How Stable is Stable Diffusion under Recursive InPainting (RIP)?

Javier Conde
ETSI de Telecomunicación,
Universidad Politécnica de Madrid
28040 Madrid, Spain

Miguel González
ETSI de Telecomunicación,
Universidad Politécnica de Madrid
28040 Madrid, Spain

Gonzalo Martínez
Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid
28911, Leganés, Madrid, Spain

Fernando Moral
Escuela de Arquitectura,
Universidad Antonio de Nebrija
28015 Madrid, Spain

Elena Merino-Gómez
Escuela de Ingenierías Industriales,
Universidad de Valladolid
47011 Valladolid, Spain

Pedro Reviriego
ETSI de Telecomunicación,
Universidad Politécnica de Madrid
28040 Madrid, Spain

ABSTRACT

Generative Artificial Intelligence image models have achieved outstanding performance in text-to-image generation and other tasks, such as inpainting that completes images with missing fragments. The performance of inpainting can be accurately measured by taking an image, removing some fragments, performing the inpainting to restore them, and comparing the results with the original image. Interestingly, inpainting can also be applied recursively, starting from an image, removing some parts, applying inpainting to reconstruct the image, and then starting the inpainting process again on the reconstructed image, and so forth. This process of recursively applying inpainting can lead to an image that is similar or completely different from the original one, depending on the fragments that are removed and the ability of the model to reconstruct them. Intuitively, stability, understood as the capability to recover an image that is similar to the original one even after many recursive inpainting operations, is a desirable feature and can be used as an additional performance metric for inpainting. The concept of stability is also being studied in the context of recursive training of generative AI models with their own data. Recursive inpainting is an inference-only recursive process whose understanding may complement ongoing efforts to study the behavior of generative AI models under training recursion. In this paper, the impact of recursive inpainting is studied for one of the most widely used image models: Stable Diffusion. The results show that recursive inpainting can lead to image collapse, so ending with a nonmeaningful image, and that the outcome depends on several factors such as the type of image, the size of the inpainting masks, and the number of iterations.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Philosophical/theoretical foundations of artificial intelligence**; • **General and reference → Evaluation**.

KEYWORDS

Stable Diffusion, Generative AI, Stability, Inpainting.

ACM Reference Format:
Javier Conde, Miguel González, Gonzalo Martínez, Fernando Moral, Elena Merino-Gómez, and Pedro Reviriego. 2024. How Stable is Stable Diffusion under Recursive InPainting (RIP)?. In *Proceedings of Workshop on Evaluation and Trustworthiness of Generative AI Models (GenAI-KDD).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Figure 1: Example of the Recursive InPainting (RIP) process on "Van Gogh, Self portrait, 1889" : original and outcome versions side by side.

1 INTRODUCTION

Generative Artificial Intelligence (AI) has taken center stage in the last two years and triggered a new technology revolution. Generative AI models can generate text, audio, images, or video and can be used in many transformative applications. Among the AI tools, Large Language Models (LLMs) such as GPT4 [1], which can answer questions, summarize, translate, and paraphrase texts, and text-to-image generators such as DALL-E [18], which can create

GenAI-KDD, August 26, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

images for almost any text description has attracted the interest of the public with hundreds of millions of users.

These tools have achieved unprecedented performance levels in most tasks, and evaluating their performance is a key issue. In the case of LLMs, many benchmarks have been proposed to assess their knowledge of different topics, their ability to solve math [4] or reasoning problems [22], or their language understanding [8]. Those benchmarks are used to compare models, and when a new model is introduced, typically performance over the most common benchmarks is reported [24]. In the case of image generation tools, a number of metrics have been proposed to evaluate performance, such as the Fréchet Inception Distance (FID) [9], precision and recall [12] or diversity and coverage [16] that try to capture how close are generated images to real ones and how well generated-images cover the range of real images. Another feature supported by some AI image generation tools and implemented with ad-hoc AI models is inpainting [27]. In this case, the AI tool is given an image with missing fragments and has to fill them to complete the image.

Evaluating the quality of the content generated by AI is not only relevant to compare AI models or to assess their progress in different tasks. The widespread adoption of generative AI is transforming the nature of the content on the Internet. AI-generated texts and images are now pervasive and, in some cases, dominant, and the trend is expected to continue in the next years. This has implications for newer AI models as they are commonly trained on data that is scraped from the Internet, so a loop is created where newer AI models are trained with data generated with previous AI models [15]. This can lead to worse performance or even the collapse of AI models [5] and has triggered research on the stability of AI models when trained with their own data [2],[3].

