Delta Denoising Score

Amir Hertz[1,2] Kfir Aberman[1] Daniel Cohen-Or[1,2]

[1]*Google Research* [2]*Tel Aviv University*

Abstract

This paper introduces Delta Denoising Score (DDS) a novel diffusion-based scoring technique which optimizes a parametric model for the task of image editing. Unlike the existing Score Distillation Sampling (SDS), which queries the generative model with a single image-text pair, DDS utilizes an additional fixed query of a reference image-text pair to generate delta scores that represent the difference between the outputs of the two queries. By estimating noisy gradient directions introduced by SDS using the source image and its text description, DDS provides cleaner gradient directions that modify the edited portions of the image while leaving others unchanged, thereby yielding a distilled edit of the source image. The analysis presented in this paper supports the power of the new score for image-to-image translation. We further show that the new score can be used to train an effective zero-shot image translation model. The experimental results show that the proposed loss term outperforms existing methods in terms of stability and quality, highlighting its potential for real-world applications.

1. Introduction

Large-scale language-vision models have revolutionizing the way images and visual content, in general, can be generated and edited. Recently, we are witnessing a surge in the development of text-to-image generative models, which utilize textual input to condition the generation of images. A promising avenue in this field is Score Distillation Sampling (SDS) [24] – a sampling mechanism that utilizes probability density distillation to optimize a parametric image generator using a 2D diffusion model as a prior.

The effectiveness of the SDS stems from rich generative prior of the diffusion model it samples from. This is in contrast to the direct use of a language-vision model, like CLIP, which was trained using contrastive loss [25]. The prior of large generative diffusion models, like Stable Diffusion [27], DALLE-2 [26] and Imagen [31] is particularly rich and expressive and has been demonstrated to be highly effective in generating visually stunning assets across various domains, including images, 3D models and more.

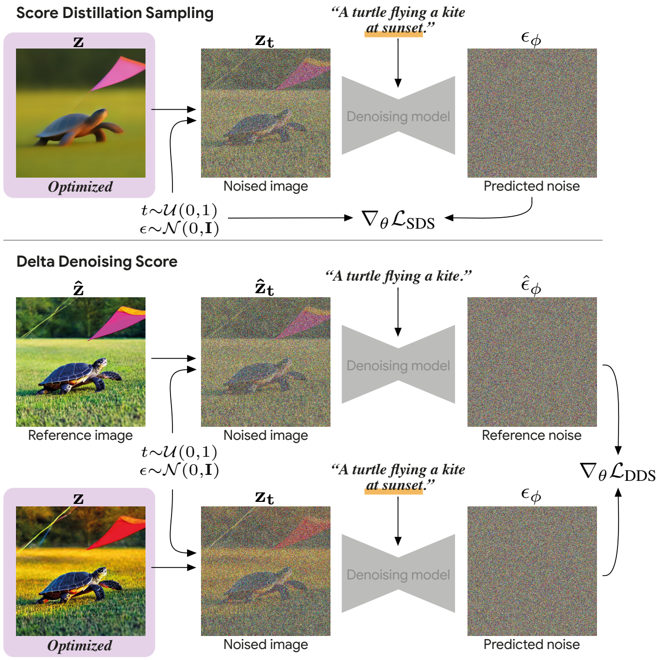

Figure 1: **Delta Denoising Score (DDS) vs. Score Distillation Sampling (SDS).** *Top: SDS allows optimizing a given image by querying the denoising model guided by a text prompt. However, it results in a blurred image away from the edited elements. Bottom: DDS, queries an additional reference branch aligned with its text-prompt, and generates delta scores that represent the difference between the outputs of the two queries. DDS provides cleaner gradient directions that modify the edited portions of the optimized image, while leaving the others unchanged.*

Despite its usefulness, one of the primary issues associated with SDS is its tendency to converge towards specific modes, which often leads to the production of blurry outputs that only capture the elements explicitly described in the prompt. In particular, using SDS to *edit an existing image* by initializing the optimization procedure from that image, may result in significant blurring of the image beyond the edited elements.

In this paper, we introduce a new diffusion-based scoring technique for optimizing a parametric model for the task of editing. Unlike SDS, which queries the generative model with a pair of image and text, our method utilizes an additional query of a reference, aligned, pair; that is, the text describes and agrees with the content of the image. Then, the output score is the difference, or *delta*, between the results of the two queries (see Figure 1). We refer to this scoring technique as Delta Denoising Score (DDS).

In its basic form, DDS is applied on two pairs of images and texts, one is a reference image-text that remains intact during the optimization, and the other is a target image that is optimized to match a target text prompt. The delta scoring provides effective gradients, which modify the edited portions of the image, while leaving the others unchanged. The key idea is that the source image and its text description, can be used for estimating undesirable and noisy gradients directions introduce by SDS. Then, if we want to alter only a portion of the image using a new text description, we can use our reference estimation and get a cleaner gradient direction to update the image.

DDS can be used as a prompt-to-prompt editing technique that can modify images by only editing their captions, where no mask is provided or computed. Beyond that, Delta Denoising Score enables us to train a distilled image-to-image model without the need of paired training dataset, yielding a zero-shot image translation technique. Training the model, requires only dataset of the source distribution, associated with simple captions that describe the source and target image distributions. As we shall show, such zero-shot training can be applied on a single or multi-task image translation, and the source distribution can include synthetically generated and real images.

To demonstrate the effectiveness of our approach, we conducted experiments comparing our model to existing state-of-the-art text-driven editing methods.

