école normale supérieure paris-saclay
université PARIS-SACLAY

Ilias Sarbout

Inpainting with denoising diffusion probabilistic models

Contents

List of Figures

1 Introduction

Image completion, or image inpainting, is the process of filling in missing regions of an image with consistent content. This task is often self-supervised in deep learning because it is difficult to accurately evaluate the quality of an inpainted image numerically. As a result, several algorithms have been developed in recent years to take advantage of large image datasets and train efficient image inpainting models. However, some models, such as denoising diffusion probabilistic models (DDPMs), can perform image inpainting without being trained on known missing regions. DDPMs that have been trained on a large number of images can generate plausible images due to their a priori knowledge of the natural image domain. It is possible to use this generation process to perform inpainting, as demonstrated by the approach described in [5] and shown in Figure 1 (RePaint).

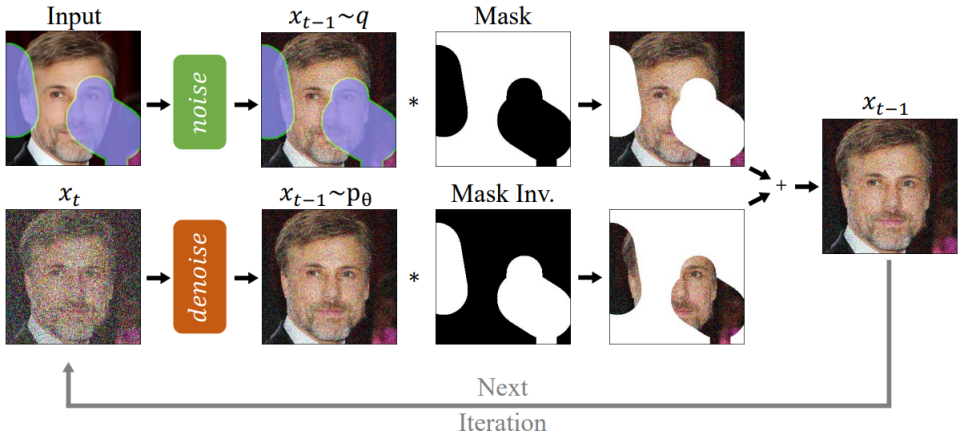


Figure 1: Overview of RePaint [5] approach.

In our experiments, we used samples from the Linnaeus 5 dataset [1], which were obtained from Pixabay, a source of royalty-free images. In addition, we frequently referenced the CelebA sample in relation to the RePaint [5] method. An example of this sample is shown in Figure 2.

Figure 2: The CelebA sample, from CelebA dataset [4].

2 Model and hardware setting

We worked with three U-Net DDPM model trained over imagenet and found on openai guided diffusion repository for 256 sized images, and also with the same architecture trained on CelebA dataset.

- 256x256_classifier.pt (256COND)
- 256x256_diffusion.pt (256COND)
- 256x256_diffusion_uncond.pt (256UNCOND)
- CelebA256_250000.pt (256CelebA)

Most of our experiments were conducted with the unconditional model, to avoid undesired effects due to overfitting. Note that those models do not generalize well to lower image resolution as shown in Figure 3.

Figure 3: 64 sized samples generated with 256UNCOND model.

To evaluate the generalization capabilities of our model, we compared the results of inpainting a face using the 256UNCOND and 256CelebA models with a sample from the CelebA dataset. Some sample comparisons are shown in Figure 4. The results show that the 256UNCOND model may generate less realistic faces compared to the 256CelebA model.

We conducted our experiments using the RePaint [5] and guided diffusion [2] GitHub repositories, adding various features such as the ability to generate samples conditionally based on timestep checkpoints, extract attention, and use different U-Net implementations. We used an NVIDIA RTX 3060 Ti GPU, which allowed us to generate or inpaint a sample in approximately 5 minutes.

3 Inpainting setting

Masks

During our experiments we use three kind of masks described in Figure 5 we refer as square, half and smile masks.

Figure 4: Inpainting a face with 256CelebA (first row) and 256UNCOND (second row).

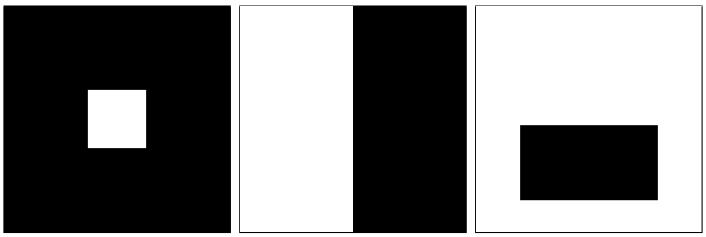


Figure 5: Square mask (left), half mask (center), smile mask (right).

Schedule

$$n = t_T + (-1 + \frac{t_T}{\text{jump length}}) * (\text{jump n sample} - 1) * (2 * (\text{jump length})) \quad (1)$$

To achieve good image inpainting, we use a technique called harmonization rather than slowing down the diffusion process. We set the following parameters: t_T: 250, jump length: 10, jump n sample: 10. Starting at timestep 250, we reverse 10 more diffusion steps (jump length) and then return to the original timesteps, repeating this process 10 times (jump n sample). The timesteps over the first 1000 iterations for our settings are shown in Figure 6. The total number of diffusion operations can be calculated using Equation 1, which is 4570 for our settings. It's worth noting that almost half of these operations are relatively fast, as forwarding diffusion is easier than reversing.

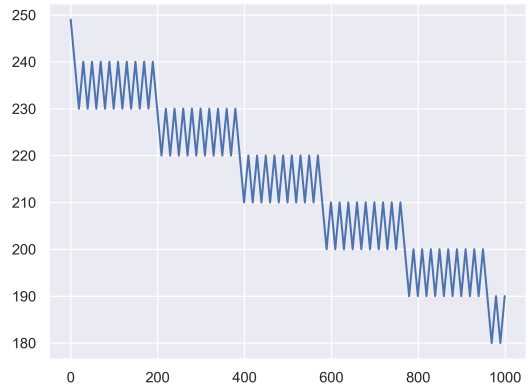


Figure 6: Timestep values for 1000 first iterations.

4 Face inpainting with 256CelebA

4.1 Experiment description

The goal of this section is to perform human smile inpainting with 256CelebA and to challenge its generalization abilities to non-celebrity faces from distinct ethnicities. We will use FairFace [3] images and labels and check the potential bias in coherent face features generation as well as smile match resulting from 4 controlled classes (Black, Indian, Southeast Asian and White).

To limit selection bias, we did not select images within each ethnic group. A corpus of 400 images containing 100 images from each class was produced, and only those corresponding to the inpainting criteria were retained. The model trained on CelebA is indeed very strict on the position of the images (centering around the nose), the scale and the orientation of the faces. Skewed faces and images at the wrong scale were therefore removed. Without alignment, face inpainting tended to generate another noise. Hence, a vertical realignment was performed to normalize the center of the images around the nose and the images were cropped to keep only 256 pixels in width and height. At the end we were left with a total 146 images with almost 35 in each class. The split is roughly balanced, as shown in Table 4.1.

Race	**Number of examples**
Black	34
Indian	43
Southeast Asian	32
White	37

Figure 7 show a few representative examples of good (close to the original) and bad inpainting outputs.

