AI Foundation Models: Redefining the Future of Image Editing

Matheus Ribeiro de Oliveira Piotr Nobis
Technical University of Munich
https://github.com/matt576/image-editing

Abstract

This project explores the development of an image-editing tool leveraging foundation models to enable sophisticated image modifications, enhancements, and transformations. The tool provides a variety of capabilities, including background blurring and replacement, object removal, inpainting, outpainting, restyling, and superresolution. By integrating technologies such as the Segment Anything Model (SAM) for mask generation, Depth Anything for depth estimation, and Stable Diffusion for image generation tasks, the tool achieves high-quality results comparable to state-of-the-art photo-editing applications.

1. Introduction

The development of foundation models has opened up new possibilities for practical applications. Modifications and transformations of images are more accessible than ever before due to advancements in deep learning methods. Powerful models allow users without prior experience or knowledge in image editing to effortlessly refine, modify, and transform images. These advanced technologies enable sophisticated real-time image manipulations, with potential applications in various domains, from digital art creation, to enhancing everyday photography.

This project focuses on the development of the AFM Image-Editing Tool, an AI-supported platform designed to modify images using various cutting-edge technologies. We demonstrate the capabilities of advanced diffusion models and provide users with a set of tools for editing images. Our final product incorporates mask generation, background blurring, background replacement, object removal, inpainting (object replacement), outpainting, restyling, superresolution, and text-to-image (Txt2Img) generation. Our main contributions are:

- Creation of an advanced image-editing tool, including integration, implementation, and enhancement of advanced functionalities powered by foundation models.
- Making sophisticated image manipulations accessible to users without prior software engineering or image editing experience, via a user-friendly GUI.
- Comprehensive qualitative comparison of various functions and models, evaluating their performance against industry-leading products, including a user survey.

2. Related Work

The AFM Image-Editing Tool utilizes several foundation models trained on large datasets, ensuring robust generalization for various tasks. These models deliver high levels of realism, quality, and variety, which are essential for effective image editing. Below, we outline the key models for image generation and editing tasks used in our project:

- **Stable Diffusion**: Image generation and transformation method enabling high-quality inpainting, outpainting, restyling, and superresolution. Stable Diffusion [8] uses latent diffusion models, which operate in latent space to efficiently handle image data. The process involves a forward phase, including gradually adding noise to the image until it resembles Gaussian noise. The reverse phase removes the noise to recover the data sample. The model is trained to learn the distributions needed to generate new samples. This approach allows for synthesis of realistic and detailed images from text prompts, existing images, or for filling in missing parts.
- **Depth Anything**: This component provides Monocular Depth Estimation (MDE) of images. By leveraging large dataset of unlabeled images and training using pseudo labels, the authors of Depth Anything [10] achieved better estimates than those presented in [5]. Moreover, the authors of [10] employ semantic-assisted perception using DINOv2 [3] for auxiliary supervision of depth estimation. Depth Anything is considered the state-of-the-art model for depth estimation.
- **Segment Anything Model (SAM)**: Developed by Meta, SAM [1] provides precise mask generation, an essential step in many image-editing applications. It enables accurate selection and manipulation of specific image regions. The model is trained on the SA-1B dataset, which contains around one billion masks, from 11 million images. Its lightweight mask decoder and various segmentation prompts allow for quick and flexible mask generation.

3. Method

The AFM Image-Editing tool incorporates a diverse set of functionalities that leverage advanced methods for intuitive image refinement and editing. In this section, we introduce the method and pipeline for image generation.