2. Related Work

Text-to-Image models [30, 26, 27], have recently raised the bar for the task of generating images conditioned on a text prompt, exploiting the powerful architecture of diffusion models [12, 32, 35, 12, 33, 27].

Recent works have attempted to adapt text-guided diffusion models to the fundamental challenge of single-image editing, aiming to exploit their rich and diverse semantic knowledge. Meng et al. [18] add noise to the input image and then perform a text-guided denoising process from a predefined step. Yet, they struggle to accurately preserve the input image details, which were preserved by a user provided mask in other works [21, 2, 1]. DiffEdit [7] uses DDIM inversion for image editing, but avoids the emerged distortion by automatically producing a mask that allows background preservation.

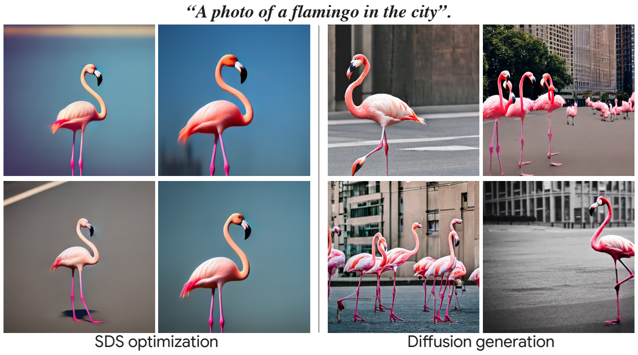

Figure 2: **Sampling text-to-image diffusion models.** *Generation via SDS optimization starting from random noises (left) vs. conventional diffusion-based image generation (right). Both samples are generated with respect to a given text prompt (top). Generating images based on SDS only leads to less diverse results and mode collapse where the main subject in the text appears in front of a blurry background.*

While some text-only editing approaches are bound to global editing [8, 17, 16, 23], Bar-Tal et al. [4] propose a text-based localized editing technique without using any mask. Their technique allows high-quality texture editing, but not modifying complex structures, since only CLIP [25] is employed as guidance instead of a generative diffusion model. Prompt-to-prompt [11] suggests an intuitive editing technique that enables manipulation of local or global details for images that were synthesized by a text-to-image network. [20] proposed an approach to invert real images into the latent space of the diffusion model, such that prompt-to-prompt can be applied to real images. Imagic [15] and UniTune[38] have demonstrated impressive text-driven editing capabilities, but require the costly fine-tuning of the model. InstructPix2Pix [5], plug-and-play [37] and [22] can get an instruction or target prompt and manipulate real images towrds the desired edit

DreamFusion [24] proposed the SDS score which is used in [34] to direct a StyleGAN generator for domain adaption. This is conceptually similar to StyleGAN-NADA [9] which uses instead CLIP [25] to translate the domain of a StyleGAN generator to other domains based only textual description.

3. Delta Denoising Score (DDS)

We begin with a brief overview of the SDS loss function and explain the challenges in sampling and editing images

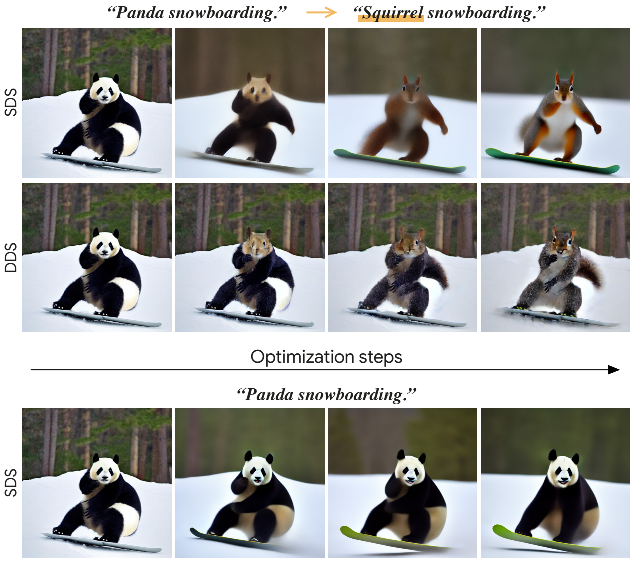

Figure 3: **Bias in SDS optimization.** *On left column is an image generated by the prompt "Panda snowboarding". Top rows show the difference between SDS to DDS optimization when changing the animal in the prompt. Bottom row shows SDS optimization applied using the original prompt. Even in this case, the image become blurry.*

with SDS, based on empirical observations. In particular, we demonstrate that SDS introduces a noisy direction when applied to the task of image editing. We then introduce our Delta Denoising Score (DDS), which utilizes a reference pair of image and text to correct the noisy direction of SDS and offers a new technique for the task of prompt-to-prompt editing [11]. We conduct all our experiments using the latent model– Stable Diffusion [27], nevertheless, in our overview and results, we refer to the models latents and output channels as images and pixels respectively.

3.1. SDS overview

Given an input image $\mathbf{z}$, a conditioning text embedding y, a denoising model ϵ_ϕ with parameters set ϕ, a randomly sampled timestep $t \sim \mathcal{U}(0,1)$ drawn from the uniform distribution, and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ following a normal distribution, the diffusion loss can be expressed as:

$$\mathcal{L}_{\text{Diff}}\left(\phi, \mathbf{z}, y, \epsilon, t\right) = w(t) ||\epsilon_\phi\left(\mathbf{z_t}, y, t\right) - \epsilon||_2^2,$$

where $w(t)$ is a weighting function, and $\mathbf{z_t}$ refers to the noisy version of $\mathbf{z}$ obtained via a stochastic noising forward process given by $\mathbf{z_t} = \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon$, with α_t being the noise scheduler. For simplicity, we omit the weighting factor in the remainder of this section.

