Text-guided Explorable Image Super-resolution

Supplementary Material

This supplementary material is organized as follows: Sec. 7 provides the algorithms for text based super-resolution introduced in Sec. 3. We conduct further experiments demonstrating the trade-off between text adherence and LR PSNR in gradient-based guidance methods in Sec. 8.1, and show the results of our experiments with other approaches. We provide detailed experimental settings in Sec. 9 and additional qualitative results in Sec. 10.

7. Algorithms for Text based Super-resolution

We provide the algorithms for text-based super-resolution methods described in Sec. 3. In the following p_1 and p_2 denote reverse diffusion processes in the text-conditioned downsampled stage and the subsequent super-resolution stage in *T2I* diffusion models. For Imagen [68], $\mathbf{c}_1$ corresponds to text embeddings from text encoder, and $\mathbf{c}_2$ contains $\mathbf{x}_{LR}$, in addition to text embeddings. For unCLIP [64] $\mathbf{c}_1$ 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.

Algorithm 1 *T2I* -DDNM sampling process

$\mathbf{x}_{LR_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{c}_1)\sqrt{1-\bar{\alpha}_t}\right)$
  $\hat{\mathbf{x}}_{LR_{0|t}} = \mathbf{A}^{\dagger}{}_{LR}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}{}_{LR}\mathbf{A}_{LR})\mathbf{x}_{LR_{0|t}}$
  $\mathbf{x}_{LR_{t-1}} \sim p_1(\mathbf{x}_{LR_{t-1}}|\mathbf{x}_{LR_t}, \hat{\mathbf{x}}_{LR_{0|t}})$
end for
$\mathbf{x}_{LR} \leftarrow \mathbf{x}_{LR_0}$
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\zeta}_{\boldsymbol{\theta}}(\mathbf{x}_t, t|\mathbf{c}_2)\sqrt{1-\bar{\alpha}_t}\right)$
  $\hat{\mathbf{x}}_{0|t} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_{0|t}$
  $\mathbf{x}_{t-1} \sim p_2(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$
end for
return $\mathbf{x}_0$

8. Further Analysis of Text based SR Methods

8.1. Gradient based guidance in T2I models

We observed in Tab. 2 of our paper that gradient-based measurement guidance could be in trade-off with the text adherence. We found that increasing the number of diffusion steps to 1000 still did not improve LR consistency to the desired level (> 45 dB) when we include classifier-free guidance for $\times 16$ super-resolution of faces. We also performed experiments with Imagen+DPS, on a subset of 25 face images, for the text prompt 'a photograph of a woman with curly hair', the results are provided in Tab. 4. While

Algorithm 2 *T2I* -DPS sampling process

$\mathbf{x}_{LR_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{c}_1)\sqrt{1-\bar{\alpha}_t}\right)$
  $\mathbf{x}'_{LR_{t-1}} \sim p_1(\mathbf{x}_{LR_{t-1}}|\mathbf{x}_{LR_t}, \mathbf{x}_{LR_{0|t}})$
  $\mathbf{x}_{LR_{t-1}} = \mathbf{x}'_{LR_{t-1}} - \rho_t \nabla_{\mathbf{x}_{LR_t}} \|\mathbf{y} - \mathbf{A}_{LR}\mathbf{x}_{LR_{0|t}})\|_2^2$
end for
$\mathbf{x}_{LR} \leftarrow \mathbf{x}_{LR_0}$
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\zeta}_{\boldsymbol{\theta}}(\mathbf{x}_t, t|\mathbf{c}_2)\sqrt{1-\bar{\alpha}_t}\right)$
  $\mathbf{x}'_{t-1} \sim p_2(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_{0|t})$
  $\mathbf{x}_{t-1} = \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} \|\mathbf{y} - \mathbf{A}(\mathbf{x}_{0|t})\|_2^2$
end for
return $\mathbf{x}_0$