1) **Mask Generation**: Process of creating a binary mask, which is a map identifying selection of parts of the image. Masking plays a big role in image processing, since its functionality is reused in pipelines such as for inpainting, outpainting, and object removal, where we are interested in modifying a certain image section. Our Image-Editing Tool enables generation of mask via pixel coordinates selection using SAM [1], and text prompt using Grounded SAM [7]. The use of masking often involves mask dilation (expanding the mask by a selected number of neighboring pixels) to improve the quality of the generative task in creating a seamless transition to the environment, specially on object replacement and removal tasks.
2) **Background Blurring**: Applies depth-based blur to the image obtained via depth map, generated by Depth Anything [10]. Image blurring involves foreground extraction using RMBG-1.4 model, which keeps the foreground elements sharp, as well as setting focal point on the foreground element and depth-based blur relative to the distance to the object. We also provide blurring given explicit mask of the foreground generated manually.
3) **Background Replacement**: Reuses the foreground extraction model, and replaces background of an image by applying stable diffusion for inpainting, based on an input text prompt.
4) **Object Removal**: Enables erasing unwanted elements in the image. Given a mask, it performs a generative fill of the selected parts of the image, completing the environment and focusing on a seamless transition to the rest of the image. LaMa [9] and Latent Diffusion for inpainting [8] are applied for this task. No text prompt is necessary.
5) **Inpainting**: Process of reconstructing parts of an image, indicated by a provided mask of the element. Involves replacing the selected object based on the environment and input text prompt.
6) **Outpainting**: Expands the original image by indicated number of pixels in each chosen direction(s), performs inpainting on the expanded frame sections. Involves filling in the outer frames following the given text prompt.
7) **Restyling**: Process of altering the appearance of an image to reflect a different artistic style or aesthetic. The applied models can change colors, textures, patterns and other style elements, while keeping the underlying content intact. The artistic style is modified using text prompt defined by the user. 'Guidance Scale' and 'Strength' parameters can be altered, influencing the creativity of the output provided by generative models.
8) **Superresolution**: Enhances the quality of an image by applying 4x superresolution. Our pipeline resizes the dimensions of an image to multiples of 128, before being cropped in patches of size 128 x 128. We apply LDM 4x Superresolution [8] and Stability AI's 4x Upscaler [8] models on the individual patches. As a final step we merge the outputs to complete the full output image, with a higher resolution.

Function	Checkpoint
Mask Generation	facebook/sam-vit-huge
Mask Generation	facebook/sam-vit-huge + GroundingDINO [2]
Background Blurring	LiheYoung/depth_anything_vitl14
Foreground Extraction	briaai/RMBG-1.4
Background Replacement	stabilityai/stable-diffusion-2-inpainting
Background Replacement	diffusers/stable-diffusion-xl-1.0-inpainting-0.1
Object Removal	ldm_inpainting/last
Object Removal	big-lama/best
Outpainitng	stabilityai/stable-diffusion-2-inpainting
Outpainting	diffusers/stable-diffusion-xl-1.0-inpainting-0.1
Restyling	runwayml/stable-diffusion-v1-5
Restyling	kandinsky-community/kandinsky-2-2-decoder
Restyling	stabilityai/stable-diffusion-xl-refiner-1.0
Superresolution	CompVis/ldm-super-resolution-4x-openimages
Superresolution	stabilityai/stable-diffusion-x4-upscaler

Table 1. Checkpoints

The aforementioned functionalities are extended by a Txt2Img pipeline, which enables optional input image generation directly from text.

4. Results & Discussion

In this section, we qualitatively compare the models for the same number of inference steps and input prompts, and evaluate the functions against available Google Photos tools. Refer to the Appendix B for a more detailed overview of the output example images. Our Gradio App, which is a user-friendly GUI incorporating the functionalities, is presented in A.

4.1. Background Blurring

The evaluation method involved visually and qualitatively comparing our results with those from Google Photos using the same blur intensity parameter and without applying any sharpening. In our approach, foreground elements are first extracted, allowing for a more precise application of the blur effect to the background. In the image examples from Appendix B.1, the foreground object was successfully isolated, and the background was blurred with varying intensity. This resulted in a stronger blur effect on image pixels that the depth map indicated were further away.

In contrast, Google Photos uses a point-based method to select the focal point. For this example, choosing a point in

the middle of the element of interest resulted in partial blurring of the element itself. Additionally, when qualitatively comparing the backgrounds between two images, the differences are barely noticeable to the naked eye, demonstrating that our method produces results comparable to Google's. If the element of interest extends along the depth dimension, our approach ensures the object remains sharp. Ultimately, the choice depends on the specific needs and use cases, such as whether the foreground element should stay sharp or whether the user prefers automatic foreground extraction over a point-based blur.

4.2. Object Removal

In this subsection, we evaluate the performance of our object removal method using the images in Appendix B.3. We compare the performance of Latent Diffusion [8], LaMa inpainting [9], and Google Photos, focusing on how these models handle object removal with both dilated and non-dilated segmentation masks.

Mask generation postprocessing plays a fundamental role in this task, significantly impacting the quality of the results. When using dilated segmentation masks (i.e. extending the white area over the black background), all models perform well, achieving seamless background integration with minimal artifacts. On the other hand, non-dilated masks, such as those generated directly by SAM [1], result in suboptimal outputs. These models often leave fragments of the original object or environment, leading to noticeable imperfections in the edited image.