The text conditioned diffusion models use classifier-free guidance (CFG [13]) that consists of two components, one that is conditioned on text input, and another that is unconditioned. During inference, the two components are used to denoise the image via

$$\epsilon_\phi^\omega\left(\mathbf{z_t}, y, t\right) = \left(1 + \omega\right) \epsilon_\phi\left(\mathbf{z_t}, y, t\right) - \omega \epsilon_\phi\left(\mathbf{z_t}, t\right),$$

where the components are balanced using a guidance parameter ω.

Given an arbitrary differentiable parametric function that renders images, g_θ, the gradient of the diffusion loss function with respect to the parameters θ is given by:

$$\nabla_\theta \mathcal{L}_{\text{Diff}} = \left(\epsilon_\phi^\omega\left(\mathbf{z_t}, y, t\right) - \epsilon\right) \frac{\partial \epsilon_\phi^\omega\left(\mathbf{z}, y, t\right)}{\partial \mathbf{z_t}} \frac{\partial \mathbf{z_t}}{\partial \theta}.$$

It has been demonstrated in [24] that omitting the U-Net Jacobian term (middle term) leads to an effective gradient for optimizing a parametric generator with diffusion models:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\mathbf{z}, y, \epsilon, t) = \epsilon_\phi^\omega\left(\left(\mathbf{z_t}, y, t\right) - \epsilon\right) \frac{\partial \mathbf{z_t}}{\partial \theta}. \tag{1}$$

Incrementally updating the parameters of the generator in the direction of the gradient, produces images that exhibit a higher degree of fidelity to the prompt. However, SDS suffers from the tendency to converge towards specific modes, resulting in non-diverse and blurry outputs that only highlight elements mentioned in the prompt. Figure 2 showcases a comparison between sampling Stable-Diffusion with SDS vs. sampling it with a standard reverse process of the diffusion model, demonstrating this issue with 2D image samples.

The original purpose of SDS was to generate samples via optimization from a text-conditioned diffusion model. It is noteworthy that g_θ can be an arbitrary parametric function that renders images. In the following sections we demonstrate our results with $g_\theta = \theta$, namely, a trivial generator that renders a single image, where the optimization variables are the image pixels themselves, however, note that the derivation is general.

3.2. Editing with SDS

The original purpose of SDS was to generate samples from a distribution conditioned solely on a text prompt. However, we now aim to extend SDS to the task of editing, which involves conditioning the sampling process on both an image and text.

Our objective is to synthesize an output image $\mathbf{z}$ that incorporates the structure and details of an input source image $\hat{\mathbf{z}}$, while conforming to the content specified in a target prompt y. This is a standard text-driven image-to-image translation problem, where modifications may be applied locally or globally [11, 5].

One potential approach to utilize SDS is to initialize the optimization variable with the source image $\mathbf{z}_0 = \hat{\mathbf{z}}$ and

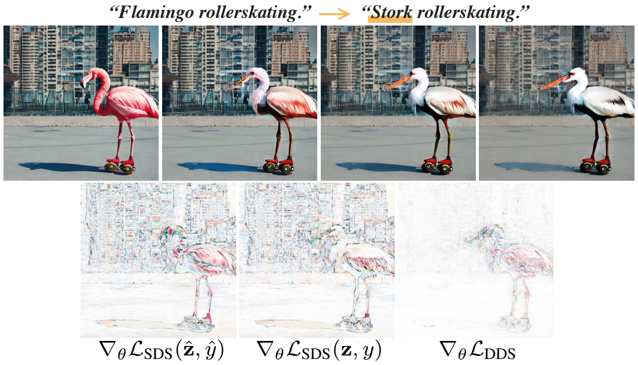
"Flamingo rollerskating." → "Stork rollerskating."
$\nabla_\theta\mathcal{L}_{\text{SDS}}(\hat{\mathbf{z}},\hat{y})$
$\nabla_\theta\mathcal{L}_{\text{SDS}}(\mathbf{z},y)$
$\nabla_\theta\mathcal{L}_{\text{DDS}}$

Figure 4: **DDS gradients.** *Bottom visualize the update of one DDS optimization. By subtracting the the SDS gradient of the reference image (left), from the SDS edited gradient of the middle image we get our cleaner DDS update (right).*

applying SDS while conditioning on y. However, we have observed that similarly to the non image conditioned SDS, this approach leads to blurred outputs and a loss of details, particularly those that are unrelated to the input prompt. Figure 3 (top row) demonstrates such example where the panda transforms into a squirrel at the cost of blurring out other details.

Based on our observations, we define a decomposition for the gradients $\nabla_\theta\mathcal{L}_{\text{SDS}}$ to two components: one component δ_{text} is a desired direction that directs the image to the closest image that is aligned with the text. And another, undesired component, δ_{bias} that interferes with the process and causes the image to become smooth and blurry in some parts. Formally:

$$\nabla_\theta\mathcal{L}_{\text{SDS}}(\mathbf{z},y,\epsilon,t) := \delta_{\text{text}} + \delta_{\text{bias}}, \tag{2}$$

where both δ_{text} and δ_{bias} are random variables that depend on $\mathbf{z}$, y, ϵ and t. Under this definition, to address this issue and enable high-quality or distilled image editing with SDS, we have to isolate and extract the text-aligned part δ_{text} and follow it during the optimization while avoiding the δ_{bias} direction that may take the image to unintended places.