Figure 7: Original (top) and unpainted (bottom) images, with two good examples (left) followed by 2 bad inpaintings (right).

Typical artifacts that we discovered displayed apparition / disappearance of mustache on male/female character and lower part of the face changes such as apparition/disappearance of a smile. While small artifacts did not disqualify the picture, more "shocking" ones will classify as incorrectly inpainted (such as showing a mustache on a female or a child).

Figure 8: Comparison of original (left) and inpainted (right) face for South East Asian (top) and Black (bottom) samples, showing respectively apparition of mustache and disappereance of smile.

4.2 Analysis of biases

As discussed, artifacts shared by all races' inpainted pictures were dismissed (presence of smile, presence or absence of mustache on males). Most of them are probably coming from the celebrity dataset features, which do not show many smiling faces,

for example. However, we see trends where females were transformed after inpainting with a mustache, or simply having very masculine traits/jaw. That phenomenon would seem to occur more often on picture representing the Indian, following by the Black class.

Ethnicity	**Correctly inpainted**
Southeast Asian	28/32 (12% incorrect)
White	35/37 (5.5% incorrect)
Black	26/34 (23.5% incorrect)
Indian	32/43 (25.5% incorrect)

Table 1: Results obtained on all classes for image fidelity or absence of detrimental artifact.

It appears that 256CelebA can be considered as biased due to the training data that is not representative of human faces diversity. It is worth noting that this check has to be visual and we only focused on output artifacts that could be considered a prejudice. To go one more step in the comprehensiveness of the bias analysis, we would need a bigger dataset and more quantitative criterias. This small study is however a good starting point for our analysis of diffusion inpainting, as it shows us that CelebA is a dataset with insufficient diversity on the varieties of human face, and that an analysis of the properties of the model trained on CelebA could be biased: in inpainted images, the model tends for instance to erase smiles. For the rest of our experiments, we will therefore use 256UNCOND, trained on more images and on a wider variety of content.

Now when analyzing only change in presence of smile, we obtain the following results presented in Table 2.

Ethnicity	**Wrong restitution on smile criteria**
Southeast Asian	10/32 (31% mismatch)
White	7/37 (19% mismatch)
Black	6/34 (17.6% mismatch)
Indian	6/43 (14% mismatch)

Table 2: Results obtained on all classes for smiles mismatch between original and inpainted.

5 Generation and inpainting with 256UNCOND

5.1 Image generation

In Appendix A, we show 34 randomly generated samples using the 256 UNCOND model. These samples were not selected specifically, but were randomly picked. Out of the 34 samples, 6 of them show dogs, likely because the Imagenet dataset is biased toward dogs. This is why the model performs particularly well at inpainting dogs, and we will use this in future experiments. Overall, most of the samples are semantically consistent.

5.2 Inpainting

We randomly selected some samples from the *other* category of the Linnaeus 5 dataset and inpainted them using a square mask. Figure 9 shows an example of 9 different inpaintings from a single sample. In Appendix B, we show more random results from different samples. While not all samples are semantically consistent, the results shown were also randomly picked. We also used samples from FairFaces [3] dataset to perform human smile inpainting.

Figure 9: Different inpaintings from an original sample (left) with 256UNCOND.

6 Class conditionning

We generated inpainted images using a single sample and a square mask in three different settings. To address the issue of inconsistent inpainted samples, we also tested class-conditioned inpainting (CCI) using the 256COND model and the *seashore* Imagenet category in two settings with jump lengths of 10 and 20. The results, shown in Figure 11, were randomly selected from the test set. The original seashore sample (Figure 11) had a unmasked region containing both water and foam. Without CCI, the inpainted image's unmasked region varied between depicting water, sky, or white surfaces. With CCI, the unmasked region consistently depicted either foam or water. The landscapes shown in the resulting inpainted images were all similar, as they consistently featured a beach at the bottom. Increasing the jump length did not significantly improve the harmonization of the results, as they were already well-harmonized using our basic settings.

Figure 10: Seashore sample

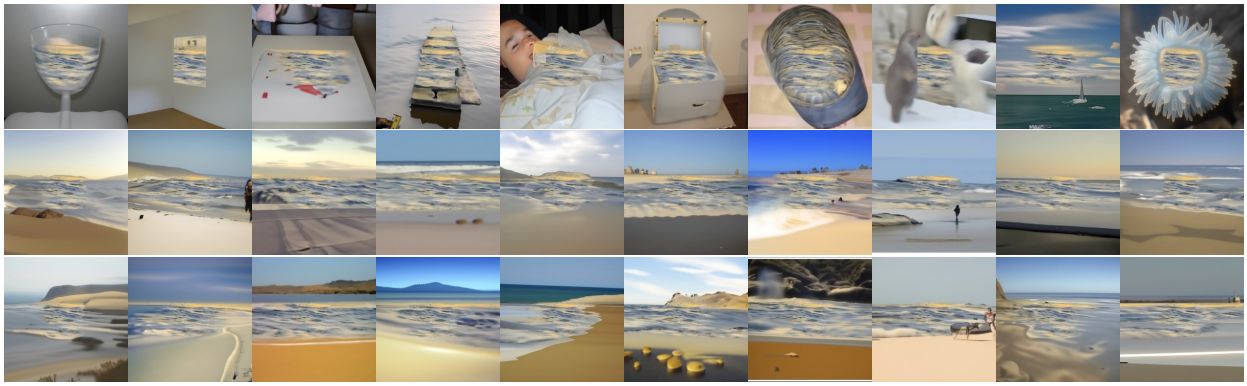


Figure 11: Inpainted samples without conditioning (first row), CCI with jump length 10 (second row), CCI with jump length 20 (third row).

7 Object generation through timesteps

In the image inpainting process with denoising diffusion probabilistic models (DDPMs), the inpainting is stochastic and sequential, allowing us to resample the inpainted region of an image from a particular timestep. Figure 12 illustrates that, even when resampling 10 timesteps before the end of the reverse diffusion process, the result can vary slightly. Our experiments suggest that the diversity of resampling increases with higher starting timesteps, as expected. While there are no quantitative metrics to evaluate this diversity, resampling from timesteps between 250 and 120 often leads to significant changes in the inpainted region in most cases. However, after timestep 110, resampled images tend to be similar to the original inpainted image.

In this section, we present our inpainting experiment using checkpoints with the 256UNCOND model. We use 34 image samples from the 'dogs' category of the Linnaeus 5 dataset and a half mask. First, we inpainted each sample three times and saved checkpoints of the inpainted samples at 23 different steps, as shown in Table 3. Then, for each inpainted sample, we inpainted it again starting from every checkpoint to observe at which point object generation is determined early in the reverse diffusion process. This experiment required 7 days of computation using

Figure 12: Original image (left). Inpainted image with half mask (middle). Resampled inpainting from timestep 10 (right).

our hardware. A few randomly selected results are shown in Appendix E.

200	390	580	770	960	1150	1340	1530	1720	1910	2100	2290	2480
230	220	210	200	190	180	170	160	150	140	130	120	110
2670	2860	3050	3240	3430	3620	3810	4000	4190	4380			
100	90	80	70	60	50	40	30	20	10			

Table 3: Checkpoint steps and corresponding timesteps (red).