Algorithm 3 *T2I* -ΠGDM sampling process

$\mathbf{x}_{LR_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{LR_{0|t}} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_{LR_t} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_{LR_t}, t|\mathbf{c}_1)\sqrt{1-\bar{\alpha}_t}\right)$
  $\mathbf{x}'_{LR_{t-1}} \sim p_1(\mathbf{x}_{LR_{t-1}}|\mathbf{x}_{LR_t}, \mathbf{x}_{LR_{0|t}})$
  $\mathbf{x}_{LR_{t-1}} = \mathbf{x}'_{LR_{t-1}} - \rho_t \nabla_{\mathbf{x}_{LR_t}} \|\mathbf{A}^{\dagger}{}_{LR}\mathbf{y} - \mathbf{A}^{\dagger}{}_{LR}\mathbf{A}_{LR}(\mathbf{x}_{LR_{0|t}})\|_2^2$
end for
$\mathbf{x}_{LR} \leftarrow \mathbf{x}_{LR_0}$
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\zeta}_{\boldsymbol{\theta}}(\mathbf{x}_t, t|\mathbf{c}_2)\sqrt{1-\bar{\alpha}_t}\right)$
  $\mathbf{x}'_{t-1} \sim p_2(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_{0|t})$
  $\mathbf{x}_{t-1} = \mathbf{x}'_{t-1} - \rho_t \nabla_{\mathbf{x}_t} \|\mathbf{A}^{\dagger}\mathbf{y} - \mathbf{A}^{\dagger}\mathbf{A}(\mathbf{x}_{0|t})\|_2^2$
end for
return $\mathbf{x}_0$

Algorithm 4 CLIP guidance with DDNM

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
  $\mathbf{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\sqrt{1-\bar{\alpha}_t}\right)$
  $\hat{\mathbf{x}}_{0|t} = \mathbf{A}^{\dagger}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})\mathbf{x}_{0|t}$
  $\hat{\mathbf{x}}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_{0|t})$
  $\mathbf{x}_{t-1} \leftarrow \hat{\mathbf{x}}_{t-1} - \rho_t \nabla_{\mathbf{x}_t} E(c, \mathbf{x}_{0|t})$
end for
return $\mathbf{x}_0$

	Step size	Steps	CFG	LR PSNR (dB)	CLIP score
Faces16×	1.0	500	✗	39.71	0.2433
	1.0	800	✗	39.83	0.2409
	1.0	1000	✗	40.94	0.2353
	0.5	500	✓	28.46	0.3130
	0.5	1000	✓	36.51	0.2798

Table 4. Effect of CFG, step size, and number of steps on consistency and text adherence in $T2I - DPS$.

Imagen-ΠGDM achieved the desired LR consistency without classifier-free guidance, we were unable to reach the desired level of LR consistency with Imagen−DPS, even without classifier-free guidance. When 1000 diffusion steps are used, we obtain an LR PSNR of around 40.9 dB. When classifier-free guidance is included, text adherence improves, with a significant drop in LR PSNR. Fig. 6 shows sample results of this experiment for both Imagen−ΠGDM and Imagen−DPS. Imagen−ΠGDM obtains a higher PSNR than Imagen−DPS. Figs. 4 and 6 indicate that these methods can also achieve consistency with text at the cost of lower LR PSNR. We did not find solutions satisfying both the constraints in 3 independent runs. Fig. 7 shows sample reconstructions of Imagen−ΠGDM and Imagen−DPS when there is no classifier-free guidance for the same text prompt. We find that Imagen−ΠGDM achieves significantly higher LR PSNR. While both methods achieve photo-realistic reconstructions, they fail to adhere to text prompts without classifier-free guidance.

Influence of image embedding in unCLIP-DDNM In our paper, we introduced an 'embeddings averaging trick' and showed that this can reduce the mismatch between image embedding and the observation. We now illustrate the impact of image embedding on unCLIP-DDNM solutions with an example in Fig. 8. The first row shows the LR image, and the images generated by unCLIP through the usual DDPM sampling process with no text prompt, i.e. decoder is conditioned only on image embeddings generated by the prior with null-text. We fix the random seed throughout this experiment. This embedding produces an image of a T-shirt. Even starting the reverse diffusion process at an earlier step initializing with a pseudo-inverse solution retains the same concept. In the second row, we observe that using the prior image embedding as it is leads to a solution where the pseudoinverse is super-imposed on an image of T-shirts. On the other hand, embeddings averaging results align the concept provided by prior embedding with the pseudoinverse solution. The benefit of embeddings averaging for text-conditioned super-resolution is further illustrated in Fig. 9. Even for text prompts that are seemingly incoherent with the low-resolution measurement such as 'a photograph of a man', embeddings averaging can reduce misalignments

