Text-guided Explorable Image Super-resolution

Kanchana Vaishnavi Gandikota* Paramanand Chandramouli*
Institute for Vision and Graphics, University of Siegen

Abstract

*In this paper, we introduce the problem of zero-shot text-guided exploration of the solutions to open-domain image super-resolution. Our goal is to allow users to explore diverse, semantically accurate reconstructions that preserve data consistency with the low-resolution inputs for different large downsampling factors without explicitly training for these specific degradations. We propose two approaches for zero-shot text-guided super-resolution - i) modifying the generative process of text-to-image (*T2I*) diffusion models to promote consistency with low-resolution inputs, and ii) incorporating language guidance into zero-shot diffusion-based restoration methods. We show that the proposed approaches result in diverse solutions that match the semantic meaning provided by the text prompt while preserving data consistency with the degraded inputs. We evaluate the proposed baselines for the task of extreme super-resolution and demonstrate advantages in terms of restoration quality, diversity, and explorability of solutions.*

1. Introduction

The goal of image super-resolution is to recover a high-quality image, given a low-resolution (LR) observation $\mathbf{y}$,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \quad (1)$$

where $\mathbf{A}$, $\mathbf{x}$, and $\mathbf{n}$ represent the down-sampling operator, ground truth image, and measurement noise respectively. Image super-resolution is highly ill-posed, especially at large super-resolution factors with many valid solutions satisfying the data consistency accurately. While most of the recent state-of-the-art supervised deep networks for super-resolution [6, 31, 80] recover only a single image from this solution space, there are also methods utilizing conditional or unconditional generative models [2, 5, 38, 42, 46, 81], which allow sampling multiple solutions. A few of these works also allow exploring the solution space, using graphical user inputs [2] or semantic maps [5]. Yet, even these methods are tailored to images of specific classes such as faces, or trained for specific super-resolution factors. On the other hand, natural language provides a simpler and more intuitive means of conveying semantic concepts. For instance, it is easier to provide detailed descriptions or convey concepts such as age, gender, emotion, and race through text rather than graphical inputs alone. Therefore, a method that guides image super-resolution through text can greatly aid the exploration of semantically meaningful solutions.

In this paper, we propose for the first time zero-shot open-domain image super-resolution using simple and intuitive text prompts. Our goal is to explore via text prompts, diverse and semantically accurate reconstructions that preserve data consistency with the low-resolution inputs for different large downsampling factors without explicitly training for these specific degradations. Towards this goal, we exploit recent advances in text-to-image (*T2I*) generative models [64, 66, 68], contrastive language image pretraining (CLIP) [61], and unsupervised zero-shot approaches to image recovery using diffusion-based generative models [12, 74, 81]. We explore two paradigms for zero-shot text-guided super-resolution. In the first approach, we adapt recent diffusion-based zero-shot super-resolution approaches to *T2I* models by appropriately modifying the generative process. We consider recent state-of-the-art text-to-image diffusion models, open-sourced versions of DALL-e2 [64], and Imagen [68], and adapt these models for zero-shot super-resolution using different zero-shot approaches for diffusion-based image recovery [12, 74, 81]. As these *T2I* models comprise of a cascade of diffusion models at different resolutions, we modify the zero-shot approaches accordingly to deal with multi-stage generation. In the second approach, we modify an existing diffusion-based zero-shot restoration approach to incorporate additional language guidance through CLIP. We focus on extreme image super-resolution with large upscale factors, as this problem is severely ill-posed, and allows exploration of a larger solution space.

Fig. 1 illustrates the benefits of text guidance in extreme super-resolution. Existing zero-shot methods such as [12, 77] cannot recover realistic details when the ground truth has complex content as seen in the super-resolution results of [12] on challenging input. On the other hand, the use

* indicates the authors contributed equally.
Code available at:https://github.com/KVGandikota/Text-guidedSR

Figure 1. **Text guided image Super-resolution.** We explore consistent reconstructions to image super-resolution problems through text prompts while achieving perfect data consistency with the given inputs for all solutions. Shown are a) extreme super-resolution of natural images (top), b) face super-resolution (bottom), with an upsampling factor of 16.

of text enables the recovery of complex scene content with specific details matching the text prompt. The use of text also effortlessly improves diversity in solutions for face super-resolution in terms of age, expression, gender, race, and other attributes over [12, 81] which recover images with limited diversity. We evaluate the proposed baselines in terms of realism, fidelity with low-resolution input, and agreement with text, in several qualitative, quantitative, and human evaluations. Extensive experimental evaluations demonstrate the benefit of using zero-shot text guidance in terms of flexibility, diversity, and explorability of solutions to extreme super-resolution. Our work opens up a promising direction of developing efficient tools for text-guided exploration of image recovery.

2. Preliminaries

2.1. Denoising Diffusion Probabilistic Models (DDPM)

DDPM generative models [27] employ two diffusion processes: *i) A forward process* slowly noising a data sample $\boldsymbol{x}_0$ into Gaussian distribution $\mathcal{N}$ in T steps, with the evolution of a sample $\mathbf{x}_t$ at time-step t given by:

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &:= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \\ \textit{i.e.,} \quad \mathbf{x}_t &= \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{2}$$

where $\{\beta_t\}_{t=0}^T$ is the noise variance schedule. *ii) A learned reverse process* using iterative denoising to generate samples from the training data distribution $q(\mathbf{x})$ in T steps given by:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2 \mathbf{I}), \quad \text{where,}$$
$$\boldsymbol{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \right), \quad \text{and,}$$
$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t, \text{ with } \alpha_t = 1-\beta_t, \text{ and } \bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i, \tag{3}$$

and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the learned neural network noise approximator.