3.3. Denoising the Editing Direction

We next aim to find the noisy direction of the SDS score, when applied for editing purposes, and remove it during the optimization process.

The gist of our method is that since we already have a source image and its text description, they can be used for estimating the noisy direction δ_{bias}, that biases the edit towards undesired directions. Then, if we want to alter only a portion of the image using a new text description, we can use our reference estimation and get a *cleaner* gradient direction to update the image. In practice, we use a reference branch that calculates the SDS score of the given image $\hat{\mathbf{z}}$ with a corresponding, aligned, text prompt $\hat{y}$, and subtract it from the main SDS optimization branch to yield a distilled edit.

Formally, given aligned and unaligned image-text embedding pairs $\hat{\mathbf{z}}$, $\hat{y}$, $\mathbf{z}$, y respectively, the delta denoising loss is given by:

$$\mathcal{L}_{\text{DD}}\left(\phi,\mathbf{z},y,\hat{\mathbf{z}},\hat{y},\epsilon,t\right) = ||\epsilon_\phi^\omega\left(\mathbf{z_t},y,t\right) - \epsilon_\phi^\omega\left(\hat{\mathbf{z}}_\mathbf{t},\hat{y},t\right)||_2^2,$$

where $\mathbf{z_t}$ and $\hat{\mathbf{z}}_\mathbf{t}$ share the same sampled noise ϵ and timestep t. Then, the gradient over $g_\theta = \mathbf{z}$, are given by

$$\nabla_\theta\mathcal{L}_{\text{DD}} = \left(\epsilon_\phi^\omega\left(\mathbf{z_t},y,t\right) - \epsilon_\phi^\omega\left(\hat{\mathbf{z}}_\mathbf{t},\hat{y},t\right)\right)\frac{\partial\epsilon_\phi^\omega\left(\mathbf{z_t},y,t\right)}{\partial\mathbf{z_t}}\frac{\partial\mathbf{z}}{\partial\theta}.$$

Again, we omit the differentiation thorough the diffusion model to obtain the Delta Denoising Score,

$$\nabla_\theta\mathcal{L}_{\text{DDS}} = \left(\epsilon_\phi^\omega\left(\mathbf{z_t},y,t\right) - \epsilon_\phi^\omega\left(\hat{\mathbf{z}}_\mathbf{t},\hat{y},t\right)\right)\frac{\partial\mathbf{z}}{\partial\theta}. \tag{3}$$

We state that DDS pushes the optimized image into the direction of the target prompt without the interference of the noise component, namely, $\nabla_\theta\mathcal{L}_{\text{DDS}} \approx \delta_{\text{text}}$.

By adding and subtracting ϵ from the term in (3), we can represent DDS as a difference between two SDS scores:

$$\nabla_\theta\mathcal{L}_{\text{DDS}} \quad = \quad \nabla_\theta\mathcal{L}_{\text{SDS}}(\mathbf{z},y) - \nabla_\theta\mathcal{L}_{\text{SDS}}(\hat{\mathbf{z}},\hat{y}). \tag{4}$$

We first claim that the score provided by the reference branch is equivalent to the noisy direction. This is because, ideally, an aligned image-text pair should have a low average SDS gradient across various timesteps and noise instances. Therefore, any non-zero gradient can be attributed to the noisy direction, thus,

$$\nabla_\theta\mathcal{L}_{\text{SDS}}(\hat{\mathbf{z}},\hat{y}) = \hat{\delta}_{\text{bias}}. \tag{5}$$

Evidently, the score of an aligned text-to-image pair is non-zero. As can be seen in Figure 3 (bottom, row), even when the optimization process starts with an image that was generated by the text, there are gradients that pull the image towards the non-desired modes. For further empirical results of the estimation of δ_{bias}, please refer to Section 5.

We next claim that the noisy component δ_{noise} of closely related images (e.g., images with similar structure that were created with close prompts) is similar. This is demonstrated in the DDS evaluation experiment in Section 5 and in Figure 8 which shows that the consine similarity between the directions of the aligned pair is high. This means that $\delta_{\text{bias}} \approx \hat{\delta}_{\text{bias}}$.

By combining the conclusions drawn from the above-mentioned experiments, we get $\nabla_\theta\mathcal{L}_{\text{DDS}} \approx \hat{\delta}_{\text{text}}$, which indicates that our DDS can be considered a distilled direction that concentrates on editing the relevant portion of the image, such that it's aligned with the target text.

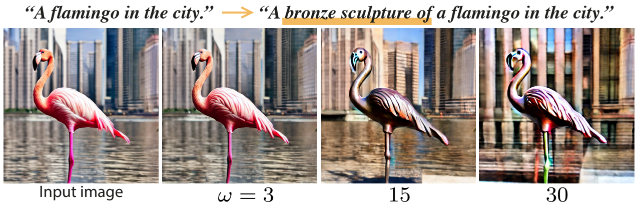
"A flamingo in the city."
"A bronze sculpture of a flamingo in the city."
Input image
$\omega = 3$
15
30

Figure 5: **DDS optimization results using different values of classifier free guidance scale** ω**.** *On the one hand, small values of* ω *led to slow convergence and low fidelity to text prompt. On the other hand, large* ω *results with low fidelity to the input image.*

Figure 4 visualizes the key idea behind DDS, The figure shows the two noisy SDS scores, of the aligned and unaligned pair, along with their difference, which comprises DDS. Notably, subtracting the two noisy scores produces a clear and concise score that concentrates solely on the targeted modification in the image.