8 Attention analysis

The 256UNCOND U-Net architecture features 16 attention blocks at small scales (32, 16, 8), which may result in the observation of spatial biases during inpainting due to the harmonization process aligning the masked region with the unmasked region. The multi-head attention mechanism used by 256UNCOND is defined below as in [6]. We conducted our experiments based on the the attention weight matrix $W = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$, where Q and K are concatenated from each head.

$$\begin{aligned} \text{Multihead}(X) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \\ \text{where head}_i &= \text{Attention}\left(XW_i^Q, XW_i^K, XW_i^V\right) \\ \text{and Attention}(Q, K, V) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \end{aligned}$$

To investigate the existence of such biases, we extracted attention weights between patches (matrix W) and measured their total received attention as the sum of the columns of W.

8.1 Total attention in mask

We conducted an experiment using the 256UNCOND model on 30 images from the Linnaeus 5 dataset with a half mask, measuring the total attention proportion in the masked region for every attention block and every checkpoint defined in Table

3. The mean total attention per block and the mean total attention per checkpoint are shown in Figure 13. Results for each block can be found in Appendix C. We observe a decrease in the attention belonging to the masked region during the process, particularly just before the process ends. One possible explanation for this is that when objects are about to be generated, masked patches rely more on the unmasked regions and therefore receive more attention. This would explain why, in turn, the received attention in the unmasked region tends to increase over the reverse diffusion process.

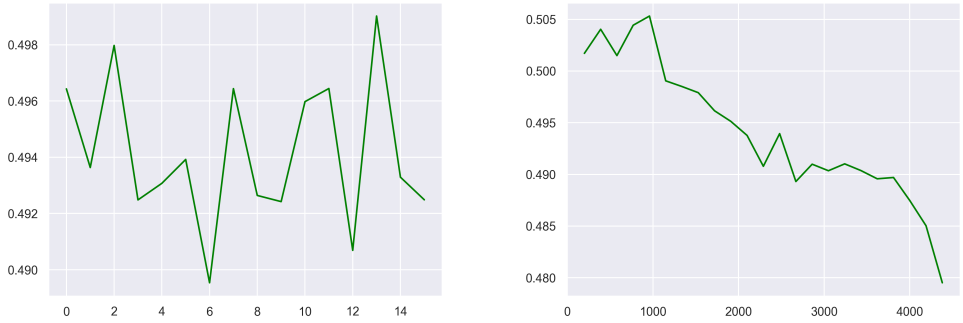


Figure 13: Mean total attention in masked region per blocks (left), per checkpoint (right).

8.2 Mask border effect

To continue our experiments, we use the concept of most important patches, which refers to the 25% patches receiving the most attention. We conducted a small experiment with the 256UNCOND model using the CelebA sample and the first attention block (size 32,32), and observed the most important patches when inpainting at step 4500. The most important patches with respect to the image and mask are shown in Figure 14. We observe that some of the most important patches tend to cluster at the border between the masked and unmasked regions. This phenomenon suggests that the harmonization process between the two regions is concentrated around the mask border, even just before the end of the process. This phenomenon suggests that DDPMs do not perform a progressive 'generation from the borders'.

We investigated the mask border effect with 30 samples from the Linnaeus 5 Dataset inpainted with the half mask. We computed the position of the most important patches at all checkpoints defined in Table 3. We computed their distance from the mask border and observed a spatial bias that is correlated with blocks, but not with timesteps. The spatial border was randomly placed on the right side, due to the even resolutions of the images. Figure 15 shows the empirical distribution of distances, as well as the distribution expected under the hypothesis of a uniform

Figure 14: Most important patches while inpainting with the CelebA sample with respect to the original image (left) and the mask (right).

spatial distribution for the most important patches, for the first attention block. In the uniform case, distances are smaller than 16, and we expect to see fewer patches with a distance of 0, due to half of the candidates being exactly on the border, and half of the patches with a distance of 16, due to half of the candidates having this distance due to the even resolution. To further examine the effect of blocks, refer to Appendix D for additional data.

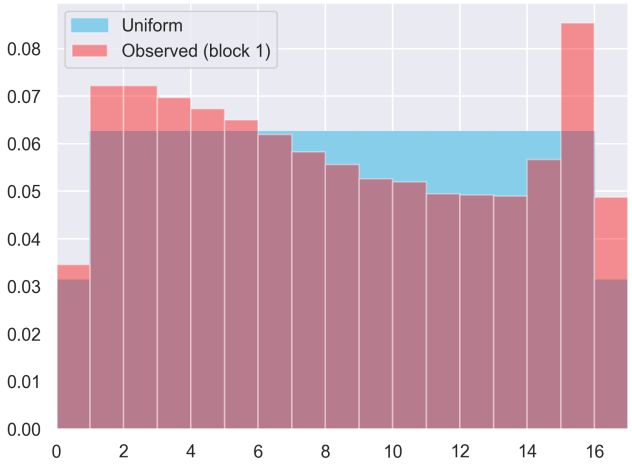


Figure 15: Most important patches distances from half mask border (orange) and theoretical distribution with uniform hypothesis (blue) for first attention block.

As shown in Figure 15, patches with a high level of attention tend to be closer to the mask border, as expected. They also tend to be more prevalent at the vertical border of the image. The exponential map of the most important patches does not clearly show spatial biases, except for the first attention block, which discriminates against three image corners, as shown in Figure 16.

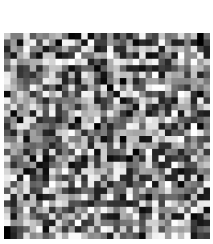


Figure 16: Exponential map of most important patches distribution over the image (mean over timesteps, block 0).

9 Conclusion

Conducting this work was a particularly challenging task due to the lengthy inference time required for the diffusion models used. However, it was also very rewarding as there are few works that focus on the generative properties of DDPMs. Generating samples from multiple checkpoints was time-consuming, and we were unable to conduct a quantitative analysis of the results. Instead, we present the results to the reader and encourage them to study the images themselves. In the future, we plan to continue this work by examining the effect of different masks and exploring the use of conditioning. Our analysis of attention patterns showed that attention tends to accumulate more in the visible area of the image during inpainting, and that it exhibits spatial biases such as a preference for patches closer to the border of the mask. Further investigation of these patterns using different masks would be of great interest.

References

[1] G Chaladze and L Kalatozishvili. Linnaeus 5 dataset for machine learning. chaladze. com, 2017.

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. CoRR, abs/2105.05233, 2021.