2.2. Range-Null Space Decomposition

When there is no measurement noise in (1), i.e. $\mathbf{y} = \mathbf{A}\mathbf{x}$, pseudoinverse operation $\mathbf{A}^\dagger \mathbf{y}$ produces the minimum norm solution with perfect data consistency. Any other solution $(\mathbf{A}^\dagger \mathbf{y} + \mathbf{x}_\delta)$ is also data consistent, as long as $\mathbf{x}_\delta$ lies in the null space of $\mathbf{A}$. Note that $\mathbf{x}$ can be decomposed as:

$$\mathbf{x} \equiv \mathbf{A}^\dagger \mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x}. \tag{4}$$

The component $(\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x}$ is in the null space of $\mathbf{A}$, (with $\mathbf{A}(\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x} \equiv \mathbf{0}$). Given an approximate solution $\bar{\mathbf{x}}$, Eq. 4 can be used to construct a data consistent solution [2, 72, 80, 81] given by $\hat{\mathbf{x}}$ as,

$$\hat{\mathbf{x}} = \mathbf{A}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\bar{\mathbf{x}}. \tag{5}$$

2.3. Zero-Shot Restoration using Diffusion Models

We now describe diffusion-based zero-shot restoration methods that we explore in this paper for text-guided restoration. Given a noisy sample $\mathbf{x}_t$ at step t, [73] obtain the clean estimate of $\mathbf{x}_0$ as:

$$\mathbf{x}_{0|t} := \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t} \right). \tag{6}$$

The works [12, 74] also utilize $\mathbf{x}_{0|t}$ of a similar form by approximating $p(\mathbf{x}_0|\mathbf{x}_t)$ as a simple Gaussian distribution. This estimate is used in a guidance function or in a projection to incorporate measurement consistency in the following methods.

Diffusion Posterior Sampling DPS [12] utilize $\mathbf{x}_{0|t}$ in reconstruction guidance as:

$$\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} \|\mathbf{y} - \mathbf{A}(\mathbf{x}_{0|t})\|_2^2, \tag{7}$$

where, the intermediate estimate at the previous step $\mathbf{x}'_{t-1}$ is obtained using the usual reverse step, and the gradient of the reconstruction loss for $\|\mathbf{y} - \mathbf{A}(\mathbf{x}_{0|t})\|_2^2$ with respect to $\mathbf{x}_t$ is obtained by differentiating through the diffusion model.

Pseudoinverse-Guided Diffusion Models ΠGDM [74] modify (7) by inverting the measurement model using pseudoinverse as:

$$\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} \|\mathbf{A}^\dagger \mathbf{y} - \mathbf{A}^\dagger \mathbf{A}(\mathbf{x}_{0|t})\|_2^2, \tag{8}$$

and demonstrate improved reconstructions with a reduced number of diffusion steps.

Denoising Diffusion Null-Space Models DDNM [81] utilize the range space-null space decomposition in the reverse diffusion process to rectify the clean estimate at each step to satisfy data consistency as:

$$\hat{\mathbf{x}}_{0|t} := \mathbf{A}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{x}_{0|t}. \tag{9}$$

This rectified data consistent estimate $\hat{\mathbf{x}}_{0|t}$ is used in subsequent sampling from $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$ in [81]. Compared to [12, 74], this method is faster as it does not require differentiation through diffusion model weights.

2.4. Text guided Image Generation with Diffusion Models

There are two approaches for text-guided image generation using diffusion models- training text-conditioned diffusion models, or incorporating text guidance into unconditional models using vision-language models such as CLIP [61].

2.4.1 Text-to-Image *T2I* Diffusion Models

We now describe the *T2I* generative models we employ for text-guided super-resolution.

DALL E-2 unCLIP [64] consists of: *i) a diffusion-based prior* to produce CLIP image embeddings [61] from encodings of the input prompt, *ii) a conditional diffusion-based decoder* to generate images conditioned on CLIP image embeddings and text prompts in a down-sampled pixel space, and *iii) a diffusion-based super-resolution module* to upsample the decoder output into a high-resolution image.

Imagen [68] utilizes a pretrained text encoder [62] to generate embeddings from input text which condition a cascade of conditional diffusion models to generate images of increasing resolutions. Different from unCLIP, Imagen uses only text embeddings which are used to condition every stage of image generation and super-resolution.

We consider unCLIP and Imagen diffusion models with two stages, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ operating on a down-sampled pixel space at resolution 64×64, and an upsampling stage $\boldsymbol{\zeta}_{\boldsymbol{\theta}}$ operating at resolution 256×256.

2.4.2 Training-free text-guided generation

The idea of training free guidance [3, 84] is to incorporate desired conditioning signal c into generation process through appropriate energy function E which measures the distance between desired condition and clean estimate $\mathbf{x}_{0|t}$ at every diffusion step:

$$\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} E(c, \mathbf{x}_{0|t}). \tag{10}$$

For text-guided generation, E can be defined using the distance between the text and image embeddings obtained through CLIP text and image encoders.