Effect of CFG on DDS As previously noted, the Classifier Free Guidance (CFG) parameter ω, regulates the relative influence of the text-conditioned and unconditional components of the denoising objective. Apparently, despite the subtraction of the two distinct branches in DDS, ω still has a discernible impact on the resulting image output. Our experiments show that small values of ω yield slower convergence rates and a correspondingly diminished fidelity to the text prompt, while larger ω values result in an attenuated fidelity to the input image. This observed phenomenon is visualized in Figure 5 and empirically evaluated in Section 5.

4. Image-to-Image Translation

With our Delta Denoising Score, we can apply a direct optimization over the image pixel space, i.e. optimizing for $z = \theta$ as illustrated in Figure 1. However, optimizing an image for each editing operation presents several drawbacks. Firstly, it necessitates captions for both the input and the desired edited image. Secondly, the results obtained on real images are inferior to those obtained from synthetic images. Lastly, the time required for inference is long (∼20 seconds per edit). To overcome these limitations, we introduce a novel unsupervised training pipeline for text-driven image-to-image translation based on our proposed DDS.

Unsupervised training with DDS Using DDS, we introduce an unsupervised training framework for a neural network that learns to translate images based on a caption that describes a *known source distribution* and another caption that describes an *unknown target distribution*. Given a dataset of source images $\{\hat{\mathbf{z}}_i\}$, source caption $\hat{y}$ and a target caption y, our goal is to learn a mapping $\mathbf{z} = g_\theta(\hat{\mathbf{z}})$ such that $\mathbf{z}$ has high fidelity to both: the input image $\hat{\mathbf{z}}$ and to the target caption y. As illustrated in Figure 6, on the bottom, we utilize the DDS formulation in (4) to optimize our network.

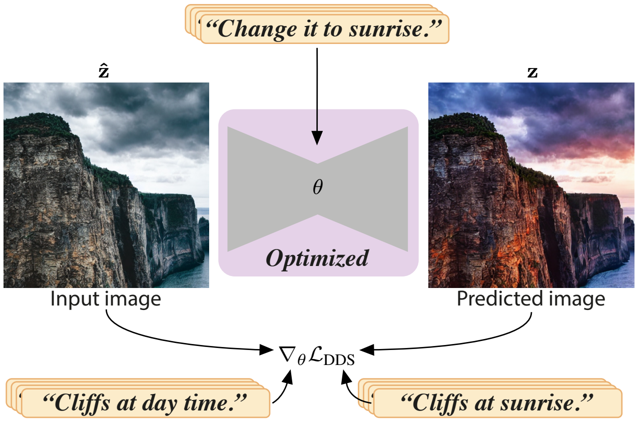

Figure 6: **Unsupervised training for multi task image-to-image translation network.** *Given an input image input image* $\hat{\mathbf{z}}$ *(left) and a sampled task embedding (on top), our network is trained by the double denoising score using the corresponding text embeddings (bottom) that describe the input image and the desired edited image result* $\mathbf{z}$*.*

Naturally, we can extend the network capabilities to be task conditioned. Under those settings, the network learns a finite set of M image-translation tasks that are defined by multiple target captions $\{y_i\}_{j=1}^{M}$ and corresponding learned task embeddings $\{k_j\}_{j=1}^{M}$, see Figure 6. At each optimization iteration, we sample a source image $\hat{\mathbf{z}}_i$ with its source caption $\hat{y}_i$, a task embedding k_j with the corresponding target caption y_j. Then the network is optimized by the DDS 4 where $\mathbf{z} = g_\theta(\hat{\mathbf{z}}_i|k_j)$.

To maintain the fidelity to the input image, we add a weighted identity regularization term:

$$\mathcal{L}_{\text{ID}} = \lambda_{id}(t)||g_\theta(\hat{\mathbf{z}}_i|k_j) - \hat{\mathbf{z}}_i||_2^2,$$

where the weight $\lambda_{\text{ID}}(t)$ is a function of the training iteration t, such that at the beginning of the training, we inject prior knowledge on the desired output, and gradually reduce it during training with a cosine decay.

DDS with CFG warmup During the training of the aforementioned network, we experienced a familiar *mode collapse* phenomena associated with the training of generative adversarial network (GAN) [10], where the network optimizations led to a local minima. In our case, the network has learned to produce a fixed object, in a fixed location within the input image, as demonstrated in Figure 7, where the same type of lion appears in the same pose and locations

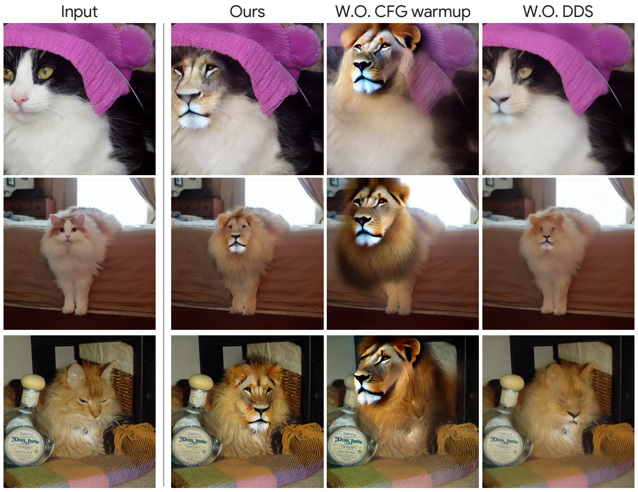

Figure 7: **Ablation study.** *We train a cat to lion image translation network under different settings. First and second columns are the input and ouuput results of our full methods. Third row are the results when training witout CFG warmup and the last column are the results when training with SDS instead of DDS.*

in all the outputs without respecting the input image. The reason for the mode collapse in our case can be explained thorough the analogy to GANs. The discriminator output score that discriminates between real and fake images can be replaced by the delta denoising score. At a local minimum point, our network succeeded to *fool* the DDS such that the output has high fidelity to y at the fixed region and high fidelity to $\hat{\mathbf{z}}$ elsewhere.