The generative AI feedback loops considered so far involve the training of newer models, so it is a loop across generations of AI models. However, there are other potential generative AI loops that, to the best of our knowledge, have not been studied before. For example, when the input to the AI model is, for instance, an image and the output is also an image, as in the case of inpainting, the AI model can be used recursively on its output, creating a loop. In this case, there is no training, only inferences that are applied recursively. Analyzing the impact of these recursive calls to the AI model on the generated content is of interest to understand whether the AI models are stable or also collapse as in the training loop [2].

In this paper, we analyze the inference feedback loop using a well-known AI image model, Stable Diffusion [19], and the inpainting functionality. An extensive empirical study is conducted to understand when the model is stable and when it collapses. The rest of the paper is organized as follows: in section 2, the inpainting functionality and the generative AI loops are briefly described. Then, the inference loop, denoted as Recursive Inpaiting (RIP), is presented in section 3 and then evaluated in section 4. The limitations of our evaluation, as well as the results, are discussed in section 5. The paper ends with the conclusion in section 6.

2 PRELIMINARIES

2.1 Inpainting

One of the functionalities implemented by some modern generative AI image tools is inpainting [27], which takes an image with missing fragments and fills in those fragments to complete the image [13]. An example of the use of inpainting is illustrated in Figure 2. In this case, Stable Diffusion was used, and we started from a complete image, applied a mask to remove some parts, and then used inpainting to complete the image. This enables a comparison between the original image and the result of inpainting. It can be seen that the tool is able to produce an image that resembles the original one. Interestingly, the AI model used, Stable Diffusion, changes the face to one that resembles a male which matches theories about the painting being a portrait of one of Leonardo Apprentices[1]. Different runs produce results with different types of faces, mostly woman-like.

The performance of inpainting depends on the model, the type of image, and the sizes and locations of the missing fragments [6]. In general, of the information lost in the image fragments, inpainting can only recover a fraction. A number of metrics can be used to measure the similarity between the original image and the reconstructed one [17]: from classical ones such as the Structural Similarity (SSIM) [25] or the multi-scale SSIM (MS-SSIM) [26] based on the pixel level, to more advanced ones such as the Learned Perceptual Image Patch Similarity (LPIPS) [29] or the Paired/Unpaired Inception Discriminative Score (P/U-IDS) [30], which use AI models to capture human-like perceptual aspects.

2.2 Recursiveness in Generative AI

The massive use of generative AI to generate text and images is creating a loop in which AI-generated content is uploaded to the Internet and then scrapped to train newer AI models [15]. This can lead to a performance degradation of AI models or even to their collapse when they are trained with data produced by themselves [5]. This has triggered interest in understanding under which conditions these generative AI models are stable when trained recursively with data produced by the AI models [2],[3]. This depends on several factors, including the model, the amount of AI-generated data used for each retraining, and whether the loop includes a single or several AI models. The study of this loop is important as it may impact both future AI models but also the nature of future content that will dominate the Internet. In all these studies, recursiveness involves training newer AI models with data generated from other AI models, however in some cases recursiveness can occur when using the same AI model for inference only. This, to the best of our knowledge, has not been studied.

3 RECURSIVE INPAINTING (RIP)

An interesting observation is that a different recursive loop for AI image models can be created when using inpainting. This is illustrated in Figure 3; we start from an image and then apply a mask to remove some parts of it and use inpainting to complete them. At this point, we have a second image that the AI image model has partly created. Then, we repeat the process using a different mask to

[1] https://www.cbsnews.com/news/male-model-behind-the-mona-lisa-expert-claims

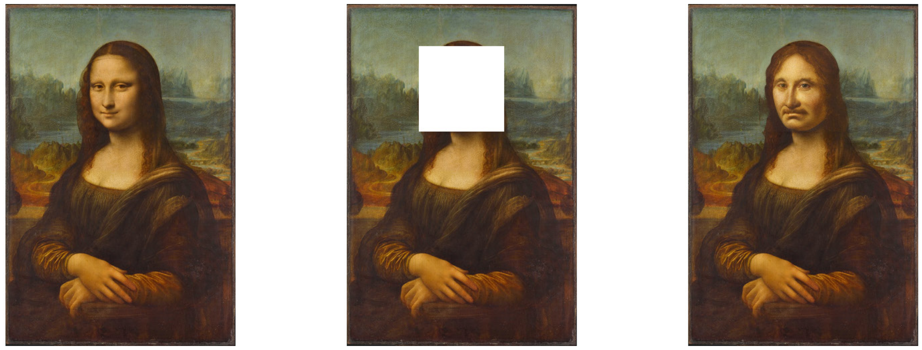


Figure 2: Example of the use of inpainting on "Leonardo Da Vinci, Mona Lisa, 1503", (left) original image, (center) image after applying a mask, (right) image after using inpainting to complete the missing fragment.

obtain a second image that, in this case, is created from AI generated content. The process continues, and we recursively apply inpainting on images that have already been inpainted. In the process, as we remove and reconstruct parts of the images, information will be lost, but will this lead to images that are completely different from the original? images that are simpler and less complex? or will the inpainting be stable and lead to images that are only variations of the original image? As in the case of recursiveness when training models with their own data, it is of interest to understand when inpainting is stable or when it collapses under recursion.