3. Methodology

Given a low-resolution image y with known degradation operator A, our goal is to generate data-consistent solutions u whose attributes can be varied using input text prompts c.

$$\begin{aligned} \textit{Data Consistency}: &\quad A\hat{u} \equiv f, \\ \textit{Semantic Consistency}: &\quad \hat{u} \sim q(u|c), \end{aligned} \tag{11}$$

where $q(u|c)$ denotes the distribution of images u with semantic meaning provided by the text prompt c. To obtain solutions satisfying semantic meaning provided by text prompt as well as measurement consistency, we explore the following methods:

1. Zero-shot super-resolution using *T2I* models.
2. Incorporating CLIP guidance into zero-shot diffusion-based restoration.

3.1. Text Guided Super-resolution using *T2I* Models

We consider two recent diffusion-based text-to-image (T2I) generative models that operate in the pixel domain DALL E-2 [64] and Imagen [68]. As these models employ a multi-stage generation process, first in down-sampled pixel-space, followed by upsampling stages, we correspondingly modify the sampling process to incorporate guidance or null-space consistency in both stages of the generation. Let $\mathbf{c_1}$ and $\mathbf{c_2}$ denote the conditioning signals in the two stages, $\boldsymbol{\epsilon_\theta}$ and $\boldsymbol{\zeta_\theta}$ denote conditioned diffusion models in the down-sampled stage, and the super-resolution stage respectively. The current estimate of the clean image at each step in the first stage and second stage are respectively given by,

$$\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon_\theta}(\mathbf{x}_{LR_t}, t|\mathbf{c}_1)\sqrt{1-\bar{\alpha}_t}\right), \tag{12}$$

$$\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\zeta_\theta}(\mathbf{x}_t, t|\mathbf{c}_2)\sqrt{1-\bar{\alpha}_t}\right). \tag{13}$$

For Imagen, $\mathbf{c}_1$ corresponds to text embeddings from the text encoder, and $\mathbf{c}_2$ contains $\mathbf{x}_{LR}$, in addition to text embeddings. For unCLIP $\mathbf{c}_1$ corresponds to a combination of CLIP image embeddings produced by the prior model and text embeddings, and $\mathbf{c}_2$ is the output $\mathbf{x}_{LR}$ of the first stage.

We first recover a lower resolution version $\mathbf{x}_{LR}$ by using a modified measurement $\mathbf{A}_{LR}$ which takes into account the downsampling operation for text-conditioned diffusion in low resolution. For the subsequent super-resolution using the model ζ_θ, we consider the actual measurement operator $\mathbf{A}$. We adapt zero-shot methods discussed in Sec. 2.3 as follows:

***T2I* -DPS:** We incorporate reconstruction guidance in both stages. In down-sampled pixel space, reconstruction guidance is given by

$$\mathbf{x}_{LR_{t-1}} \leftarrow \mathbf{x}'_{LR_{t-1}} - \rho_t \nabla_{\mathbf{x}_{LR_t}} \|\mathbf{y} - \mathbf{A}_{LR}\mathbf{x}_{LR_{0|t}})\|_2^2. \tag{14}$$

In the second stage, we incorporate reconstruction guidance following the standard DPS method given in Eq. (9).

***T2I* -ΠGDM:** We incorporate pseudoinverse guidance in both stages. In down-sampled pixel space, pseudoinverse guidance is given by

$$\begin{aligned} &\mathbf{x}_{LR_{t-1}} \leftarrow \mathbf{x}'_{LR_{t-1}} - \rho_t \nabla_{\mathbf{x}_{LR_t}} E(y, \mathbf{A}^\dagger{}_{LR}, \mathbf{x}_{LR_{0|t}}), \text{ where,} \\ &E(y, \mathbf{A}^\dagger{}_{LR}, \mathbf{x}_{LR_{0|t}}) = \|\mathbf{A}^\dagger{}_{LR}\mathbf{y} - \mathbf{A}^\dagger{}_{LR}\mathbf{A}_{LR}(\mathbf{x}_{LR_{0|t}})\|_2^2. \end{aligned} \tag{15}$$

The second stage incorporates pseudoinverse guidance following the standard ΠGDM approach given in Eq. (8).

***T2I* -DDNM:** We impose null-space consistency in both stages. In down-sampled pixel space, $\mathbf{x}_{LR_{0|t}}$ is rectified at each step as

$$\hat{\mathbf{x}}_{LR_{0|t}} = \mathbf{A}^\dagger{}_{LR}\mathbf{y} + (\mathbf{I} - \mathbf{A}^\dagger{}_{LR}\mathbf{A}_{LR})\mathbf{x}_{LR_{0|t}}. \tag{16}$$

The second stage has the usual DDNM null space rectification given by Eq. (9).

In the supplementary material, we explore text-guided super-resolution using Stable diffusion [66]. It is not straightforward to impose null-space consistency on the intermediate estimates in the Stable Diffusion model similar to DDNM, as the diffusion process happens in the latent space. We show in the supplementary that this does not lead to desirable solutions.