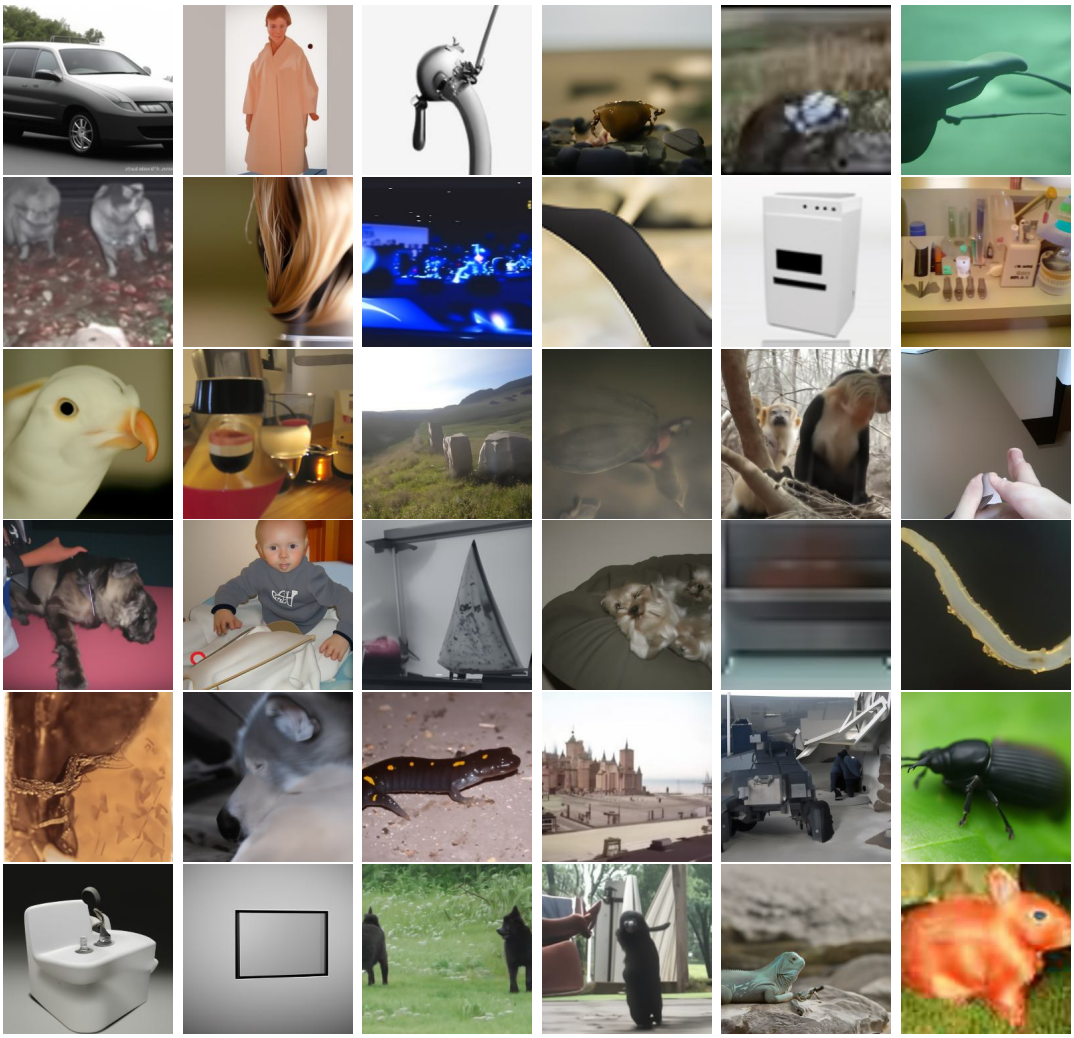
[3] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. pages 1548–1558, 2021.

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[5] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. CoRR, abs/2201.09865, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

A Generated samples with 256UNCOND



B Inpainted samples with 256UNCOND

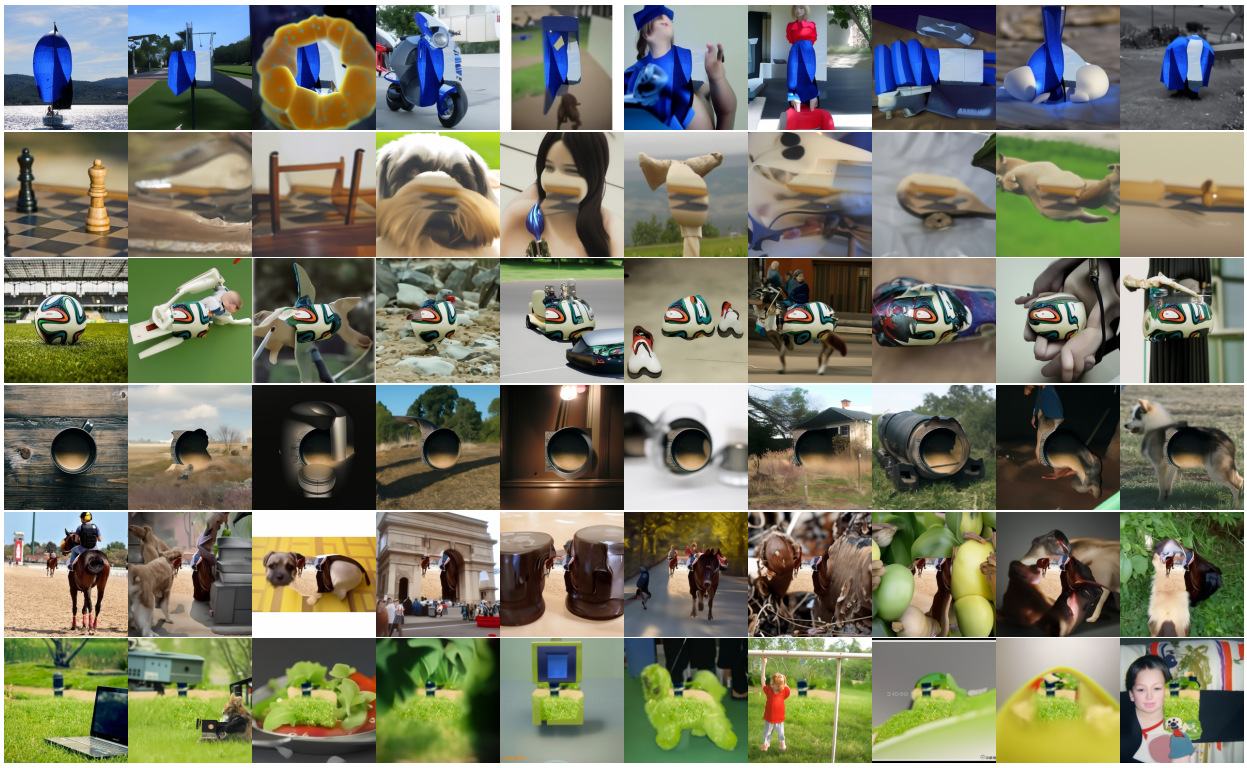


Figure 17: Square mask is used. Images to inpaint are on the left (first column).

C Mean total attention for each group of blocks

256UNCOND has 16 attention blocks with tensor width and height (32,32,16,16,8, 8,8,8,8,8,16,16,16,32,32,32). We observe the proportion of total attention belonging to the masked region for 30 samples of Linnaeus 5 Dataset, for each block and each checkpoint as defined in Table 3 and display the results here.

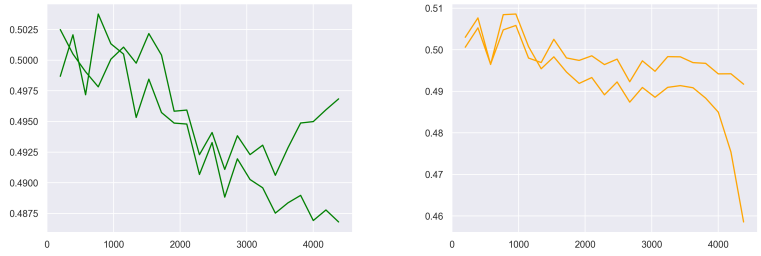


Figure 18: Blocks 1,2 (left, resolution 32), 3,4 (right, resolution 16).