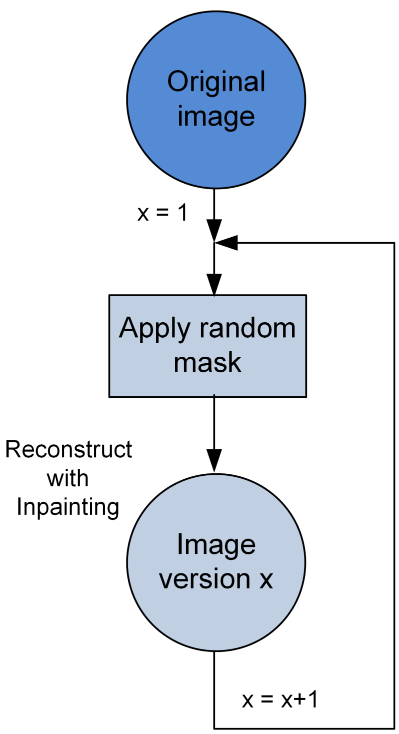


Figure 3: Illustration of the Recursive InPainting (RIP) process.

An example of recursive inpainting is shown in Figure 4. The top left plot corresponds to the original image, in this case, a portrait of Pope Innocent X by Velázquez. The other images correspond to the results after applying inpainting two, four, six,..., up to sixteen times on one-fourth of the image. It can be seen that as the iterations progress, the image starts to depart from the original, and significant changes are introduced. However, even after the sixteen iterations, the final image still resembles the original one. Instead, when the same process is done for a sketch by Vincent van Gogh, as shown in Figure 5 the lady in the original image no longer appears in the last image, which is completely different from the initial one.

The impact of recursive inpainting depends on many factors, such as the AI model, the type of image used, or the masks applied at each iteration. Intuitively, more complex images or masks that remove larger parts of the image will be more likely to lead to collapse. In the following section, the findings of an extensive empirical study of recursive inpainting with Stable Diffusion are presented as a first step toward understanding the key factors that determine the impact of recursive inpainting.

4 EVALUATION

The main parameters for the recursive inpainting are

(1) The AI model.
(2) The input images.
(3) The masks applied at each step.
(4) The number of iterations.

In our experiments, Stable Diffusion [19], a latent text-to-image diffusion model [20],[21], has been used because it is an open model and one of the most widely used AI image models. In particular, a version of Stable Diffusion 2 fine-tuned for inpainting was used[2]. This model employs a mask-generation technique [23] where the masked regions, along with the latent VAE representations of the masked image, serve as additional conditioning for the inpainting process. The model parameters were set to the default values. No text prompt was used to guide the inpainting to make the model focus on reconstructing the missing parts from the remaining visual elements with no text guidance.

[2]https://huggingface.co/stabilityai/stable-diffusion-2-inpainting

Figure 4: Example of recursive inpainting on "Diego Velázquez, Portrait of Innocence X, 1650". On the left, the original image is shown. Each subsequent image to the right displays the result after applying two recursive inpainting operations, up to the final image after sixteen inpainting operations.

Figure 5: Example of recursive inpainting on "Vincent Van Gogh, Lugekone, 1885", On the left, the original image is shown. Each subsequent image to the right displays the result after applying two recursive inpainting operations, up to the final image after sixteen inpainting operations.

As for the images, to try to avoid bias in the selection, they have been selected randomly from a large dataset with more than 81,0000 art images of several types and made by different artists[3]. From this dataset, 100 images were randomly chosen to create our evaluation set[4]. The input images are 512x512 pixels. when their original form factor is not square, blank bands are added on the sides to fit the 512x512 pixels format.

To generate the masks for inpainting, the images are divided into squares of a given size, and in each iteration, a square is randomly selected and used as the mask. The generation of the masks is illustrated in Figure 6 for the case of a 128x128 square and two iterations. It can be observed that one square is removed in each iteration. Then, inpainting is run, and the results obtained for the pixels in the mask are used to replace the ones in the initial picture. This modified picture is then used as the input image for the next iteration. This procedure guarantees that at each iteration, the inpainting only modifies the pixels in the selected mask.