To address this issue we have found that implementing a warmup scheduler for the classifier free guidance parameter ω, utilized in the estimation of the DDS gradient can be effective. As we have demonstrated earlier, adopting a low value for ω during zero-shot optimization is associated with a notably slow convergence rate. Conversely, high values push the image aggressively towards y and lead the training to mode collapse. By gradually increasing the guidance scale, the network gradually learns to make larger changes to the input image with respect to the translation task and avoids local minima.

5. Evaluations and Experiments

In this section we evaluate our observation regarding the SDS and DDS scores, compare our approach to other state-of-the-art zero-shot editing methods and conduct an ablation study to show the effectiveness of different choices in our system.

SDS evaluation We measure the expected SDS norm as a function of the timestamp t for aligned and unaligned image-text pairs. The aligned pairs obtained by generating images using Stable Diffusion [27] with subset of 100 captions from COCO validation dataset [6]. Then, for each image $\mathbf{z}$, caption y and timestep t we estimate the value $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}||\nabla_{\mathbf{z}}\mathcal{L}_{\text{SDS}}(\mathbf{z},y)||_2$ by averaging the result of 200 measurements and report the average value of the 100 estimations. To provide a reference, we also perform the experiment on 100 unaligned image-text pairs obtained by permuting the captions of the aligned set. The results are shown in Figure 8 (left). As can be seen, SDS exhibits non-negligible high gradients for aligned pairs. In addition, the gap between aligned and unaligned pairs supports our observation in Section 3 that there is an inherent noise direction δ_{bias} in the SDS gradient.

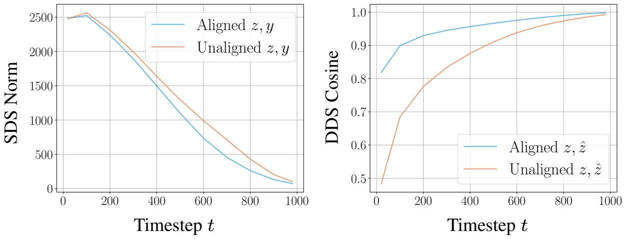

Figure 8: **Expected SDS gradients.** *On left we estimate the Expected SDS norm $||\nabla\mathcal{L}_{SDS}(z,y)||_2$ across different timesteps for aligned (blue) and unaligned (orange curve) synthetic image-text pairs. On right we measure the Cosine similarity between the SDS directions in Eq (4) on aligned (blue) and unaligned (orange) images from the Instruct-Pix2Pix dataset.*

DDS evaluation Next, we evaluate our estimation that for an aligned pair of similar images with their corresponding text, the SDS noise directions, δ_{bias} and $\hat{\delta}_{\text{bias}}$, are correlated. For this experiment we use a subset of 10000 synthetic image pairs $\mathbf{z}$ and $\hat{\mathbf{z}}$ with their corresponding captions y and $\hat{y}$ from InsturctPix2Pix [5] dataset . For each timestamp, we estimate the cosine similarity between $\nabla_{\mathbf{z}}\mathcal{L}_{\text{SDS}}(\mathbf{z},y)$ and $\nabla_{\hat{\mathbf{z}}}\mathcal{L}_{\text{SDS}}(\hat{\mathbf{z}},\hat{y})$ and report the average result across all pairs. Here again, we applied the same experiment to unaligned pairs for reference. Note that the caption for each SDS estimation remained aligned to its image. The results are summarized in Figure 8, on right. As can be seen, the aligned pairs are strongly correlated which supports our assumption that an estimation for δ_{bias} of reference image and text can be used the eliminate the same term from similar pair. More evaluations for on the effect of CFG over DDS can are showed in the appendix.

Comparison to zero-shot editing methods To evaluate our editing capability using a direct DDS optimization over the pixel space of a synthetic generated image, we use a randomly selected subset of 1000 pairs of source and target prompts from the dataset of InsturctPix2Pix [5]. The

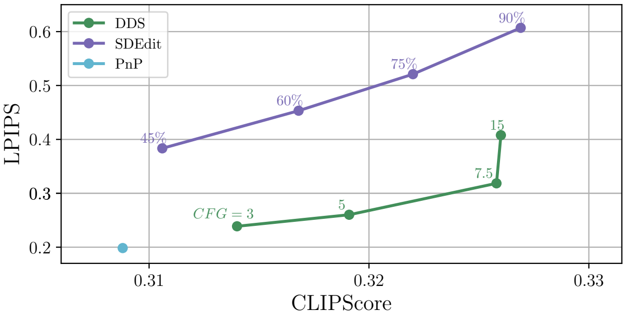

Figure 9: **Zero shot image editing comparison.** *We apply our DDS optimization with different values of CFG over* 1000 *images and prompts from the InsturctPix2Pix training dataset. Our method achieves high fidelity to edits described in the text prompts (high CLIP score) while maintaining high fidelity to the source images (low LPIPS)*