Figure 19: Blocks 5 to 10 (resolution 8).

Figure 20: Blocks 11,12,13 (left, resolution 16), 14,15,16 (right, resolution 32).

The first 2 blocks have a smoother drop of attention belonging to the masked region, while the other blocks have a sudden drop after step 4000, at the end of the reverse diffusion process.

D Mask border effect

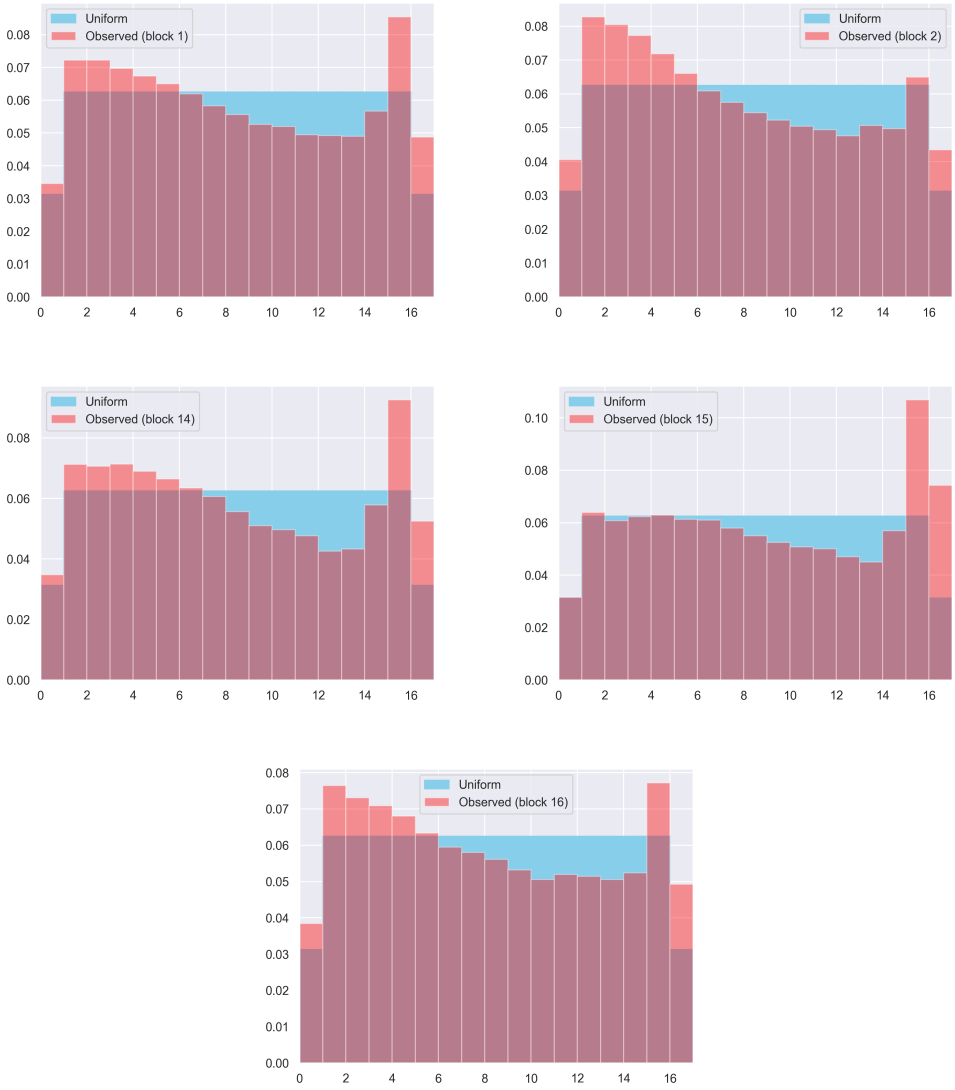
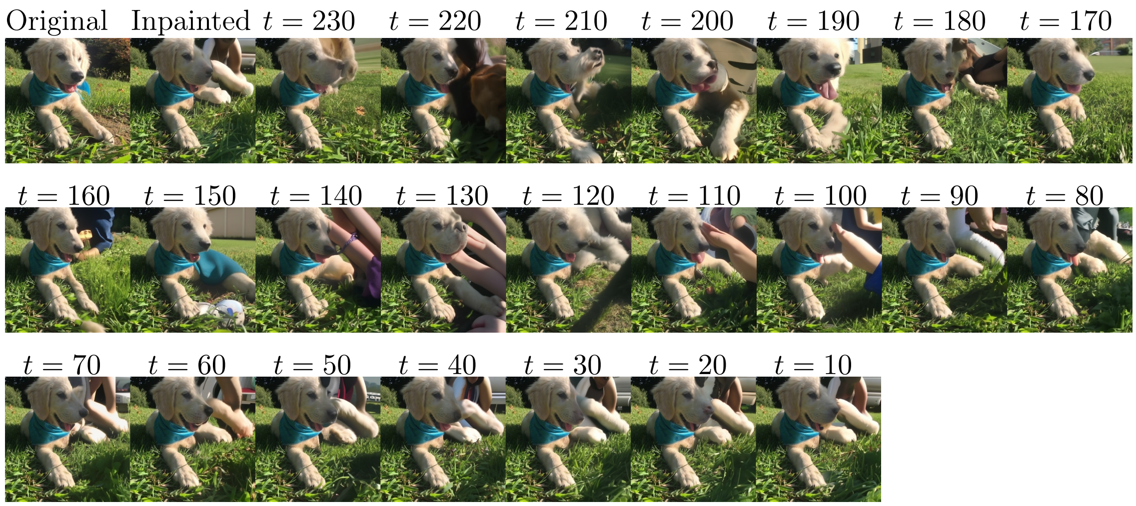
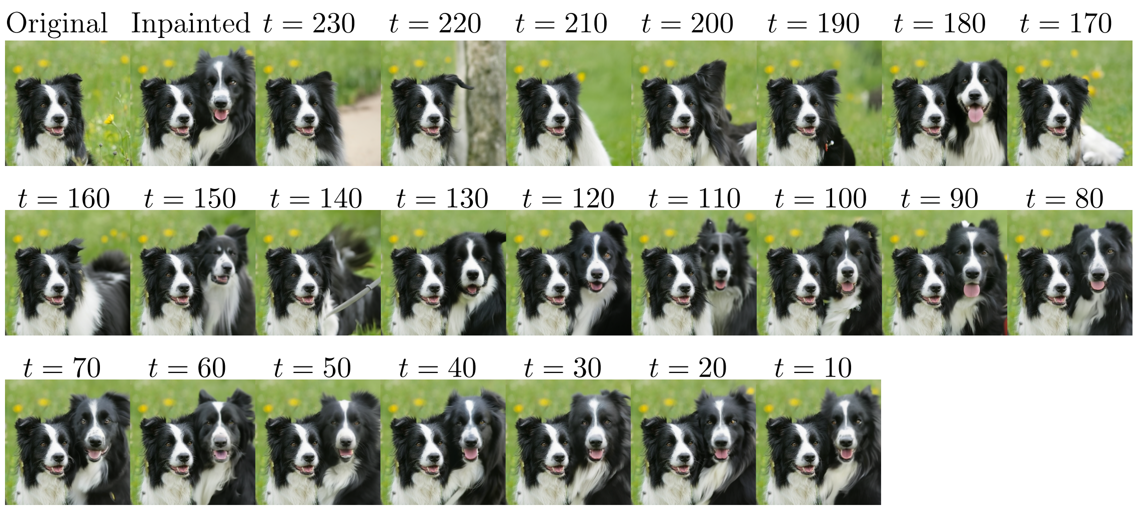