To estimate the similarity with the original image across iterations, we use the Learned Perceptual Image Path Similarity (LPIPS) [29] metric widely used to assess the quality of inpainting[5]. In the implementation used, the features of three neural networks can be used to compute the metric: SqueezeNet [10], AlexNet [11], and VGG [28].

To enable a direct comparison of inpainting with different mask sizes, our experiments use as the main parameter not the number of inpainting operations but the number of pixels on which inpainting is done. For example, for a 256x256 mask, four inpainting operations correspond to changing a number of pixels equal to those in the original 512x512 images. Instead, for a 128x128 mask, sixteen inpainting operations are needed to change 512x512 pixels[6]. Using as a metric the number of pixels that are inpainted relative to the image size makes comparisons easier across different masks and image sizes.

In the first experiment, we take the 100 random images and perform recursive inpainting for 400% of the pixels with masks of 64x64, 128x128, 256x256. To quantify the degradation as inpainting operations are done, the LPIPS metric between the original image and each generation has been computed using the three neural networks (SqueezeNet, AlexNet, and VGG) features. The average distances on the 100 images at each step of 50% inpainting are shown in Figure 7. The bars show the standard deviation observed on the samples on each of the data points. Several initial observations can be made from the results:

(1) As recursive inpainting progresses, the distance with the original image increases. This could eventually lead to an image that bears no resemblance to the original.
(2) The slope of the distance tends to become smaller but does not seem to stabilize even when the distance is large.
(3) The difference with the original image is larger when the size of the mask used for inpainting is larger which as discussed before is in line with the intuition that it is harder to inpaint larger blocks.
(4) The three networks used to compute the LPIPS (SqueezeNet, AlexNet, and VGG), provide similar results.
(5) The standard deviation is significant which suggests that different behaviors will be observed for different images.

To better understand the variability of the distances for each image, scatter plots of the LPIPS distances of the 100 images for each of the neural networks are shown in Figure 8. It can be observed that there is significant variability across images but the trends are similar to the ones observed in the mean: distance is larger with more inpainting and with larger masks. Comparing the three networks (SqueezeNet, AlexNet, and VGG), the last one, VGG is the one with fewer outliers. VGG is also the most complex network and thus should be expected to better capture the features of the

[3]https://huggingface.co/datasets/huggan/wikiart
[4]The results of our experiments are available, both image and metrics as well as the scripts to run the experiments at https://zenodo.org/doi/10.5281/zenodo.11532111
[5]The implementation used is available in a public repository https://github.com/richzhang/PerceptualSimilarity
[6]Note that the pixels changed in two iterations can be the same as each iteration selects a mask randomly.

Figure 6: Example of recursive inpainting on "El Greco, The Nobleman with his Hand on his Chest, 1580".

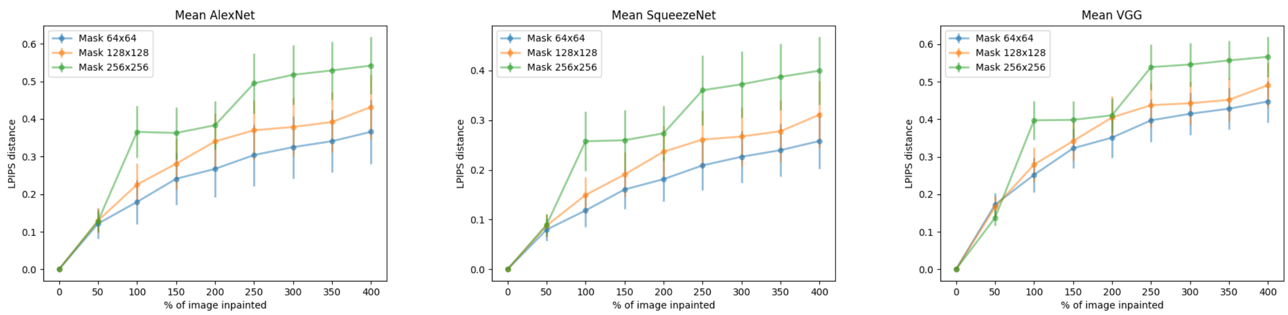


Figure 7: Mean LPIPS across the 100 images versus the inpainting done for AlexNet (left), SqueezeNet (middle) and VGG (right) for different mask sizes (64x64, 128x128, 256x256).

images [28]. Therefore, in the following, we only report results for VGG although all the metrics are available in the repository along with the images.

Another factor that can impact the degradation is the image used as the starting point for the process. To analyze this, LPIPS distance plots were generated for each image and analyzed manually. A few illustrative examples are shown in Figure 9. In the first one (left), the distance tends to stabilize as recursive inpainting progresses. In the second, there is a large difference in the distances with mask size and finally, in the last one, the distances are similar for all mask sizes. The image sequences for the three images are shown in Figures 10,11,12.