3.2. CLIP guided Image Super-resolution

We incorporate CLIP guidance into DDNM-based image super-resolution. For a given image $\mathbf{x}_0$ and text prompt c, we define an energy function $E(c, \mathbf{x}_0)$ which measures similarity between the given image and text prompt using CLIP model, through cosine similarity between the CLIP image embeddings and CLIP text embeddings. At each step t, we obtain a clean estimate $\mathbf{x}_{0|t}$ using Eq. (6), and compute the gradient $\nabla_{\mathbf{x}_t} E(c, \mathbf{x}_{0|t})$. We rectify $\mathbf{x}_{0|t}$ to satisfy null space consistency using Eq. (9) to obtain $\hat{\mathbf{x}}_{0|t}$ and use it to compute an intermediate estimate of the previous step:

$$\hat{\mathbf{x}}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t}). \tag{17}$$

This intermediate previous step is then modified to incorporate CLIP guidance as

$$\mathbf{x}_{t-1} \leftarrow \hat{\mathbf{x}}_{t-1} - \rho_t \nabla_{\mathbf{x}_t} E(c, \mathbf{x}_{0|t}). \tag{18}$$

While one could also define an energy function combining CLIP guidance with reconstruction guidance or pseudo-inverse guidance, we observe that this creates a trade-off between the two objectives of minimizing reconstruction loss and maximizing similarity with text, and does not provide satisfactory results.

4. Experimental Evaluation

We perform experiments on extreme super-resolution of images with large super-resolution factors $\times 8$, $\times 16$, as this problem is severely ill-posed and allows exploration of a larger solution space, and is therefore an ideal setting to test our method on exploring diverse solutions. In contrast, input imposes stronger constraints on the solutions for super-resolution at smaller scale factors, limiting their diversity and explorability. We generate low-resolution images using bicubic downsampling and compute the pseudoinverse operator $\mathbf{A}^\dagger$ using SVD following [38, 81]. To evaluate consistency between the generated result and the input text prompt, we use CLIP score [61] using the ViT-B/16 CLIP model. For super resolutions with large factors, PSNR/SSIM which measure consistency with ground truth are not effective metrics to measure reconstruction performance, as multiple solutions can lead to the same low-resolution image. We instead evaluate consistency by calculating PSNR between the input LR image and the downsampled version of the solution, as also used in recent challenges for learning super-resolution space [24, 47]. We evaluate the reconstruction quality in terms of NIQE score [54]. We also conduct a user study to evaluate how the users rate the plausibility of reconstruction and semantic consistency with the text prompt.

We perform experiments on face images from CelebA-HQ dataset [35] (a subset of 200 images) and open domain images with captions from NoCaps dataset [1] (a subset of 100 images). For face image super-resolution, we manually provide different text prompts with varying personal attributes such as age, gender, smile, glasses, and curly hair. All our experiments are performed for an output resolution of 256×256. For text-guided restoration with *T2I* models, we use the open-source versions of Imagen Deep-Floyd IF [16], unCLIP Karlo-unCLIP [40], and Stable Diffusion v1.4 [66]. We describe the detailed settings and hyper-parameters for each method in the supplementary material.

We first compare the performance of proposed baselines against existing restoration methods, vanilla DDNM and DPS using unconditional diffusion models trained on CelebA-HQ faces [35] and Imagenet [17]. For this experiment, we provide a neutral prompt for text-based methods 'a high-resolution photograph of a face' for face images, and utilize images and captions from the NoCaps dataset for open domain images. The results are summarized in Tab. 1. Among the zero-shot diffusion-based image restoration methods, DDNM achieves the best LR PSNR owing to exact consistency imposed by projection operation, where DPS achieves visually sharper results, which is reflected in the lower NIQE scores. The text-based reconstruction

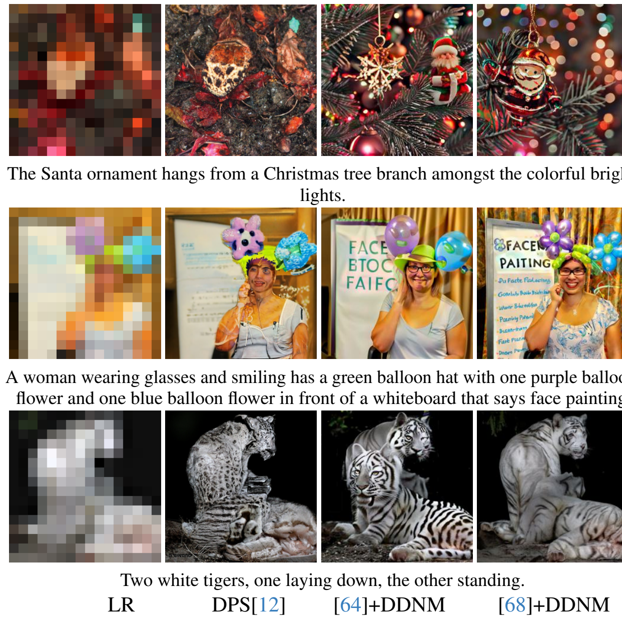

Figure 2. Visual comparison of $16\times$ SR on open domain images.

baselines from Sec. 3 achieve comparable performance to specialized models trained on faces on neutral prompts. Further, on open-domain images from the NoCaps dataset, we observe significantly better image quality in terms of NIQE score, and semantic matching as measured by CLIP score in comparison to DDNM and DPS using diffusion model trained on Imagenet. Further, we note that using CLIP guidance along with DDNM significantly improves both semantic matching with text as well as NIQE, with a reduction in LR consistency when compared to vanilla DDNM which yields rather blurred results. All the methods still maintain a good consistency with the low-resolution measurement, with LR PSNR > 45 dB. In addition to the baselines in Tab. 1, we also experimented with the baselines of [9, 31], we did not include these results as these methods do not achieve the desired level of LR consistency. Among these [31] is trained for the task of extreme super-resolution, yet, we find that it cannot handle very low-resolution inputs well.