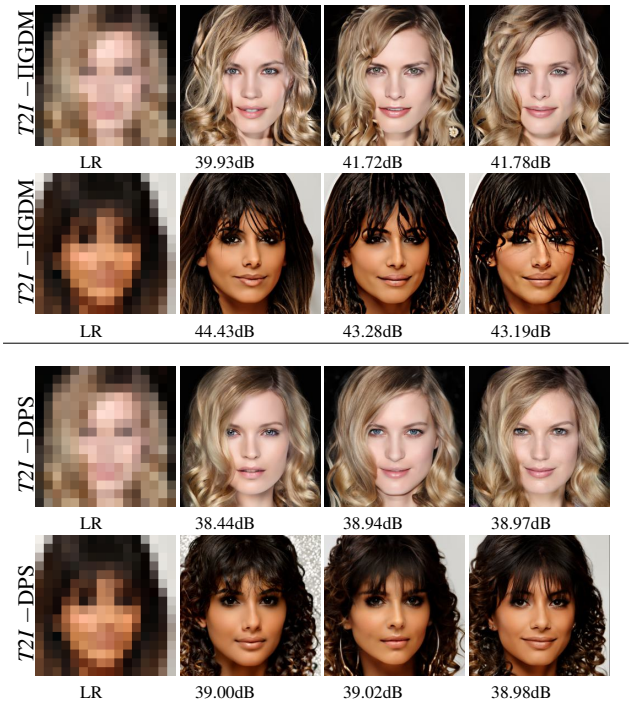

Figure 6. Results of multiple runs of Imagen-ΠGDM (top) and Imagen-DPS (bottom) with 1000 steps including classifier-free-guidance and corresponding LR PSNR values. Text prompt 'Photograph of a face of a woman with curly hair'.

and weird backgrounds and improve photorealism.

8.2. Experiments with Stable Diffusion

Using Text inputs in PSLD [67] Recently [67] extends posterior sampling for inverse problems [12] to latent diffusion models such as Stable diffusion [66]. While this method also addresses super-resolution, they consider super-resolution with small SR factors $2\times$, $3\times$, and $4\times$, and do not provide text prompts. Further, as Stablediffusion produces images at higher resolution 512×512, [67] apply the forward operator after upsampling inputs to this resolution, run posterior sampling, and then downsample images to resolution 256×256. Therefore, even $4\times$ SR has a higher resolution input. We instead experimented with a much lower input resolution of 32×32, which we interpolate to 64×64 and provide as input to PSLD for $8\times$ SR task. While the original setting of $4\times$ SR with 128×128 resolution provided satisfactory outputs, we could not find a satisfactory hyperparameter setting which results in a sharp output image with very low-resolution inputs.

Incorporating Null-space Consistency We also studied the applicability of Stable diffusion [66] for text-guided image super-resolution using DDNM. In the case of Stable Diffusion, the diffusion process happens in the latent space

