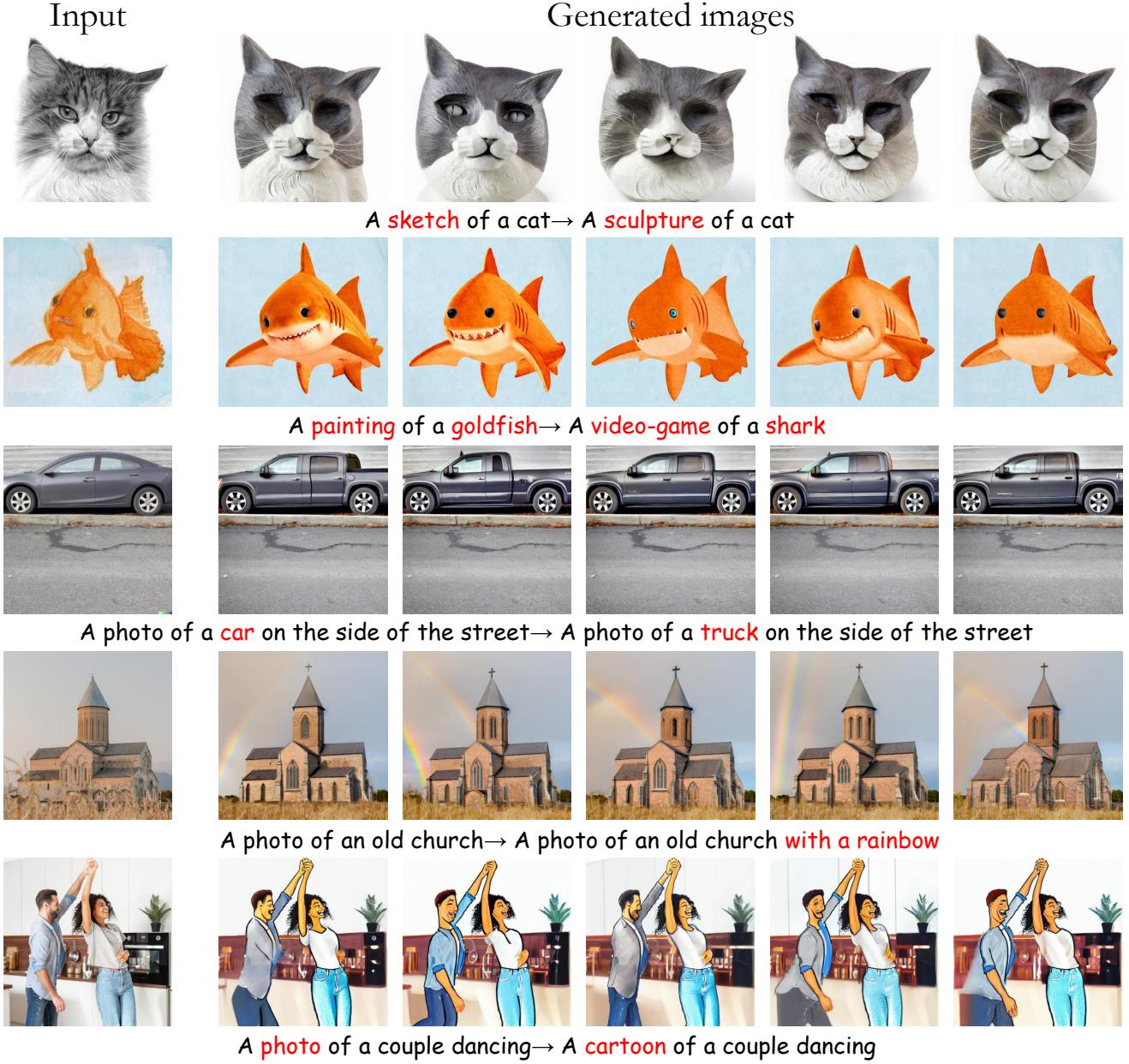


Figure 16: **Diverse text-based editing with our method.** We apply our inversion five times with the same source and target prompts (shown beneath each example). Note how the variability between the results is not negligible, while all of them conform to the structure of the input image and comply with the target text prompt. Notice e.g. the variability in the sculpture cat's eyes and mouth, and how the rainbow appears in different locations and angles.