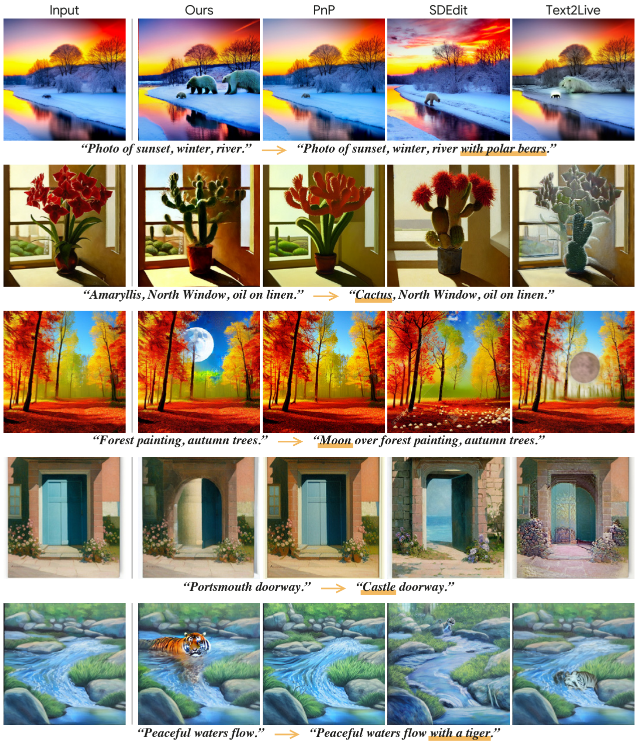
Input
Ours
PnP
SDEdit
Text2Live
"Photo of sunset, winter, river." → "Photo of sunset, winter, river with polar bears."
"Amaryllis, North Window, oil on linen." → "Cactus, North Window, oil on linen."
"Forest painting, autumn trees." → "Moon over forest painting, autumn trees."
"Portsmouth doorway." → "Castle doorway."
"Peaceful waters flow." → "Peaceful waters flow with a tiger."

Figure 10: **Zero shot image editing.** *Our method, on the left, better apply structural or color changes over the input image as described in the target text prompt while preserving high fidelity to the input image.*

dataset already includes the paired image results obtained by Prompt-to-Prompt (P2P) [11], from which we took only the source images. For each editing result we measure the text-image correspondence using CLIP score[25]. In addition, we evaluate the similarity between the original and the edited images using the LPIPS perceptual distance [39]. We compare our method to additional zeros shot methods: SDEdit [18], Plug-and-Play (PnP) [37] and Text2Live [4]. It can be seen in Figure 10 that comparing to other methods, our approach demonstrates higher fidelity to the text prompt and to the source image on average. The quantitative results are summarized in Figure 9 where we show the metrics of our method for different numbers of classifier free guidance scale. Notice, that as observed in Figure 5, the improvement to the fidelity to text that obtained by using a large value of CFG is negligible compared to the deterioration in the fidelity to the source image.

Table 1: User evaluation for zero shot image editing quality (see Sec. 5). We report the percentage of judgments in our favor over 800 answers (2400 in total).

PnP	SDEdit	Text2LIVE
70.9%	72.1%	75.5%

In addition, we conducted a user study. In each question we randomly selected an example and two editing results and the user was asked to choose which one is a better editing result in terms of fidelity to the target text, source image and overall quality. We collected 2400 answers from 100 users using Amazon Mechanical Turk service. We report in table 1 the percentage of judgments in our favor. As seen, our method outperforms all baselines by a large margin.

Image-to-image translation training We train different multi-task networks as described in Section 4. For each training instance, we generate a synthetic dataset of 5000 images using the Stable Diffusion model conditioned on manually written captions (5-20 captions for each dataset). Each training starts from a pre-trained Stable Diffusion model, modified as follows: The latent noise inputs are replaced with latents of images from our synthetic dataset. The text embedding condition is replaced with our learned task embeddings, initialized with text embedding that describes the task embedding. For example, for the task of adding snow in an image, we use the phrase "snowing". Finally, the timestep of the diffusion process is no longer used since our model inference process contains a single feed-forward. Therefore, we re-use the timestep condition as additional per-task *learned* embedding which is initialized with a positional embedding of $t = 0.5$. While the text condition is injected by cross attention, time by adaptive group normalization (ADAGN). Additional implementation details are provided in the supplementary material.

Image-to-image translation comparison We evaluate a *Cat-to-Other* network trained to translate images of cats to

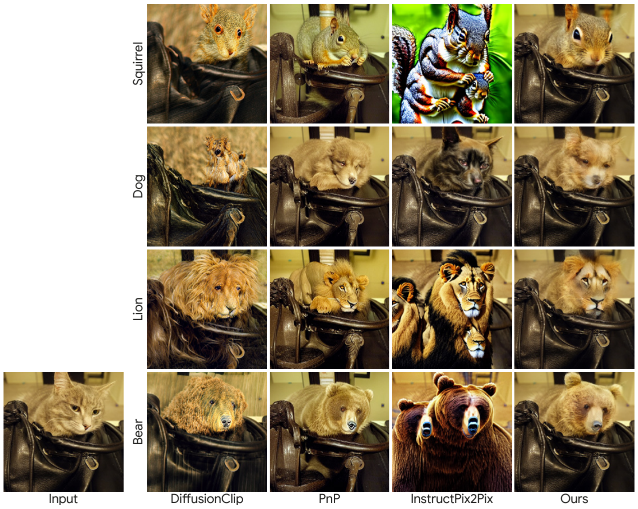

Figure 11: **Image-to-Image translation comparison.** *Our multi-task network was trained to change a cat to different animals by the DDS. It was trained on synthetic cat photos and evaluate on subset of photos from COCO and Imagenet datasets. Our results (right column) better preserve the structure of the cat in the input image and its background.*

Table 2: **Quantitative comparison for the multi-task image-to-image translation network.** We measure text-image correspondence using CLIP [25]. In addition, we evaluate the similarity between the original and the edited images using the LPIPS [39] perceptual distance.