- Latent Diffusion: This model excels at handling hidden color transitions within the environment originally behind the removed object. It produces smooth results and exhibits natural continuity in color and texture, surpassing the quality achieved by Google Photos.
- LaMa: While effective, this model tends to produce less vibrant colors compared to the original image. This might be attributed to the use of Fourier convolutions in their method, which could affect color vibrancy.
- Google Photos: Background transitions are more abrupt, particularly where there are variations in lighting, such as between darker grass in shadow and lighter grass under sunlight on the dog picture in the park.

One common challenge across all models, including Google Photos, is handling shadows and lighting continuity in the area where the object was removed. The resulting images often display darkness in the previously occupied space, contrasting with the surrounding environment. This issue highlights a limitation in current models' ability to accurately simulate shadow removal and maintain lighting balance.

4.3. Background Replacement

To evaluate our background replacement method, we compared the performance of two models: Stable Diffusion [8] and Stable Diffusion XL (SDXL) [4], both tested with the same number of inference steps to ensure a fair comparison (see Appendix B.2).

SDXL, a larger model with more parameters, required a slightly longer processing time compared to Stable Diffusion. However, the additional computational effort resulted in noticeably superior image quality. Upon close examination, the SDXL output exhibits much finer details and a higher level of realism in background integration. In contrast, the standard Stable Diffusion model produced less consistent results, especially for less inference steps, where artifacts and blending inconsistencies were more apparent, along with less detailed structures for the same prompts used with SDXL.

4.4. Inpainting (Object Replacement)

In this task, we evaluate object replacement capabilities using Stable Diffusion, Stable Diffusion XL (SDXL), and Kandinsky v2.2 [6]. The models were tested using identical prompts, negative prompts, and the same number of inference steps. The prompts emphasized realism, detail, and accurately depicted limbs. As a test case, an image of a bear was used, with the bear being replaced by an astronaut (details in Appendix B.4).

- SDXL: The model emerged as the top performer, delivering superior textures, color accuracy, and detail. It effectively maintained realism and produced highly detailed images, making it the preferred choice for tasks requiring high-quality and realistic inpainting.
- Stable Diffusion: While not as detailed as SDXL, Stable Diffusion (with and without ControlNet [11]) generated consistent and acceptable results with shorter inference times. This makes it a viable option when computational efficiency and time constraint are prioritized over absolute detail and realism.
- Kandinsky v2.2: The outputs from Kandinsky were generally less detailed than those from SDXL. Unless specifically adjusted by negative prompts, Kandinsky often produced more inconsistent results and images with vibrant and flashy colors. This characteristic can be beneficial in applications where more vivid and artistic representations are desired.

4.5. Restyling

For the task of Restyling (Img2Img), we divide the results discussion into two parts: model comparison for same (negative) prompts and inference steps, and the effects of parameters 'strength' and 'guidance scale' on the image generation (see Appendix C).

The same models mentioned in the subsection 4.4 were applied here. The qualitative comparison from the previous section 4.4 holds, with the addition of an extra observation: for the Kandinsky model, adding terms such as 'flashy and vibrant colors' to the negative prompt resulted in a more grounded and realistic output, as desired. For more details, please refer to the Appendix B.5.

4.6. Outpainting

For the outpainting task, we compared the performance of Stable Diffusion and SDXL using an image of a woman in front of a mirror, visible up to her shoulders (details in Appendix B.6). The goal was to extend the image by 200 pixels in all four directions, ensuring she is depicted wearing a red dress in her bedroom.

- SDXL: This model performed almost flawlessly, meticulously following the prompt. The extended image depicted the woman wearing a red dress in her bedroom, maintaining a high level of detail and continuity. However, despite specifying keywords such as "deformed" and "extra limbs" in the negative prompt, a third forearm was erroneously generated, highlighting a known issue in AI models regarding accurately rendering human limbs.
- Stable Diffusion: This model showed more limitations. It incorrectly depicted the woman in a white dress instead of red, indicating a misunderstanding of the prompt. Additionally, the bedroom extension was less seamless than that produced by SDXL, with noticeable inconsistencies in detail and integration, especially involving mirrors and door frames. Surprisingly, no deformed or extra limbs were generated.

These results demonstrate the superior capability of SDXL in following complex prompts and achieving high-quality outpainting results. However, the persistent issue with rendering human limbs accurately suggests an area for further improvement. Stable Diffusion, while faster, struggled with prompt accuracy and seamless integration in this task, making it less suitable for scenarios requiring precise adherence to detailed instructions.