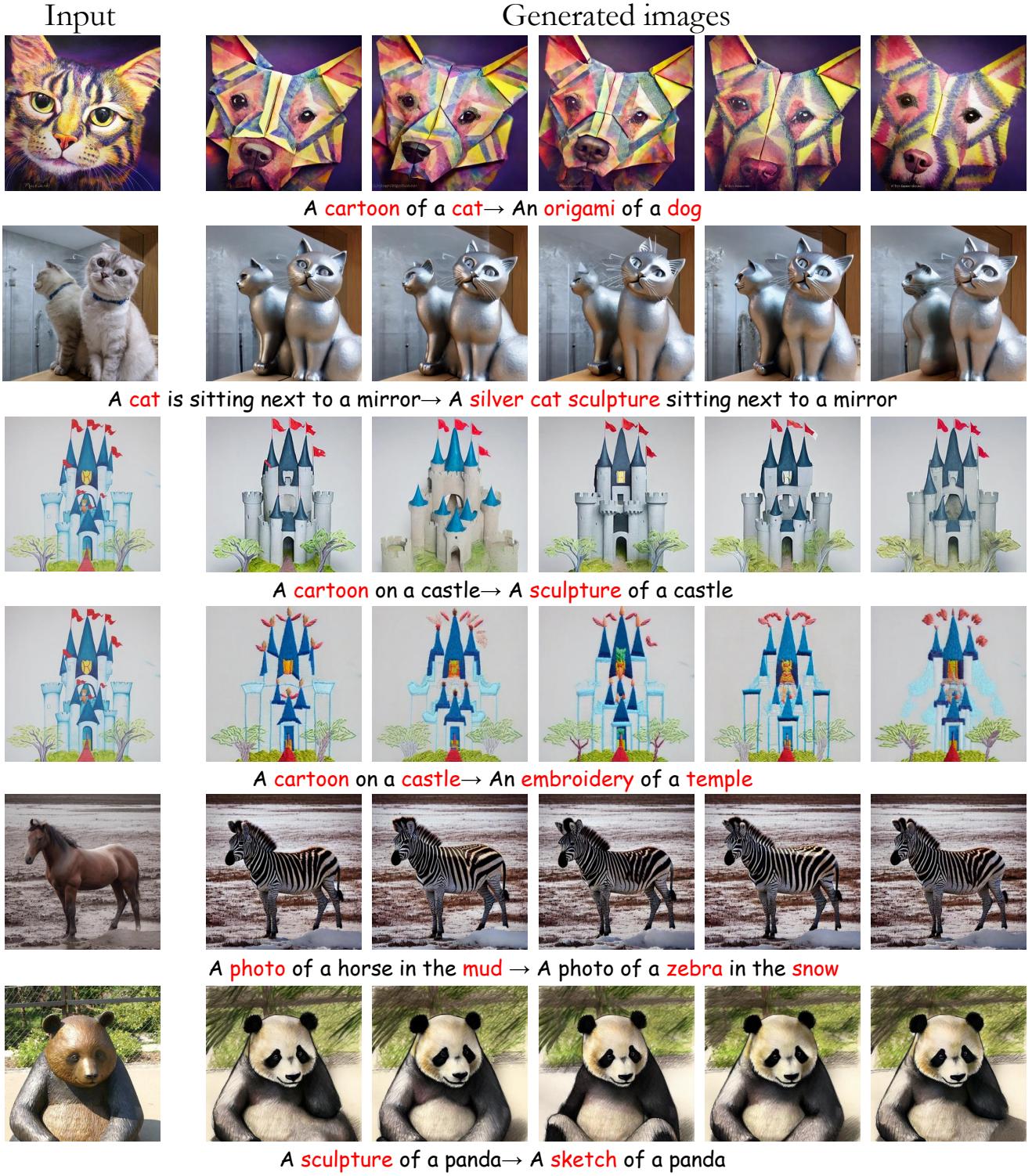


Figure 17: **Additional results for diverse text-based editing with our method.** Notice that each edited result is slightly different. For example, the eyes and nose of the origami dog change between samples, and so do the zebra's stripes.