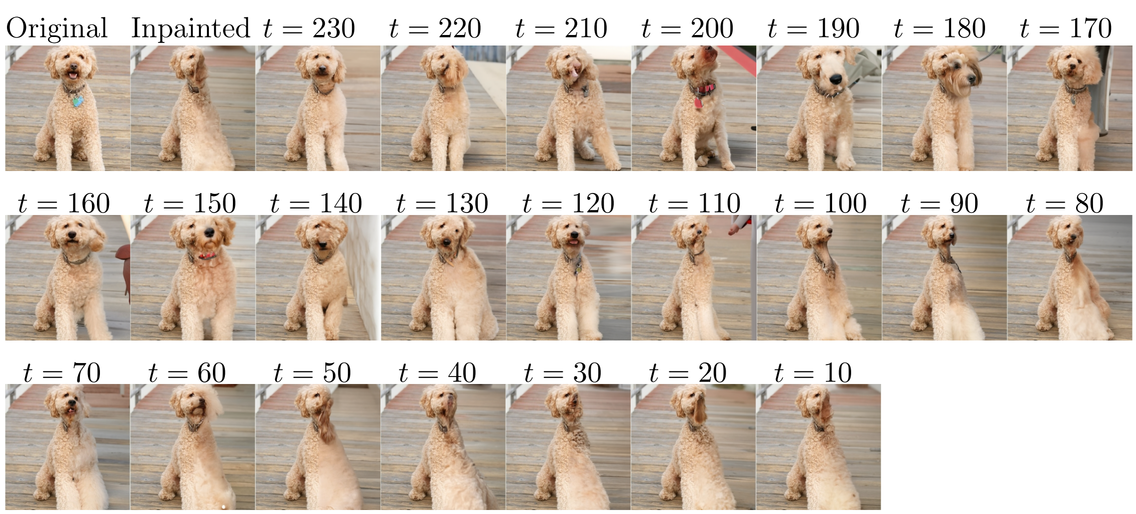
Figure 21: Most important patches distance to mask border for all 32 sized attention blocks.

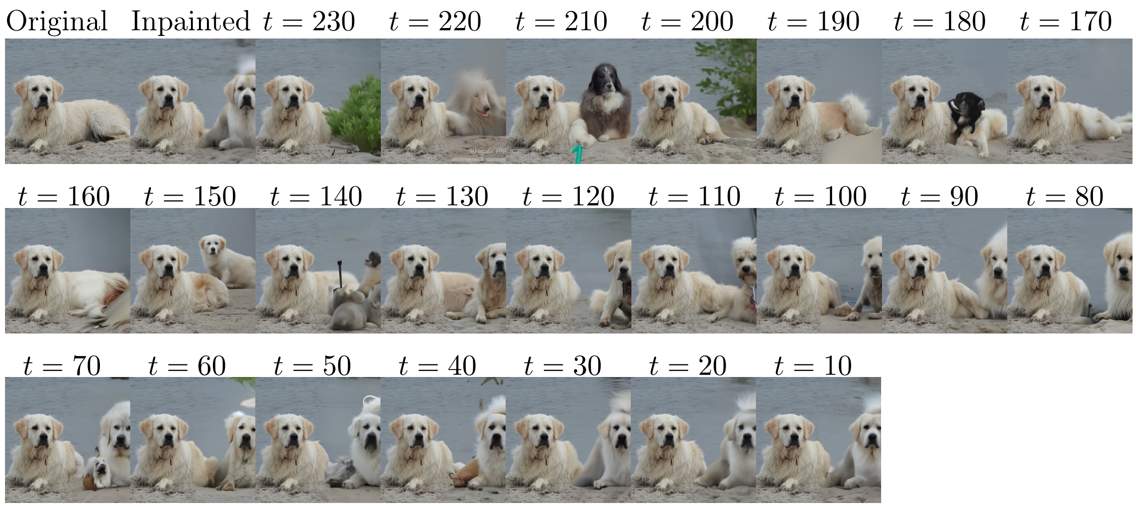
Attention effect near mask border seems to be different for each 32 sized attention block, and is the strongest for block 2.