4.7. Superresolution

In the superresolution task, we compared the performance of the LDM x4 Superresolution pipeline and the Stability AI's 4x Upscaler using an image of a white cat (details in Appendix B.7). The objective was to enhance the image resolution while improving detail and realism.

The Latent Diffusion model often provided superior results for the same number of inference steps, producing high-resolution images with enhanced detail and clarity. It also required considerably less inference time compared to the Stability AI Upscaler, making it an efficient choice for generating detailed images quickly.

While the Stability AI Upscaler model sometimes added a degree of realism to the upscaled image, it also intensified the colors, making them more vibrant than in the original image. This change in color vibrancy may not always be desirable, depending on the user's requirements for maintaining color fidelity.

The comparison highlights that the Latent Diffusion model is particularly effective for tasks requiring quick processing and high detail enhancement, making it well-suited for applications where speed and clarity are required. On the other hand, the Stability AI Upscaler's overall inconsistency with results, its tendency to enhance color vibrancy, and long inference time make it an inferior model to LDM 4x Superresolution pipeline.

4.8. User Survey

As an additional evaluation method, we conducted a user survey in which participants selected their preferred image for each function. While this approach is highly subjective, we ensured that the focus was on evaluating the overall realism and detail of the results. The survey images can be found in Appendix B.

1) Background Blurring B.1: Out of the provided images, our method was the clear winner, with only 20% of participants choosing Google Photos' results as preferred. In a direct comparison across all of four images, our method outperformed Google Photos every time. Participants appreciated our approach, which emphasizes the foreground object and applies a stronger blur to the more distant background elements.
2) Background Replacement B.2: Stable Diffusion XL (SDXL) was the preferred model, receiving 80% of the votes and winning in each direct comparison. The high level of detail and seamless transitions to the new background were the main factors influencing the choice of the participants.
3) Object Removal B.3: For erasing objects, Latent Diffusion was the top choice with 55% of the votes, followed closely by LaMa with 40%, and Google Photos with only 5%. Latent Diffusion performed better in managing shadow and lighting transitions, while LaMa in images with cluttered backgrounds and further objects.
4) Inpainting B.4: In the task of object replacement, Stable Diffusion with ControlNet and SDXL were tied, each receiving 45% of the votes, while Kandinsky received 10%, and Stable Diffusion received no votes. Despite of expectations for SDXL to be preferred, Stable Diffusion with ControlNet proved to be extremely competitive for inpainting tasks, showing strong approval from participants for its good transition to the environment, even though not as detailed results as SDXL.
5) Restyling B.5: For applying different artistic styles to the input images, SDXL was the clear winner, gain-

ing 80% of the votes. It was followed by Stable Diffusion and our enhanced version of Kandinsky (using extra negative prompts for flashy colors), each receiving 10%. SDXL's consistent adherence to prompts and detailed textures made it the best performer for this task.

6) Outpainting B.6: Similar to Restyling, SDXL was the preferred method with 75% of the votes, compared to 25% for Stable Diffusion. The high level of detail and consistent extension of the environment with similar textures were its strongest attributes.
7) Superresolution B.7: Among the enhanced images, 70% of users preferred the result of the LDM pipeline over Stability AI's upscaler. This can be attributed to its consistent high level of detail in this task. While Stability AI's model produced some good results, worth 30% of the votes, it particularly struggled with images of lower original resolution, despite offering vibrant colors and good detail.

5. Conclusion & Future Work

In conclusion, the AFM Image-Editing Tool successfully demonstrates the potential of foundation models to deliver sophisticated image-editing capabilities. The app achieves high-quality results across multiple tasks, including super-resolution, background- and object manipulation, by leveraging state-of-the-art models like SAM and Latent Diffusion. Our comparative analysis confirms the tool's effectiveness against industry-leading applications and highlights areas for enhancement.

Future work should focus on improving limb rendering, shadow removal, and lighting consistency to further enhance realism in images. Additionally, exploring the integration of more model architectures can expand the tool's functionality, along with optimizing inference time. Studying the potential of extra input image conditioning in more detail (e.g., using ControlNet combined with Stable Diffusion) could also be a point to consider for more flexibility and robustness of our tool, specially for the inpainting task. User feedback will be essential in guiding the development of new features and refining the app to make advanced image editing even more accessible.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 3

[2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 3

[5] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020. 1

[6] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion, 2023. 3

[7] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[8] 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, 2, 3

[9] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2, 3

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1, 2

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

Appendix

A. Gradio App

To integrate all the functions discussed in this paper, we developed an application from scratch using the Gradio Python library. The final graphical user interface (GUI), shown in Figure 1, is designed to be both intuitive and user-friendly, allowing users to easily leverage foundation models for image editing. Our app is accessible to beginners while also offering features that appeal to more experienced users.