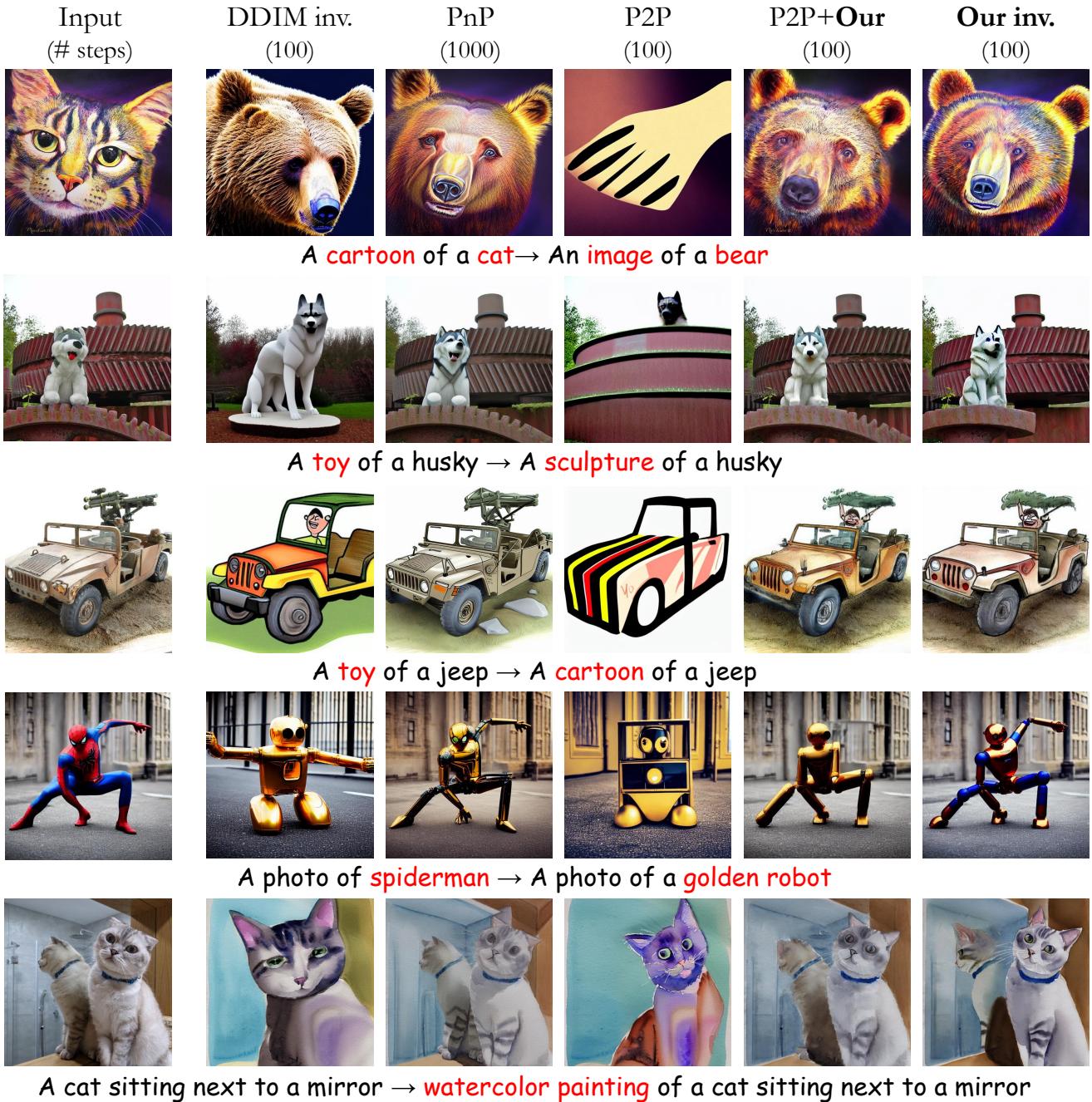


Figure 18: Qualitative comparisons between all methods.