Fig. 2 qualitatively compares the super-resolution performance of vanilla DPS with *T2I* -DDNM approaches using unCLIP and Imagen. On this test set, DDNM using an unconditional diffusion model trained on Imagenet produces blurry results. While DPS recovers sharp images in where the image content is simpler, Figs. 2 and 1 show that it struggles with complex scene content. On the other hand, the use of powerful *T2I* models in zero-shot restoration can recover data consistent solutions matching complex text prompts.

***T2I* −ΠGDM and *T2I* -DPS** Among the *T2I* model based approaches, Imagen [68]+ΠGDM has the least CLIP score in Tab. 1. This is because this method is evaluated without

Dataset	SR	Metric	DPS	DDNM	[68]+DDNM	[64]+DDNM	CLIP guided	[68]+ΠGDM
Faces	8×	LR PSNR(dB)(↑)	50.42	75.40	51.68	67.02	50.16	51.08
		NIQE(↓)	5.59	8.41	6.17	5.54	6.12	6.86
	16×	LR PSNR(dB)(↑)	51.98	80.91	51.86	66.30	51.79	52.02
		NIQE(↓)	5.54	9.77	5.94	5.43	6.38	6.98
Nocaps	8×	LR PSNR(dB)(↑)	47.01	72.94	50.34	66.33	47.86	48.75
		NIQE(↓)	9.66	10.27	4.62	4.88	5.37	5.10
		CLIP(↑)	0.2592	0.2326	0.3102	0.3344	0.2564	0.2811
	16×	LR PSNR (dB)(↑)	48.07	78.42	53.05	70.01	50.97	49.67
		NIQE(↓)	4.81	13.24	4.72	5.21	5.71	5.33
		CLIP(↑)	0.2418	0.2162	0.3037	0.3381	0.2517	0.2788

Table 1. Quantitative evaluation of baselines DPS[12],[81], and Imagen[68]+DDNM, [64], CLIP guided DDNM, Imagen[68]+ΠGDM.

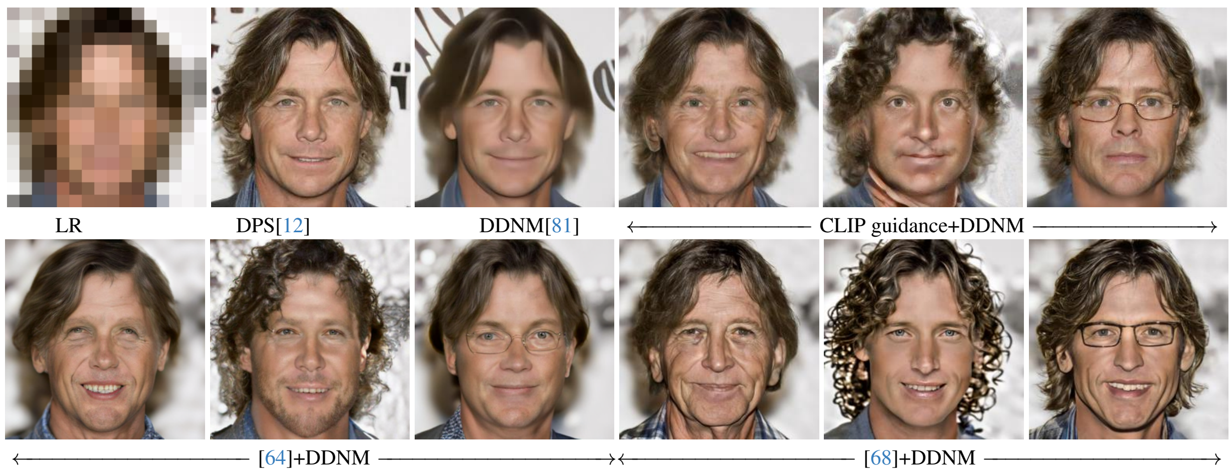

Figure 3. Exploring solution for 16× SR on a face image for the prompts: 'Elderly smiling man', ' Man with curly hair', 'Man with glasses',

classifier-free guidance (CFG) [26], whereas DDNM-based methods included CFG. It is known that including CFG in *T2I* models improves adherence to text. However, we find that it conflicts with gradient-based measurement guidance, reducing data consistency in *T2I* −**Π**GDM and *T2I* -DPS. The use of CFG requires a lower stepsize parameter, and as seen in Tab. 2, the use of the same number of diffusion steps drastically decreases LR-PSNR while improving CLIP score on the NoCaps dataset. We investigate this further by evaluating [68]+ΠGDM with varying numbers of reverse diffusion steps on a subset of 25 CelebAHQ images for the prompt 'a photograph of a woman with curly hair'. While increasing the number of diffusion steps improves LR PSNR, it also reduces text adherence. We see that improving LR PSNR always does not lead to desired results in Fig. 4. In the supplementary material Sec. 8.1, we include a detailed study and include results of similar experiments with Imagen [68]+DPS. We observe that Imagen−DPS does not provide the desired level of LR consistency even without classifier-free guidance. When classifier-free guidance is included, text adherence improves, with a significant drop in LR PSNR.