	CLIP score ↑	LPIPS ↓
DiffusionClip	$\mathbf{0.251 \pm 0.022}$	0.572 ± 0.0594
PnP	0.221 ± 0.036	0.31 ± 0.075
InstructPix2Pix	0.2190 ± 0.037	0.322 ± 0.215
DDS (ours)	0.225 ± 0.031	$\mathbf{0.104 \pm 0.061}$

images of four different animals: a dog, a lion, a bear, and a squirrel. We tested our network using a collection of 500 cat images from ILSVRC [29] and COCO [6] validation set; overall, we tested the results of 2000 image translations. We use the same LPIPS and CLIP scores to estimate fidelity to the source image and target text that describes the target distribution, for example, “A photo of a lion”. We compare our method to DiffusionCLIP [16], PnP [37] and InstructPix2Pix [5] which also utilize the generative prior of a pre-trained diffusion model for the task of image-to-image translation. Unlike our method, InstructPix2Pix fine tune a diffusion model using synthetic pairs of images and therefor it is sensitive to quality of the pairing method.

The results are summarized in Table 2 and Figure 11. As can be seen, our method achieves both: better fidelity to the input image and to desired target domain. Additionally, our method operates via a single feed-forward pass during inference, making it ×50 faster than the other diffusion-based methods that require a full iterative diffusion process in each inference sampling. A qualitative comparison is shown in Figure 11. As can be seen, our method better preserves the structure of the cat when translating to other animals. Moreover, our distilled training results in a more robust network that can better distinguish between regions in the image that had to be changed and areas to preserve.

Ablation Study We evaluate key components of our image-to-image translation network on a single task of cat-to-lion image translation. First, we show our results without the CFG scaling warmup. As shown in Figure 7 (third column), it results in mode collapse where roughly the same lion appears in the same location regardless of the cat in the input. In addition, we train a network with the *Vanilla* SDS term instead of our DDS while the other components, the $\mathcal{L}_{\text{ID}}$ term and the CFG warmup, remain untouched and prevent the mode collapse. As can be seen (right column in Figure 7), the quality of the translation to a lion is worse than our full settings. Moreover, the SDS training struggles to preserve high-frequency details in the input image. For example, see the patterns of the purple wool hat in the first row.

6. Conclusions , Limitations and Future work

We have presented, Delta Denoising Score, a new diffusion scoring technique that allows optimizing a given image as means to edit it with respect to a text-prompt. Delta Denoising Score uses the SDS score applied to input image to calculate cleaner gradients during the optimization, which leads to a distilled edit. We have also showed an image-to-image translation model trained with our new score. The model is training with no supervision, requires no pairs of images, and thus can be trained on real images.

Our Delta Denoising Score works well in distilling text-driven image-to-image translations. However, there are cases that the results are imperfect. For example, our method struggles in making significant pose changes or moving an object. In addition, DDS fails to preserve the subject’s identity while changing its posture, see Fig. 12 in the middle. However, our method can be used on top of personalised, fine-tuning step [28], see Fig. 12 on right.

We also acknowledge that the multi-task model can be better trained and may be further improved by combining multiple experts training [3], which uses multiple network, or utilize subset of paired data and train our network under semi-supervised settings.

The scope of Delta Denoising Score is wide, and its generalization across various editing tasks [5] should be explored in the future. Furthermore, we believe that it can be extended to other modalities, such as text-driven 3D shape editing, video editing and motion editing [19, 14, 36].

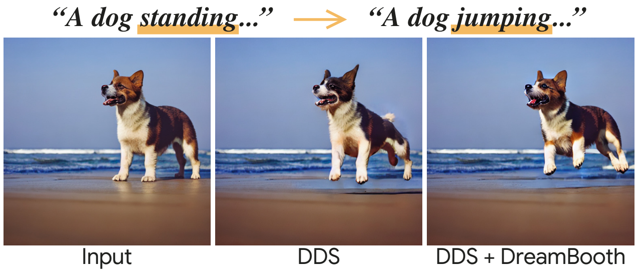

Figure 12: **Limitation.** *DDS struggles to preserve the identity of a dog (left) while modifying its pose (middle), without, first, applying personalization fine tuning [28] over the diffusion model (right).*

The objective of this work is to extract efficient and clean gradients that can facilitate the optimization of an image towards a distilled edit. This, we believe, is an important step towards enhancing our understanding of how to effectively extract and utilize the rich knowledge that is concealed within large-scale generative models.

7. Acknowledgement

We thank Ben Poole, Elad Richardson, Ron Mokady, Jason Baldridge and Yael Pritch for their valuable inputs that helped improve this work.

References

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

[2] 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*, pages 18208–18218, 2022. 2

[3] 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. 8

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022. 2, 7

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 3, 6, 8

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 6, 8

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[8] 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. *arXiv preprint arXiv:2204.08583*, 2022. 2

[9] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[11] 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, 3, 7

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[14] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 8

[15] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, abs/2210.09276, 2022. 2

[16] 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*, pages 2426–2435, 2022. 2, 8

[17] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021. 2

[18] 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, 7

[19] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 8

[20] 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

[21] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[22] 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. 2

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021. 2

[24] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 7, 8

[26] 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

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3, 6

[28] 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. 8, 9

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 8

[30] 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, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[31] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 1

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2

[34] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022. 2

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[36] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022. 8

[37] 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. 2, 7, 8

[38] 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

[39] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7, 8