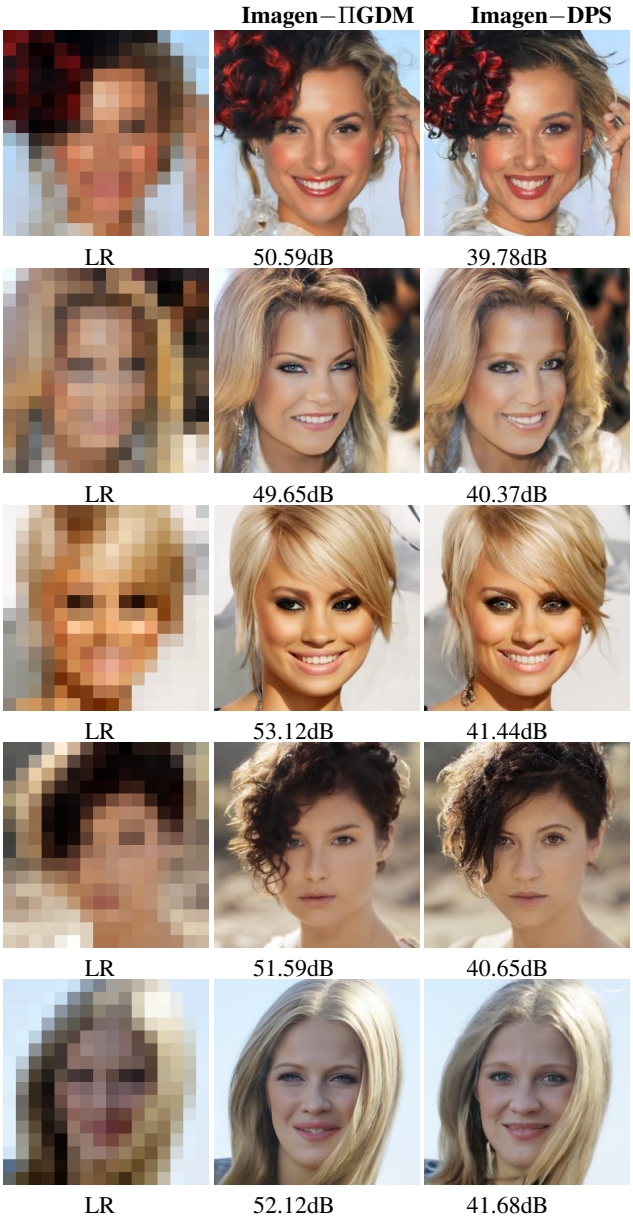

Figure 7. Example reconstructions of Imagen-ΠGDM and Imagen-DPS without classifier-free-guidance, and corresponding LR PSNR values. Text prompt 'Photograph of the face of a woman with curly hair'.

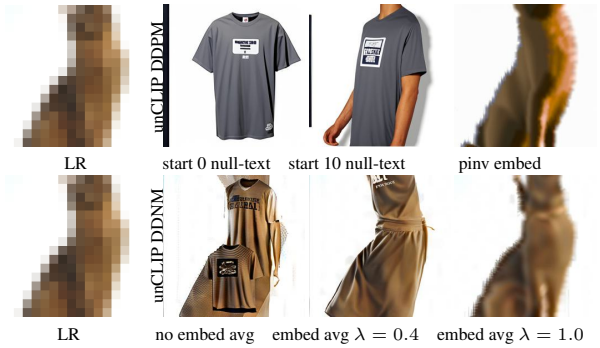

Figure 8. Effect of image embedding in unCLIP-DDNM.

'A photograph of a face of{*}' for $* \in$ {a man, a woman}

Figure 9. Embeddings averaging trick improves photorealism in difficult examples $\times 16$ face super-resolution.

of a variational auto-encoder, and it is not straightforward to adapt DDNM to this model. We attempted to enforce DDNM consistency in the text-conditioned Stable diffusion. At each step in the reverse diffusion process, we estimate the clean latent variable z_0, and decode it to image space and enforce data consistency. This data-consistent image is then encoded again to latent space to resume reverse diffusion. While this approach achieves data-consistency, we find that it is highly unstable due to the lossy nature of the variational autoencoder. As the number of inference steps increases, it results in unrealistic images with heavy artifacts. When the number of inference steps is low, the artifacts reduce, however the resulting images are blurry, see the top row of Fig. 10. On the other hand, we observed that using the interpolated low-resolution image as an initial estimate in the diffusion process can lead to a totally different image without any guidance or consistency enforcement in the in-

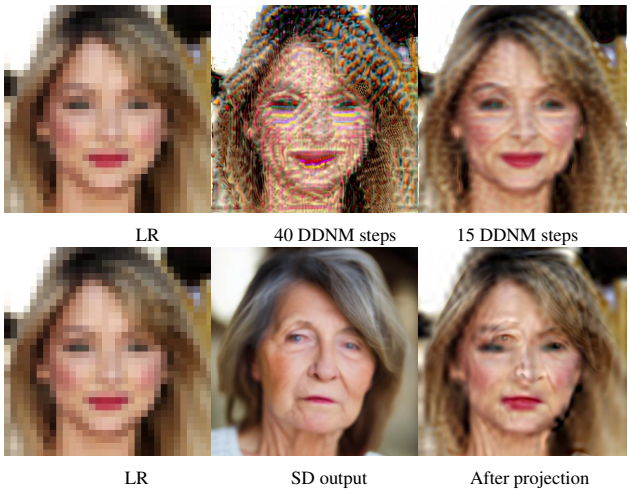

Figure 10. (Top) Stable Diffusion with DDNM using VAE decoder & encoder. (Bottom) Stable Diffusion results without DDNM in reverse diffusion. Results for the text prompt 'a photograph of a face of an elderly woman' for $\times 8$ SR