In the first image even when using a small mask, see Figure 10a, we observe that the image gradually becomes blurred and loses quality as the recursive inpainting progresses. However, it still manages to maintain the overall style of the original image. With a medium-sized mask, Figure 10b, the initial iterations show a similar trend, but as the blurring errors accumulate, the image eventually degenerates completely in the final iterations. When using a large mask, Figure 10c, the style of the image is quickly lost, resulting in a completely different image by the end of the process. Therefore, it seems that the stabilization of the error can be at least partially attributed to the collapse of the process producing images unrelated to the original one.

For the second image, the results in Figure 11 show the impact of the mask size, when it is small, Figure 11a, the inpainting focuses on reconstructing the texture of the image as there are no objects. Instead, for the largest mask size Figure 11c, the inpainting starts adding new objects which leads to more and more objects ending with a collage. Instead in the third image, the results in Figure 12 are similar for all mask sizes and only towards the end differences can be observed for the largest mask size in Figure 12c. In this case, removing a large part of the image does not cause the insertion of new objects, leading to a different behavior. In summary, the type and features of the initial image seem to be an important factor for the outcome of recursive inpainting.

As the parts removed are randomly chosen, it is of interest to see whether the degradation is similar on different runs. To understand the variability of the degradation with the run, 10 images have been selected from the set of 100, and each has been run 10 times. The LPIPS metrics across runs for three different images are shown in Figure 13 when using the VGG network which again tends to have the lowest deviations. It can be observed that the variations are larger for larger masks which is expected as the larger the mask, the fewer the iterations to reach a given percentage of inpainting which causes more variability. The variations are also reduced as the percentage of inpainting increases showing again, that the larger the number of inpainting operations the lower the variability. This means that recursive inpainting seems to converge in terms of LPIPS distance as the process advances.

Finally, looking at the results in qualitative terms from an aesthetic perspective, the results raise several important concerns,

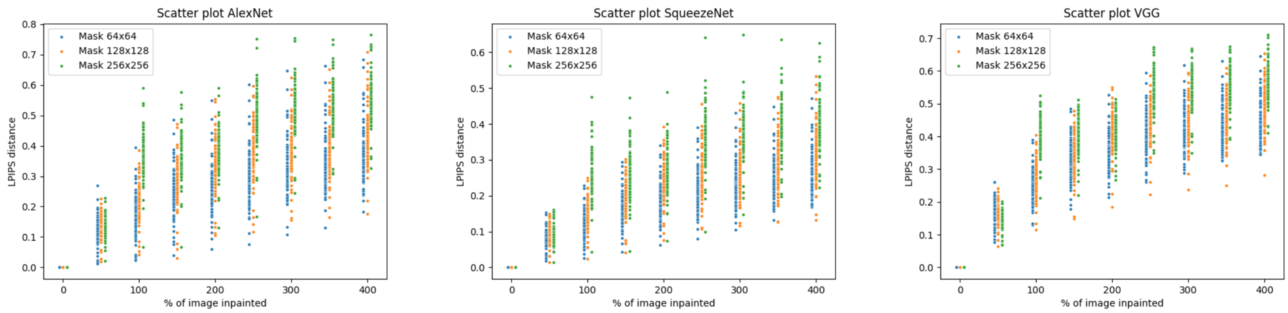


Figure 8: Scatter plots of the LPIPS across the 100 images versus the inpainting done for AlexNet (left), SqueezeNet (middle) and VGG (right) for different mask sizes (64x64, 128x128, 256x256).

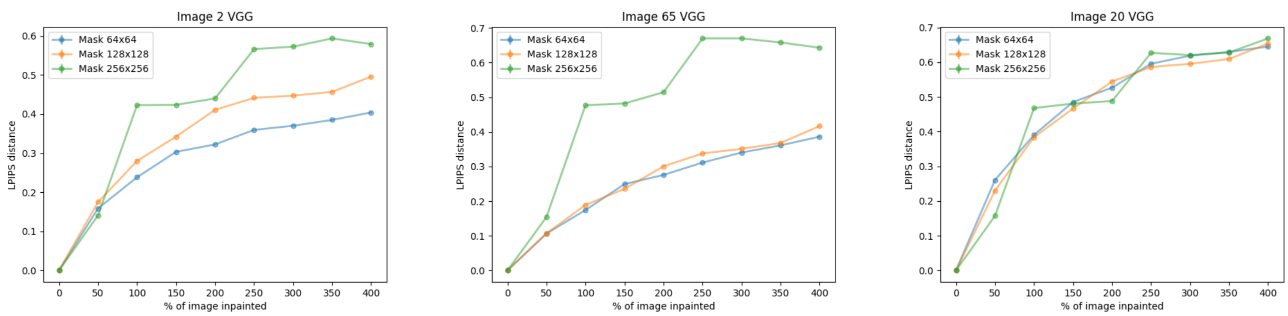


Figure 9: Examples: Distance stabilizes (left), Large distance and variation with mask size (middle), similar distance with mask size (right).