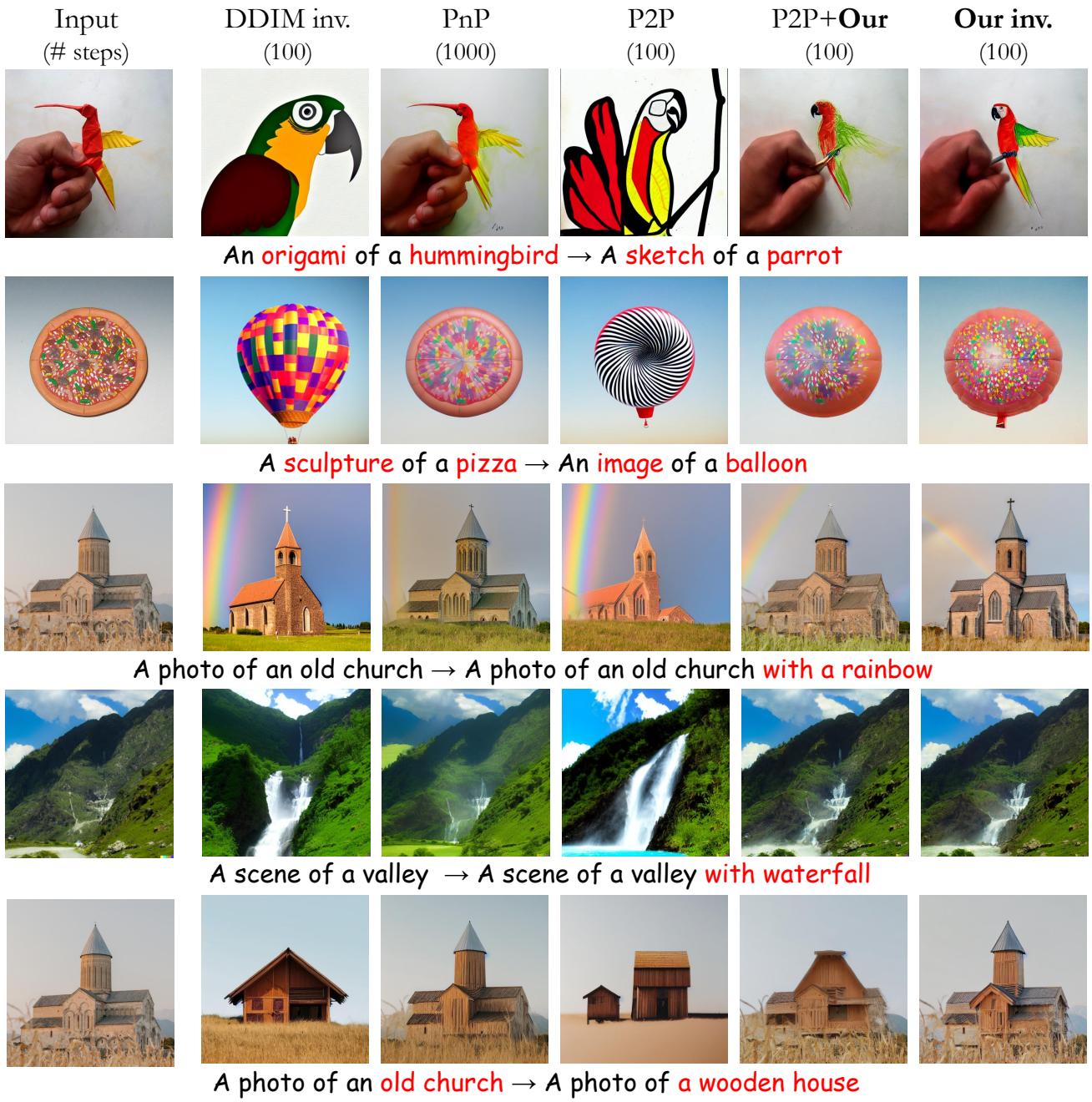


Figure 19: Additional qualitative comparisons between all methods.