Metric	DPS	[68]+DDNM$^+$	[68]+ΠGDM	[68]+DPS
LR PSNR	30.05	31.40	31.52	31.33
NIQE	5.92	7.09	6.07	5.57
CLIP score	0.290	0.297	0.295	0.290

Table 5. Results of noisy super-resolution on CelebA HQ for $\times 8$ SR.

termediate steps. While the output of the Stable Diffusion model in this case is well-aligned with the text prompt, it is not consistent with the measurement. As a result, the corresponding null-space contents are not aligned with the pseudo-inverse solution. An example is illustrated in the bottom row of Fig. 10, where the null space projection adds high-frequency details of the elderly woman onto the pseudo inverse reconstruction. In contrast, using the pixel domain *T2I* models allows us to perform DDNM for text-guided reverse diffusion in the (down-sampled) pixel space, which easily generates data-consistent images that are aligned with text.

8.3. Noisy SR and Real-world SR:

We provide the results of $8\times$SR with Gaussian noise (σ=0.05) for the neutral prompt on CelebA HQ 'a high-resolution photograph of a face'. The methods $T2I$-DPS/ΠGDM can directly handle noisy inputs. For $T2I$-DDNM we use DDNM$^+$ following [81] and report LR PSNR, NIQE andCLIP score in Tab. 5. Zero-shot diffusion-based methods [12, 74, 81] require the knowledge of the degradation operator which is not available directly in real-world SR, but needs to be estimated. When the real LR images are well approximated by our downsampling operator we can still recover good quality solutions, see Fig. 11 for results on

Figure 11. Real world super-resolution on sample image from [8].

real LR images from the QMUL tiny face dataset [8] using [68]+DDNM$^+$.

9. Experiment Settings

***T2I* +DDNM** We investigate the use of Imagen [68] and unCLIP [64]with DDNM sampling face super-resolution and open-domain image super-resolution. We use the open-source versions of Imagen Deep-Floyd IF [16] and un-CLIP Karlo-unCLIP [40]. We use these models for super-resolution without further fine-tuning. For Imagen+DDNM, we use 200 reverse diffusion steps at the resolution of 64×64, followed by 50 reverse diffusion steps at resolution 256×256. Instead of starting at random noise, we start reverse diffusion at timestamp t_{stop} as 850, where noisy estimate at t_{stop} is obtained by adding noise of suitable variance to the pseudo-inverse solution. We utilize classifier-free guidance with guidance scales of 7.0 and 4.0 respectively in the first and second stages of the Imagen Deep Floyd model. When compared to CFG using null-text, we found use of negative prompts {ugly, disfigured, broken, caricature} improves the quality of reconstruction. The guidance scale controls the strength of the attribute indicated by text and can be varied to obtain attributes of varying intensity. We found that not including classifier-free guidance can still produce realistic results, but the text adherence reduces significantly.

For experiments with unCLIP+DDNM, we start reverse diffusion at timestamp t_{stop} as 850 out of the total 1000 steps and use fewer steps between $[1, t_{stop}]$ in both the reverse diffusion process (80 for text conditioned decoding and 7 for super-resolution). The noisy estimate at t_{stop} is obtained by adding noise of suitable variance to the pseudo-inverse solution. For obtaining the image embeddings we perform 25 reverse diffusion steps using the prior model. We utilize an optional embeddings average trick when there is consistent misalignment between measurement and embeddings. For this, we utilize the CLIP image encoder used to train the unCLIP model. We utilize classifier-free guidance both for the prior and the unCLIP decoder. The super-resolution stage is conditioned only on the output of the first stage, and therefore, there is no classifier-free guidance in the SR module. We use classifier-free guidance of scale 4.0 in the prior. We find that a high value of classifier-free guidance scale for text conditioned decoder results in images which adhere well to text, yet they are often smoothed out, and

appear unrealistic. We found lower values of classifier-free guidance between 1.0 and 3.0 leads to more photo-realistic results. Further, similar to Imagen+DDNM, we find better reconstructions when negative prompts are used for classifier-free guidance.