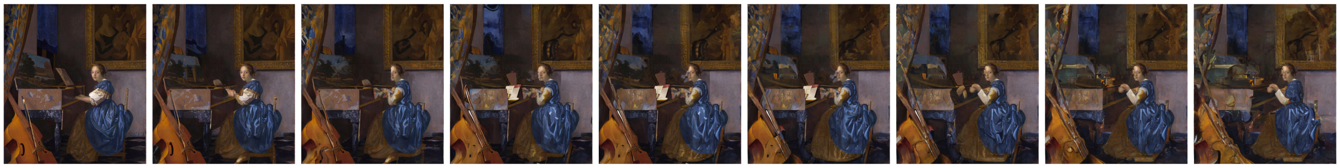
independent of the quantitative analysis of the metrics. It is surprising when the AI, instead of fixing the missing parts, creates new things that don't fit in the painting. This likely happens when small bits left over from the erased sections make the AI assume an object was there, even if it wasn't. This shows the AI doesn't truly recognize the painting it is dealing, with which raises doubts about the entire result. Moreover, the AI seems to not know the rules for making things look realistic in terms of perspective. It twists angles in odd ways at times. The worst cases happen when the AI can't comprehend how faces are rotated and tries to reconstruct them haphazardly, just to make it resemble a head or a human component. In some images, the AI attempts to reconstruct a sort of color palette in the best situations, particularly when the alterations are minimal. However, in other cases, it simply uses arbitrary colors or elements to try to resemble the original image which results in a pixelated appearance.

5 LIMITATIONS AND DISCUSSION

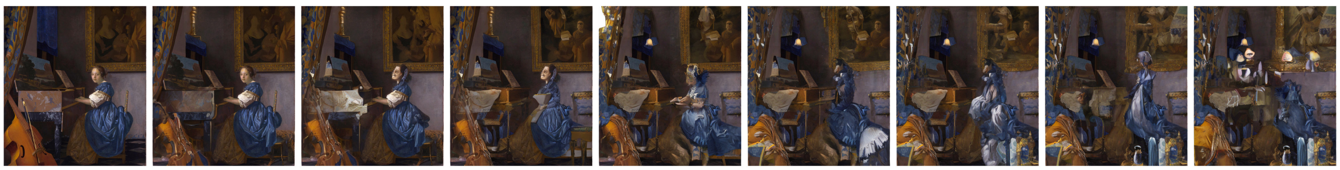
The evaluation conducted is just an initial step to understand the effects of recursive inpainting. Additional experiments with different configurations of Stable Diffusion are needed to evaluate the impact of each of the parameters. Ablation studies that replace Stable Diffusion with other AI inpainting models should also be conducted to understand if the observed behavior is specific to Stable Diffusion or occurs to inpainting models in general. The same reasoning applies to the input images, a larger number of images, possibly with different features in terms of the objects represented and their shapes and sizes should be evaluated as in our experiments we have focused on paintings.

In addition to these extensions of the empirical evaluation, theoretical models that can explain the impact of recursive inpainting have also yet to be developed. Another area for study would be to compare the results of AI to those produced by humans when presented with the same problem. However, doing experiments with humans would require a significant effort and also depend on their painting capabilities. To apply recursive inpainting, several persons, one per iteration would be needed to make sure that they have not seen the original or previous images in the series which makes the procedure rather complicated. In summary, as discussed before, this paper is just the first step in the analysis of recursive inpainting that is primarily intended to present the problem and motivate further work.

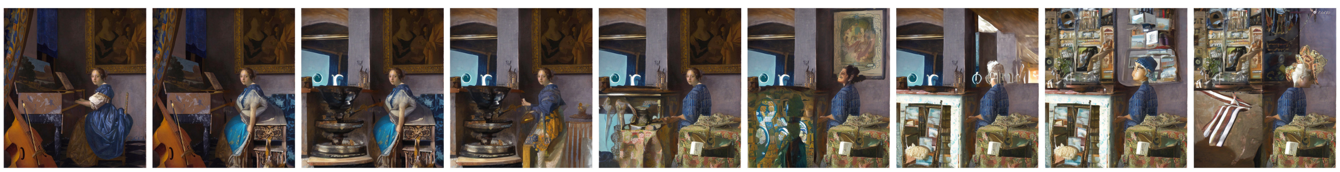
Even with the limitations discussed, the results presented show how recursive inpainting can lead to images that are completely different from the original ones. This is similar to the model collapse observed when training generative AI models with their own data [14] for which techniques to avoid collapse are being proposed [7]. Analyzing the similarities and differences between recursive inpainting and recursive training loops is another avenue for future research. Exploring modifications to the AI models to avoid the