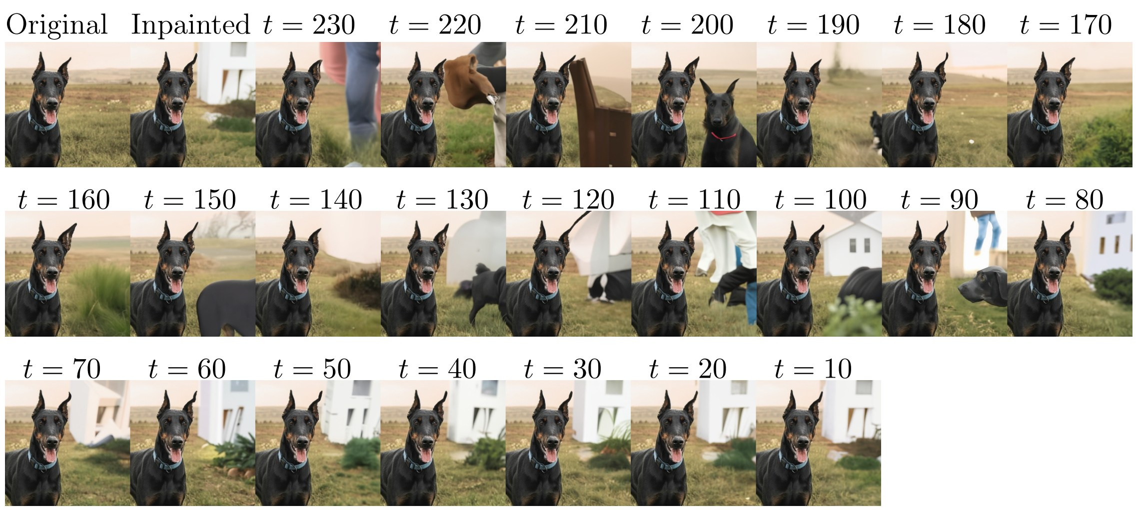
E Inpainting from checkpoint

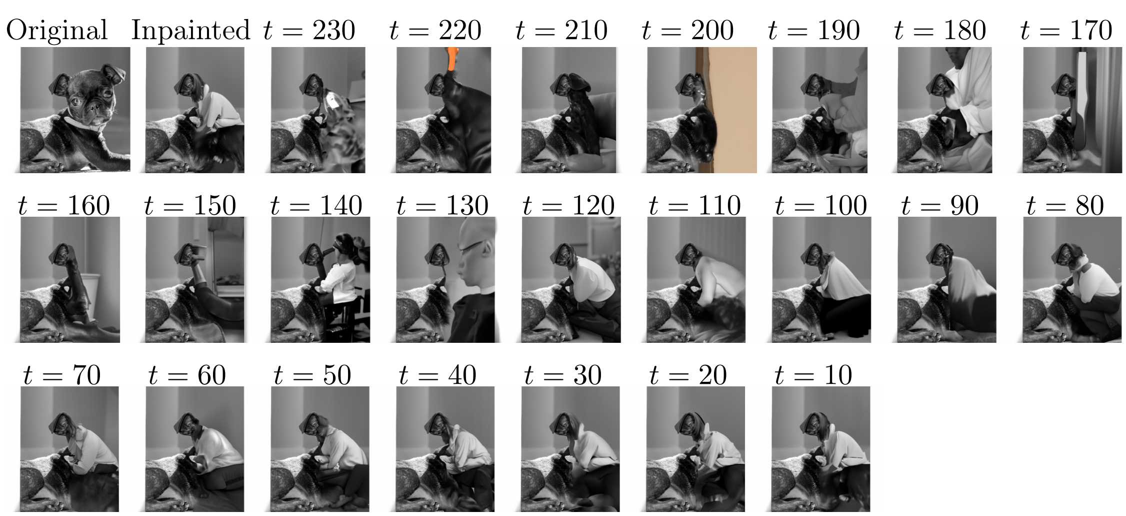


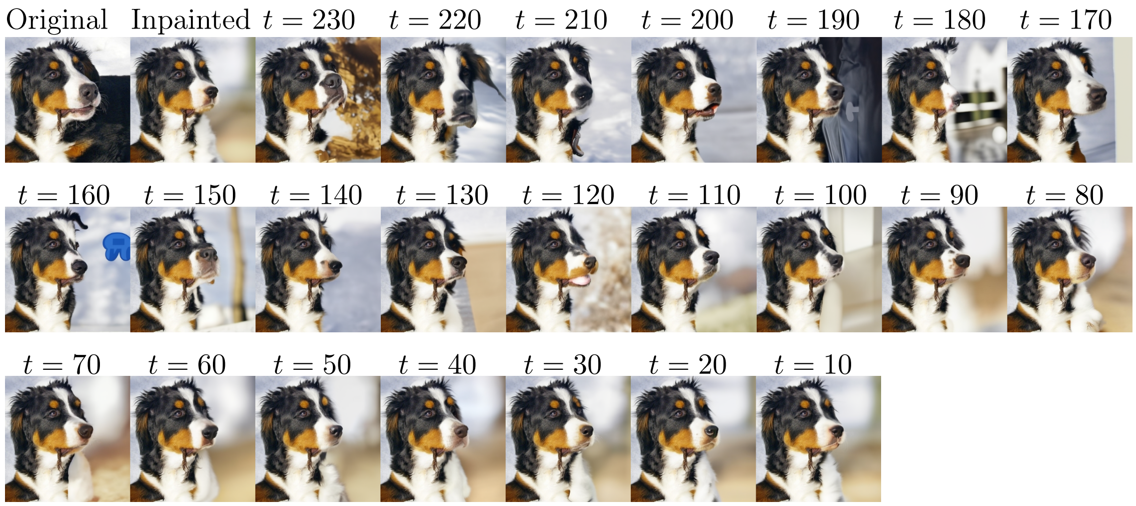


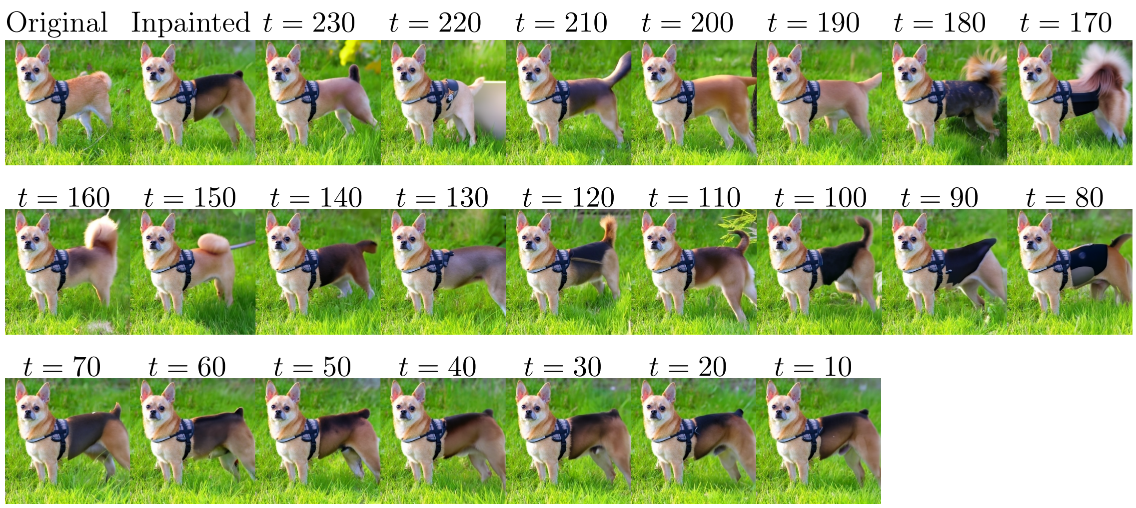


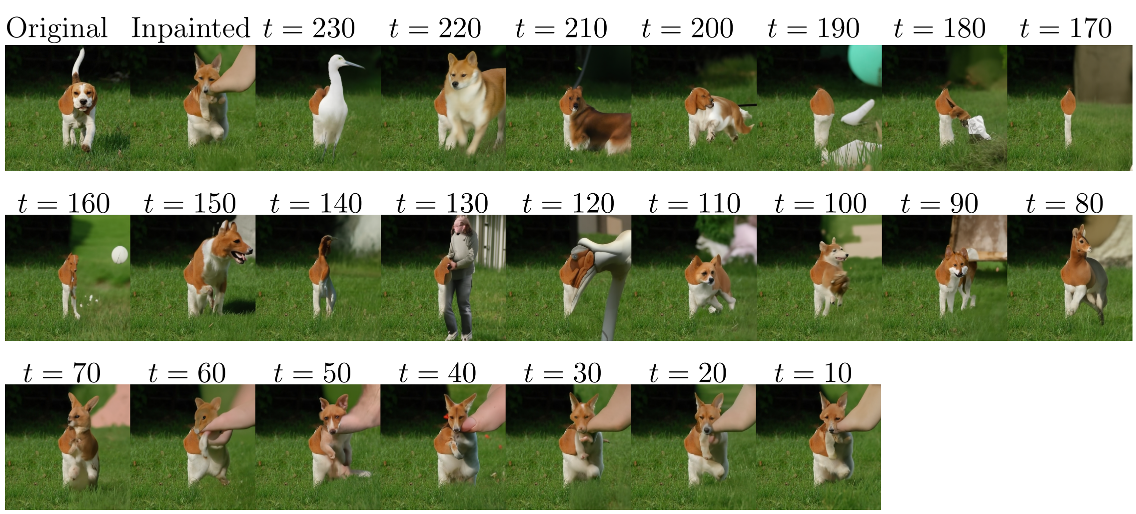

Original
Inpainted
t = 230
t = 220
t = 210
t = 200
t = 190
t = 180
t = 170
t = 160
t = 150
t = 140
t = 130
t = 120
t = 110
t = 100
t = 90
t = 80
t = 70
t = 60
t = 50
t = 40
t = 30
t = 20
t = 10