While we provided hyperparameters used in experiments, these are not fixed and can be varied by the user to explore solutions. Oftentimes, a high number of reverse diffusion steps (200) is not required for Imagen+DDNM, and most solutions satisfying both text and low-resolution consistency can be obtained for fewer steps($\sim$100). All experiments with *T2I* +DDNM were run on a single GPU office machine, containing an NVIDIA RTX 3090 GPU.

CLIP guidance+DDNM We investigate the use of CLIP guidance along with DDNM for both face super-resolution and open-domain image super-resolution. For the task of face super-resolution, we use a diffusion model trained at a resolution of 256×256 on CelebAMask-HQ dataset from available from SDEdit repository. For the task of open domain image super-resolution, we use an unconditional imagenet model trained at a resolution of 256×256 from Guided Diffusion. We use the ViT-B/16 CLIP model for CLIP guidance. We incorporate DDNM null-space rectification into CLIP guided generation of [84] using their publicly available code. We retain their strategy for learning rate and time-travel strategy when using imagenet model for open domain super-resolution. We use 300 ddim steps with a learning rate hyperparameter of 0.05. Instead of ℓ_2 distance used in [84], we utilize cosine distance measurement in the energy function for CLIP guidance. All experiments with CLIP guidance were run on a single GPU office machine, containing an NVIDIA RTX 3090 GPU.

10. Additional Results

10.1. Open domain Image super-resolution

We provide additional qualitative results of open domain text guided super-resolution in Figs. 12 and 13. While the solutions provided by DPS [12] are consistent with the measurements, the lack of additional information to guide the reconstruction results in unsatisfactory results while recovering complex scene content. The use of CLIP guidance with DDNM improves results over vanilla DPS, and *T2I* -DDNM improves adherence to text even further. Imagen-[68]+DDNM can nearly recover images adhering to complex prompts involving scene text, for instance, 'A man with salt and pepper hair, a tie, and glasses is sitting behind a table with a sign that says Mexico in front of him.', or expression 'A red panda is sitting on a tree squeezing its eyes shut and sticking out its tongue'. It is also better at compositionality, for instance for text prompt 'The woman is wearing a colorful kimono and carrying a floral print purple umbrella', Imagen-[68]+DDNM can recover a floral print purple umbrella, where as unCLIP[64]+DDNM recovers a purple umbrella without floral patterns, which appear instead on the woman's head and clothing.

10.2. Exploring SR solutions through text

We provide qualitative results on exploring the solutions to face super-resolution to upsampling factors $8\times$, $16\times$ and $32\times$. Figs. 14 and 15 show examples of exploring the solutions of $8\times$ super-resolution. The reconstructions of DPS [12], DDNM [81] and *T2I* -DDNM methods are compared. The results demonstrate improved diversity using *T2I* -DDNM by exploring a variety of attributes. In contrast, when compared to vanilla DPS and DDNM which exhibit limited diversity. Similarly in Figs. 16 and 17 and Fig. 18 we illustrate examples of exploring the solutions of $16\times$ and $32\times$ super-resolution respectively. We can see in Fig. 18 that severe ill-posedness of $32\times$ SR task allows a wide range of solutions with varying personal attributes such as perceived age, gender, race, accessories etc.

https://github.com/ermongroup/SDEdit
https://github.com/openai/guided-diffusion
https://github.com/vvictoryuki/FreeDoM

A man with salt and pepper hair, a tie, and glasses is sitting behind a table with a sign that says Mexico in front of him.

A large white Lufthansa plane is sitting on the runway.

A red panda is sitting on a tree squeezing its eyes shut and sticking out its tongue.

A stately grey and white house with an American flag sits among green trees.

A black Land Rover is next to a building.

A gorgeous red headed woodpeker perched on a branch with red flowers.

LR DPS[12] DDNM+CLIP guidance unCLIP[64]+DDNM Imagen[68]+DDNM

Figure 12. Visual comparison of 16× SR on open domain images.

The woman is wearing a colorful kimono and carrying a floral print purple umbrella.

A clock with a gold bezel reading ten minutes to 2-o-clock.

A man in a hat and blue striped shirt carrying wood

Two bicyclists are racing through a blurry background.