Exploring solutions through text Figs. 1 and 3 provide qualitative comparisons of the proposed text-based baselines with DPS [12] and [81] on face images. While vanilla DPS and DDNM using a diffusion model trained on faces achieve realistic and data-consistent solutions, they offer little scope for exploration and produce solutions with limited diversity. On the other hand, the proposed baselines can recover images with great diversity in attributes such as curly hair, glasses, expression, and age. As the ill-posedness of the recovery problem becomes more severe at high SR factors (×32), it is possible to recover a wide variety of outputs with challenging attributes in age, race, and appearance. We show more examples in the supplementary material.

Limitation of *T2I* -DDNM While *T2I* -DDNM methods do not face a trade-off between text adherence and measurement consistency, they can still result in unrealistic images

	Step size	Steps	CFG	LR PSNR (dB)	CLIP score
Nocaps	1.0	100	✗	49.67	0.2788
	0.5	100	✓	24.31	0.2923
Faces	1.0	100	✗	48.705	0.2560
	0.5	100	✓	30.675	0.3011
	0.5	200	✓	40.025	0.2742
	0.5	300	✓	42.431	0.2773
	0.5	500	✓	43.093	0.2695

Table 2. Effect of CFG, step size, and number of steps on consistency and text adherence in *T2I* -ΠGDM .

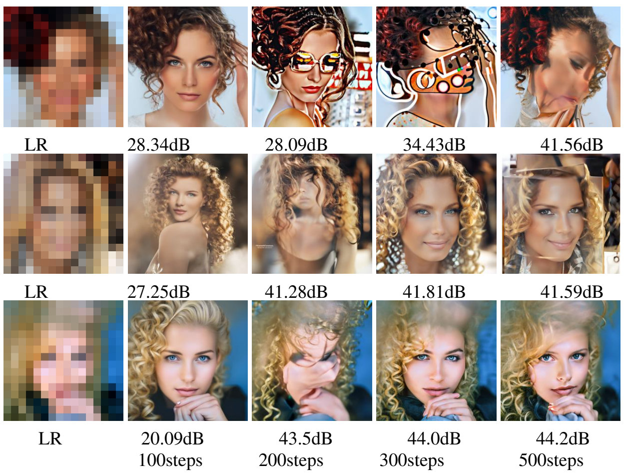

Figure 4. Effect of classifier-free guidance and stepsize in Imagen-ΠGDM.

Figure 5. $\times 16$ SR results with (bottom) and without (top) averaging trick with λ=0.4, and the text prompt 'a high-res photo of a cat'.

that are both consistent with the text as well as measurement. This is because any image hallucinated by the *T2I* model can still be made consistent with the measurement through null-space component rectification. We sometimes encounter this problem in unCLIP[64]-DDNM, when image embedding imagined by prior does not structurally align with the observation. This can be mitigated to an extent by modifying the image embedding which conditions the diffusion model.

	Imagen-DDNM		unCLIP-DDNM		CLIP guidance	
	Faces	nocaps	Faces	nocaps	Faces	nocaps
Text-Similarity	90.89%	92.88%	73.20%	54.42%	60.64%	25.38%
Photo-realism	69.35%	83.07%	30.89%	35.57%	30.77%	10.38%

Table 3. Results of user survey on text guided super-resolution

We introduce an embeddings averaging trick by considering a convex combination of embedding provided by the prior, and CLIP image embedding of the pseudo-inverse solution. The extent of this averaging can be controlled by an additional hyperparameter λ which determines the weight of pseudoinverse embedding. This embeddings averaging trick can improve the structural consistency of the solution with the input observation, as seen in Fig. 5.

User Study We performed a user study to evaluate the realism and semantic matching with text prompts on the results of Imagen-DDNM, unCLIP DDNM, and CLIP-guided DDNM using an online survey platform. The survey included 50 reconstructions for each method for the task of $\times 16$ super-resolution along with the corresponding text prompts. The LR images were generated using 30 face images from CelebA HQ, and 20 open-domain images from the NoCaps dataset. The users were asked to evaluate separately whether each reconstruction semantically matches with input text prompt and whether the solution appears photo-realistic. The results of this survey are found in Tab. 3. Both Imagen DDNM and unCLIP DDNM score better in terms of user preference in comparison to CLIP-guided recovery. This is to be expected, as both [40] and [16] are powerful *T2I* models trained over webscale data. Among the three methods, Imagen DDNM has the best user preference in terms of both semantic matching with text as well as perceived realism of the recovered solution. This human evaluation is in contrast with the quantitative evaluation in Tab. 1 where unCLIP achieves higher semantic matching with text in terms of CLIP score. As unCLIP was trained to invert CLIP embeddings, it possibly produces images with higher CLIP scores.

5. Related Work

Diverse Solutions to Image Super-resolution Deep networks have become popular tool for image super-resolution in the past decade[9, 39, 53, 78], where many state of the art methods [6, 31, 79, 80] employ supervised training to recover a single solution. Deep learning based solutions which allow sampling multiple solutions to the ill-posed SR problem also exist, [2, 5, 9, 11, 33, 38, 46, 46, 53, 55, 74, 81]. These methods utilize conditional or unconditional generative models to sample solutions. Among these conditional generative model based approaches [2, 32, 42, 46, 47, 49, 60, 69] are trained for the specific super-resolution task. Zero-shot approaches[11, 38, 53, 74, 81] on the other hand, utilize image generative models directly for image recovery.