(a) 16x16

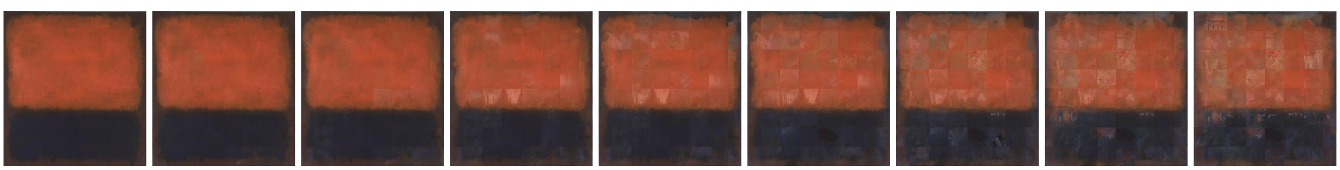


(b) 64x64



(c) 128x128

Figure 10: Results of the recursive inpainting with different mask sizes for the image corresponding to Figure 9 (left), "Johannes Vermeer van Delft, Lady Seated at a Virginal, c. 1672".

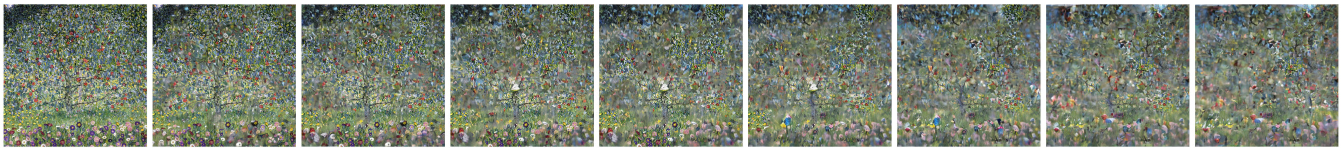


(a) 16x16

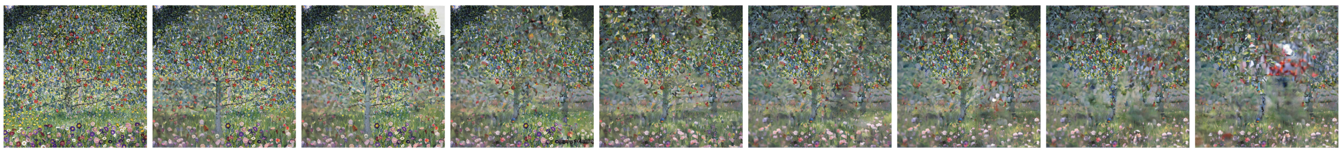
(b) 64x64

(c) 128x128

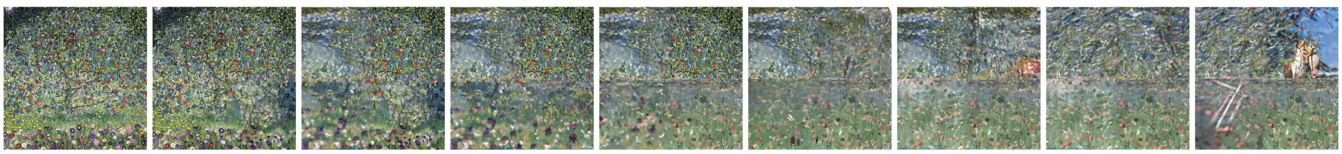
Figure 11: Results of the recursive inpainting with different mask sizes for the image corresponding to Figure 9 (middle), "Mark Rothko No. 14, 1960".



(a) 16x16



(b) 64x64



(c) 128x128

Figure 12: Results of the recursive inpainting with different mask sizes for the image corresponding to Figure 9 (right), "Gustav Klimt, The Apple Tree, 1912".