A man in a blue shirt and orange shorts ride a motorized scooter.

A smiling girl with glasses and a hat with a clock on it.

LR DPS[12] DDNM+CLIP guidance unCLIP[64]+DDNM Imagen[68]+DDNM

Figure 13. Visual comparison of 16× SR on open domain images.

Figure 14. Exploring solutions to $8\times$ face super-resolution

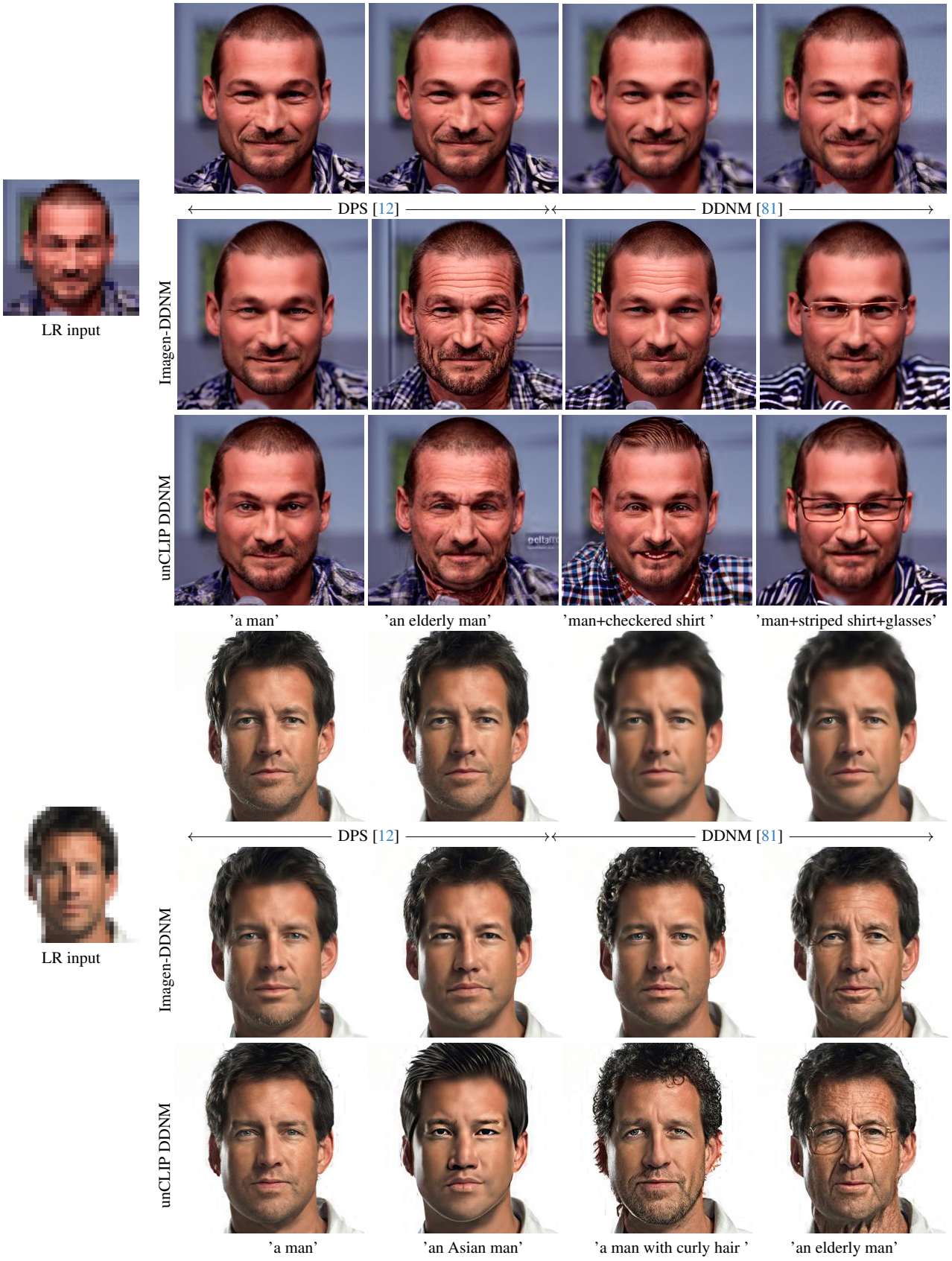

Figure 15. Exploring solutions to $8\times$ face super-resolution.