Explorable Image Super-resolution A few prior works attempt to explore solutions space using graphical inputs [2] or semantic maps [5]. However, they are still restricted to specific classes e.g. faces, or trained for specific degradation, e.g. specific super-resolution factors. A recent work [50] also combines text features into super-resolution network architectures using attention, and trains separate models for text-guided image super-resolution in an end-to-end manner for each dataset and super-resolution factor. Yet, this approach cannot handle open domain images and arbitrary super-resolution factors. To the best of our knowledge, there is no existing method that allows zero-shot exploration of solutions space for different restoration tasks on open-domain images through text.

Diffusion Models for Image Super-resolution One could utilize diffusion models for image super-resolution and other restoration tasks either by training a conditional diffusion model for specific tasks [42, 69, 82, 88], or by leveraging diffusion models for zero-shot image recovery [9–11, 29, 34, 37, 38, 48, 56, 81]. We are concerned with the latter variety, which exploit the knowledge of degradation operator to modify the sampling process. Earlier works [29, 34] adopt Langevin dynamics for linear inverse problems and incorporate measurement guidance through the gradient of the least-squares data fidelity term. [9, 11] alternate between a standard reverse diffusion step and a projection step promoting measurement consistency. Recent works utilize an estimate of clean sample at each reverse step to modify the sampling process via a consistency enforcing projection operation [38, 48, 81] or through guidance through the gradient of the least-squares data fidelity term [12], or least squares measurement guidance [74] or both [10]. While projection-based approaches are faster and do not need to backpropagate through diffusion model weights, they are restricted to inverse problems where a pseudo-inverse or its approximation can be computed. On the other hand, gradient-based measurement guidance can be applied for any inverse problems, or even arbitrary guidance [3, 84], yet it is more expensive as it requires back-propagation through the diffusion model weights at each iteration. More recently, [52] adopt diffusion models in a regularization by denoising (RED) framework, and [91] demonstrate their utility for plug-and-play image restoration as an effective alternative to the standard Gaussian denoisers.

***T2I* Generative Models** Starting from [51] many works proposed different methods to generate images from text prompts. Initial works trained RNNs [51] and GANs [41, 65, 83, 85, 85–87, 89, 90] using attention on smaller captioned image datasets [43, 76]. Recent developments in image generation [18, 21, 57], contrastive learning[7] and large-scale training on massive internet-scale datasets of paired text prompts and images [70, 71] accelerated research in vision-language learning [30, 58, 61, 63, 64, 68]. Many recent works train text-to-image (T2I) models directly on large scale datasets using autoregressive transformers [19, 22, 63] or diffusion based models[58, 64, 68]. Some of these T2I models perform diffusion in a low dimensional latent space [4, 20, 25, 28, 66, 75], or in a down-sampled pixel space [64, 68] for computational efficiency. An alternative paradigm is employed by [13–15, 23, 44, 45, 59] using text-image encoder CLIP [61] approaches to guide pretrained generative models [18, 21] towards the input text prompt.

6. Discussion, Limitations, and Conclusions

In this paper, we introduced the challenging task of zero-shot open-domain extreme super-resolution for different scale factors guided by text prompts. We explored two approaches to deal with this challenge- utilizing pretrained diffusion-based *T2I* models for zero-shot recovery and by guiding an image diffusion model with CLIP for zero-shot diffusion-based restoration. We showed that the proposed methods improve adherence to input text prompts while maintaining consistency with the observation. We demonstrated significantly improved diversity in solutions using the proposed methods. Among these methods, CLIP guidance is naturally outperformed by the more powerful *T2I* diffusion model-based methods. Further, we found that gradient-based reconstruction guidance could be in trade-off with text adherence. Moreover, the generated results are not always realistic, and it can require several attempts to realize the desired output, as also observed in text condition image generation [36]. It must be noted that not every text prompt is meaningful for every observation. When certain patterns or objects indicated by text cannot be present in the image, the corresponding objects or patterns cannot be recovered without severe artifacts or unrealistic images. In this case, it is not the failure of the approach or the model, rather it can help the users determine the plausibility of a solution. In view of this, the evaluation of text-guided restoration is highly subjective, and any quantitative evaluation in terms of image quality metrics is meaningful only when the input text prompts are well aligned and plausible for the given degraded measurement. The performance of all the proposed methods depends on and is limited by the generative capabilities of the pre-trained generative model. The method inherits the biases of the data used to train the *T2I* model. Finally, our work opens up a promising direction of developing efficient user-guided tools for text-based exploration of image recovery.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 5

[2] Yuval Bahat and Tomer Michaeli. Explorable super resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2716–2725, 2020. 1, 3, 7, 8

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. *arXiv preprint arXiv:2302.07121*, 2023. 3, 8

[4] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *17th European Conference on Computer Vision*, pages 170–188. Springer, 2022. 8

[5] Marcel C Buhler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 7, 8

[6] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 1, 7

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 8

[8] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *14th Asian Conference on Computer Vision*, pages 605–621. Springer, 2018. 16

[9] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 5, 7, 8

[10] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, 2022. 8

[11] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. 7, 8

[12] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 5, 6, 8, 14, 16, 17, 18, 19, 20, 21, 22, 23

[13] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18270–18279, 2022. 8

[14] Katherine Crowson. CLIP guided diffusion HQ 256x256. Colab Notebook, 2021.