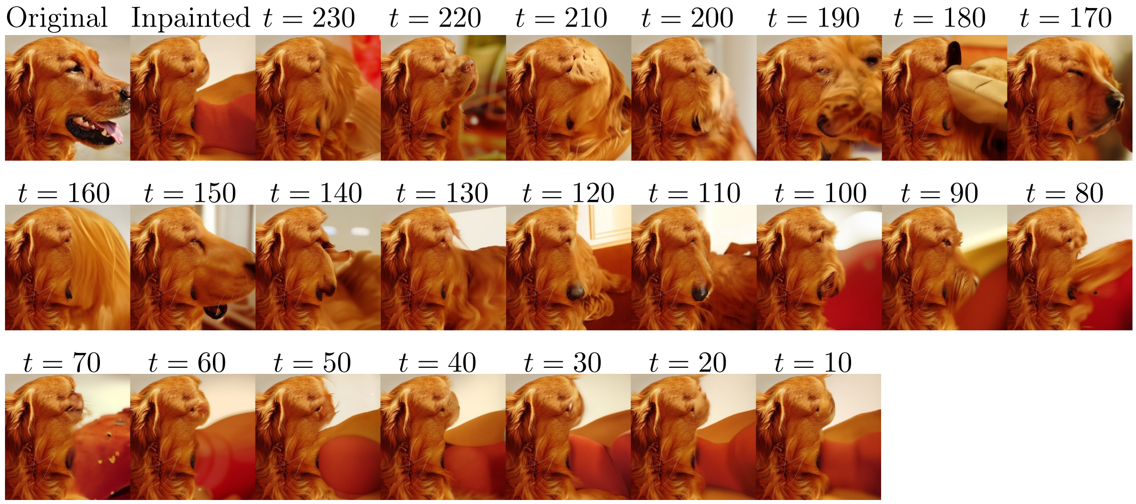

Original
Inpainted
t = 230
t = 220
t = 210
t = 200
t = 190
t = 180
t = 170
t = 160
t = 150
t = 140
t = 130
t = 120
t = 110
t = 100
t = 90
t = 80
t = 70
t = 60
t = 50
t = 40
t = 30
t = 20
t = 10

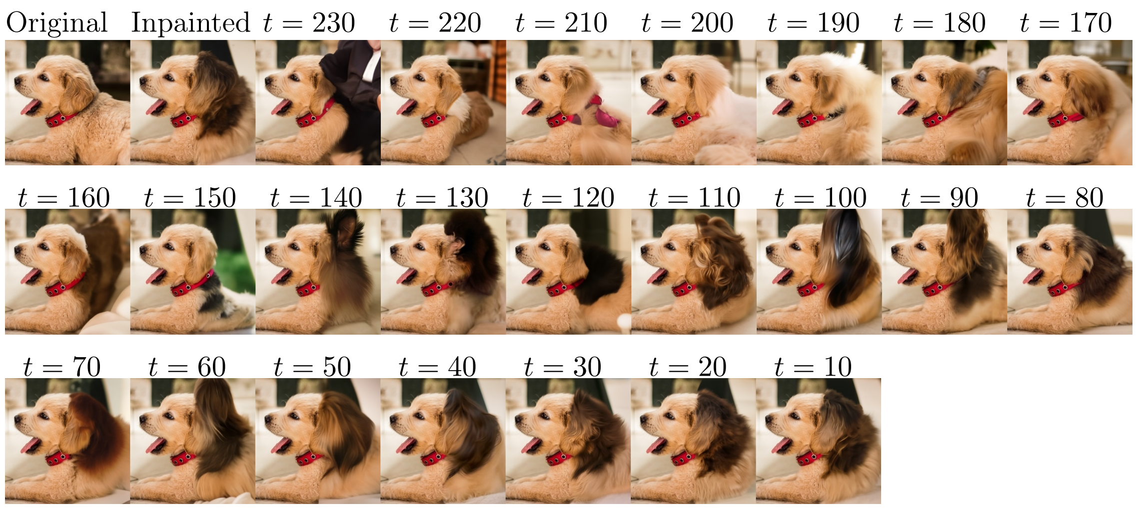

Original
Inpainted
t = 230
t = 220
t = 210
t = 200
t = 190
t = 180
t = 170
t = 160
t = 150
t = 140
t = 130
t = 120
t = 110
t = 100
t = 90
t = 80
t = 70
t = 60
t = 50
t = 40
t = 30
t = 20
t = 10

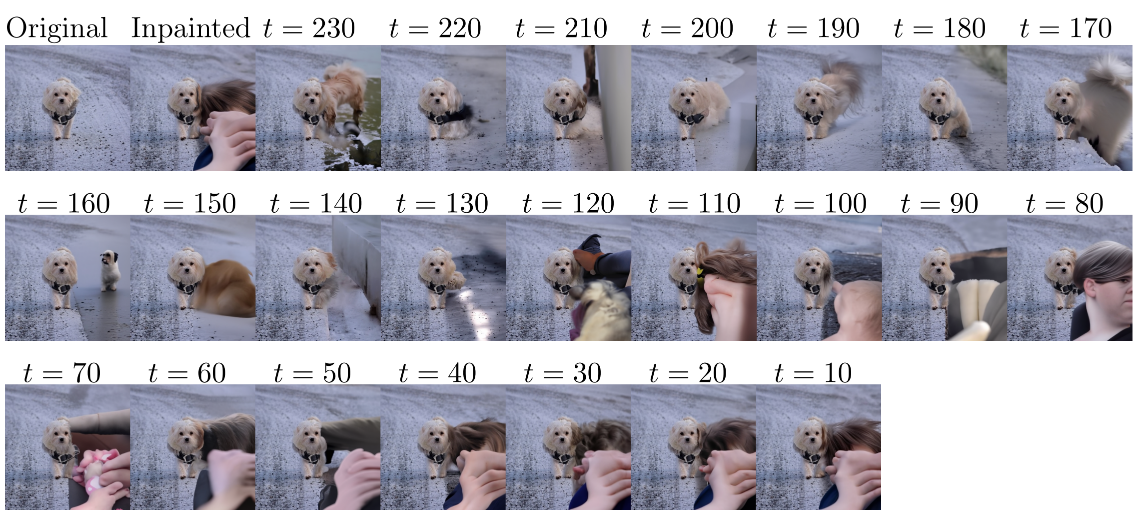

Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$


Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$


Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$


Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$


Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$


Original
Inpainted
$t = 230$
$t = 220$
$t = 210$
$t = 200$
$t = 190$
$t = 180$
$t = 170$
$t = 160$
$t = 150$
$t = 140$
$t = 130$
$t = 120$
$t = 110$
$t = 100$
$t = 90$
$t = 80$
$t = 70$
$t = 60$
$t = 50$
$t = 40$
$t = 30$
$t = 20$
$t = 10$