Figure 16. Exploring solutions for 16× SR of face images.

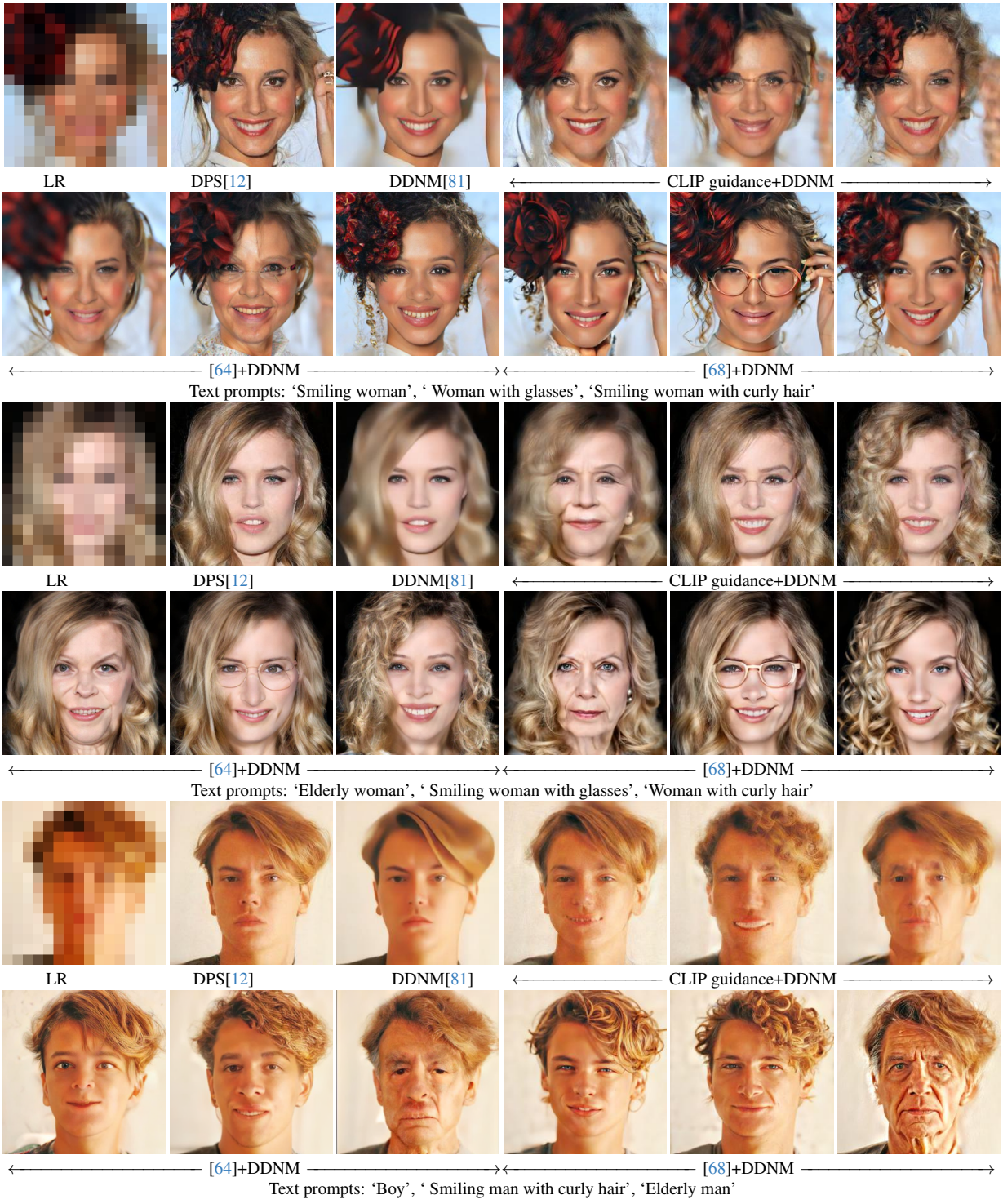

Figure 17. Exploring solutions for 16× SR of face images.

Figure 18. Exploring solutions for 32× SR of face images.