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

[16] deepfloyd.ai. Deepfloyd if: A modular cascaded diffusion model, 2023. `https://github.com/deep-floyd/IF`. 5, 7, 16

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 5

[18] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 8

[19] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[20] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021. 8

[21] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 8

[22] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *17th European Conference on Computer Vision*, pages 89–106. Springer, 2022. 8

[23] Federico A Galatolo, Mario GCA Cimino, and Gigliola Vaglini. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv preprint arXiv:2102.01645*, 2021. 8

[24] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 951–967, 2022. 5

[25] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 8

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information*

Processing Systems, pages 6840–6851. Curran Associates, Inc., 2020. 2

[28] Minghui Hu, Yujie Wang, Tat-Jen Cham, Jianfei Yang, and P.N. Suganthan. Global context with discrete diffusion in vector quantised modelling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11502–11511, 2022. 8

[29] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. In *Advances in Neural Information Processing Systems*, pages 14938–14954. Curran Associates, Inc., 2021. 8

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 8

[31] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 424–425, 2020. 1, 5, 7

[32] Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Tackling the ill-posedness of super-resolution through adaptive target generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16236–16245, 2021. 7

[33] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Srflow-da: Super-resolution using normalizing flow with deep convolutional block. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 364–372, 2021. 7

[34] Zahra Kadkhodaie and Eero P Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Advances in Neural Information Processing Systems*, 2021. 8

[35] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[36] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection, 2023. 8

[37] Bahjat Kawar, Gregory Vaksman, and Michael Elad. SNIPS: Solving noisy inverse problems stochastically. In *Advances in Neural Information Processing Systems*, 2021. 8

[38] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 1, 5, 7, 8

[39] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 7

[40] Donghoon Lee, Jiseob Kim, Jisu Choi, Jongmin Kim, Minwoo Byeon, Woonhyuk Baek, and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. `https://github.com/kakaobrain/karlo`, 2022. 5, 7, 16

[41] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[42] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1, 7, 8

[43] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 8

[44] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv preprint arXiv:2112.01573*, 2021. 8

[45] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 289–299, 2023. 8

[46] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *Proceedings of 16th European Conference on Computer Vision (ECCV)*, pages 715–732. Springer, 2020. 1, 7

[47] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 596–612, 2021. 5, 7

[48] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 8

[49] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Kangwook Kim, Younggeun Kim, Jae-young Lee, Zechao Li, Jinshan Pan, Dongseok Shim, Ki-Ung Song, Jinhui Tang, Cong Wang, and Zhihao Zhao. Ntire 2022 challenge on learning the super-resolution space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 786–797, 2022. 7

[50] Chenxi Ma, Bo Yan, Qing Lin, Weimin Tan, and Siming Chen. Rethinking super-resolution as text-guided details generation. *arXiv preprint arXiv:2207.06604*, 2022. 8

[51] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 8

[52] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023. 8

[53] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 7, 8

[54] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 5

[55] Antonio Montanaro, Diego Valsesia, and Enrico Magli. Exploring the solution space of linear inverse problems with gan latent geometry. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1381–1385, 2022. 7

[56] Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M. Patel, and Tim K. Marks. Steered diffusion: A generalized framework for plug-and-play conditional image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20850–20860, 2023. 8

[57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 8

[58] 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*, pages 16784–16804. PMLR, 2022. 8

[59] Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. In *17th European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, 2022. Springer-Verlag. 8

[60] Shichong Peng and Ke Li. Generating unobserved alternatives: A case study through super-resolution and decompression. *OpenReview*, 2020. 7

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5, 8

[62] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 8

[64] 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, 3, 4, 5, 6, 7, 8, 13, 16, 17, 18, 19, 22, 23

[65] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 8

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 4, 5, 8, 14

[67] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 14

[68] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3, 4, 5, 6, 8, 13, 16, 17, 18, 19, 22, 23

[69] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. 8

[70] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*. Jülich Supercomputing Center, 2021. 8

[71] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 8

[72] Johannes Schwab, Stephan Antholzer, and Markus Haltmeier. Deep null space learning for inverse problems: convergence analysis and rates. *Inverse Problems*, 35, 2018. 3

[73] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3

[74] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 1, 3, 7, 8, 16

[75] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 8

[76] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 8

[77] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1

[78] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 7

[79] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 7

[80] Yinhuai Wang, Yujie Hu, and Jian Zhang. Panini-net: Gan prior based degradation-aware feature interpolation for face restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2576–2584, 2022. 1, 3, 7

[81] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 5, 6, 7, 8, 16, 17, 20, 21, 22, 23

[82] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G. Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16293–16303, 2022. 8

[83] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 8

[84] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 8, 17

[85] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916, 2017. 8

[86] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[87] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[88] Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, and Matthias Grundmann. Towards authentic face restoration with iterative diffusion models and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7312–7322, 2023. 8

[89] Junchen Zhu, Lianli Gao, Jingkuan Song, Yuan-Fang Li, Feng Zheng, Xuelong Li, and Heng Tao Shen. Label-guided generative adversarial network for realistic image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 8

[90] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 8

[91] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (NTIRE)*, 2023. 8