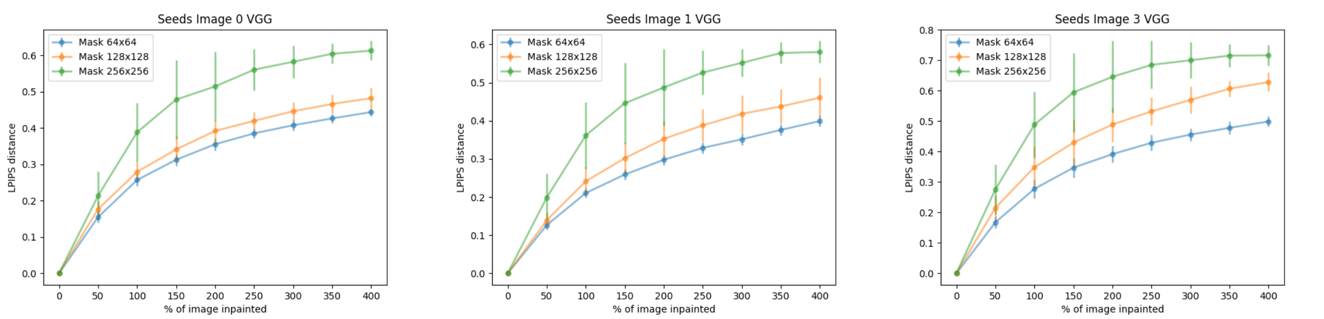


Figure 13: LPIPS for ten runs with different seeds on the same image for three different images and mask sizes (64x64, 128x128, 256x256).

collapse of recursive inpainting is also of interest and could lead to better image-generation AI models. More broadly, understanding if there is a relationship between the model collapse effect on recursive training and the degeneration of the image in the recursive inpainting is also of interest.

Finally, although recursive inpainting has been proposed as a tool to study the stability of AI inpainting models, its use to produce variations or transformations of an image may also be an interesting area for future work that to the best of our knowledge has not been explored so far.

6 CONCLUSION AND FUTURE WORK

In this paper, the effect of recursive inpainting on AI image models has been presented and studied empirically. The results show that recursiveness can lead to the degradation and, eventually, the collapse of the image. This is similar to what has been observed in the recursive training of generative AI models, which is attracting significant interest from the community. Therefore, this paper opens another area in the research of the impact of the recursive use of generative AI, in this case only in the inference phase, that can complement existing research efforts and lead to further insights on the causes of collapse. This can, in turn, lead to improvements in the AI models to mitigate the impact of recursiveness.

The analysis of recursive inpainting presented in this paper is just the first step. Additional AI models, images, and model configurations should be tested to better understand the impacts of recursive inpainting. Beyond empirical results, it is also of interest to develop theoretical models that can explain the impacts of recursive inpainting. Exploring the links between recursive training and recursive inpainting is also an interesting area for future research.

ACKNOWLEDGMENTS

This work was supported by the FUN4DATE (PID2022-136684OB-C22) and ENTRUDIT (TED2021-130118B-I00 projects funded by the Spanish Agencia Estatal de Investigación (AEI) and by the Chips Act Joint Undertaking project SMARTY (Grant no. 101140087).

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. 2023. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429* (2023).

[3] Martin Briesch, Dominik Sobania, and Franz Rothlauf. 2023. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822* (2023).

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[5] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. A Tale of Tails: Model Collapse as a Change of Scaling Laws. *arXiv preprint arXiv:2402.07043* (2024).

[6] Pedro Reviriego Fernando Moral-Andrés, Elena Merino-Gómez and Fabrizio Lombardi. 2023. Can Artificial Intelligence Reconstruct Ancient Mosaics? *Studies in Conservation* 0, 0 (2023), 1–14. https://doi.org/10.1080/00393630.2023.2227798 arXiv:https://doi.org/10.1080/00393630.2023.2227798

[7] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. 2024. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. *arXiv preprint arXiv:2404.01413* (2024).

[8] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[10] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[12] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11461–11471.

[14] Matteo Marchi, Stefano Soatto, Pratik Chaudhari, and Paulo Tabuada. 2024. Heat Death of Generative Models in Closed-Loop Learning. *arXiv preprint arXiv:2404.02325* (2024).

[15] Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. 2023. Towards understanding the interplay of generative artificial intelligence and the Internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*. Springer, 59–73.

[16] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*. PMLR, 7176–7185.

[17] Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. 2024. Deep Learning-Based Image and Video Inpainting: A Survey. *International Journal of Computer Vision* (2024), 1–34.

[18] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[21] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2256–2265.

[22] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[23] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust Large Mask In painting with Fourier Convolutions. In *2022 Ieee Winter Conference On Applications Of Computer Vision (Wacv 2022)*. IEEE COMPUTER SOC, 3172–3182.

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[26] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.

[27] Shiyuan Yang, Xiaodong Chen, and Jing Liao. 2023. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3190–3199.

[28] Wei Yu, Kuiyuan Yang, Yalong Bai, Tianjun Xiao, Hongxun Yao, and Yong Rui. 2016. Visualizing and comparing AlexNet and VGG using deconvolutional layers. In *Proceedings of the 33 rd International Conference on Machine Learning*.

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[30] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428* (2021).