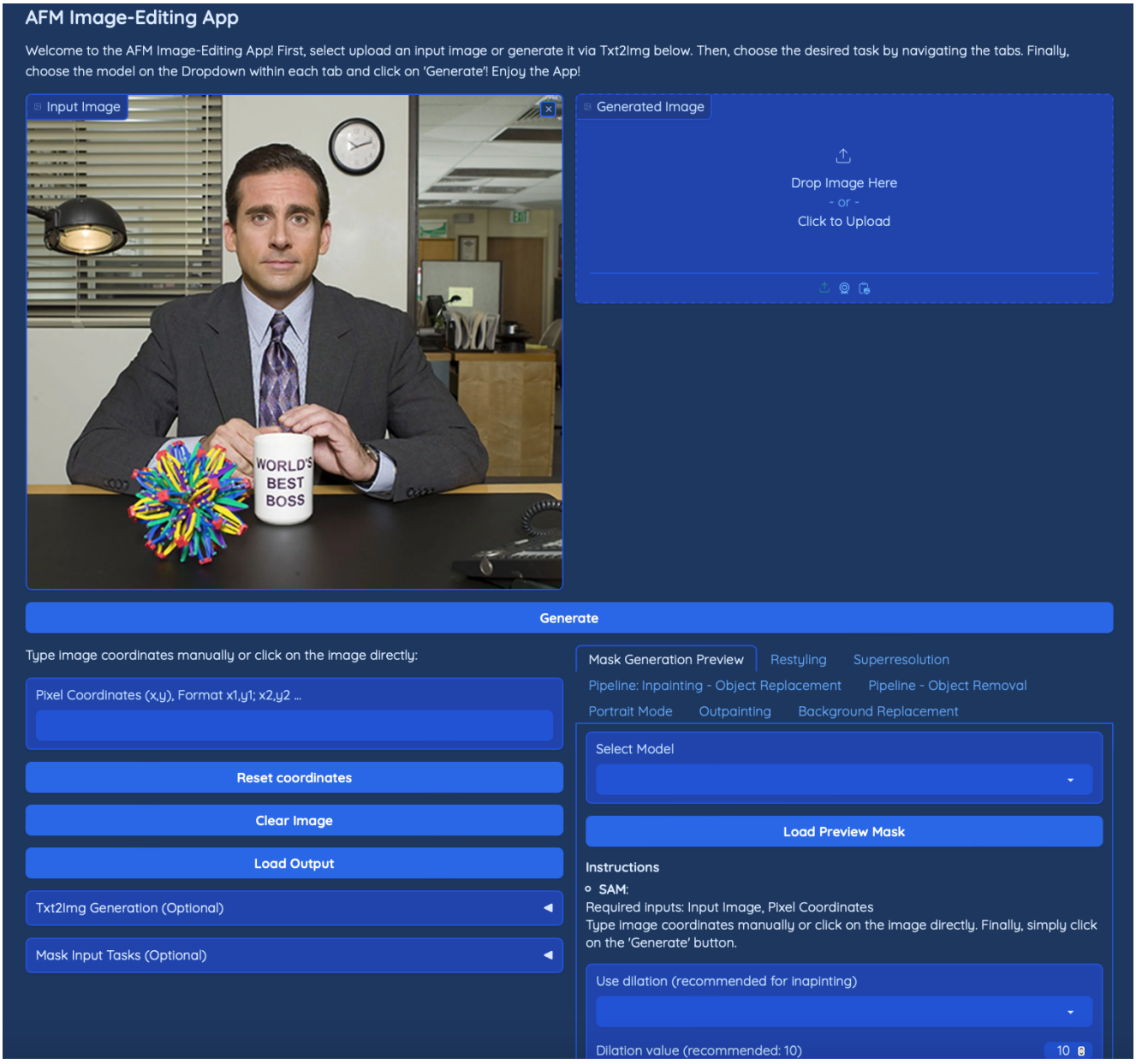


Figure 1. Gradio App

B. Survey of the results

Qualitative comparison of the images generated using different models. Each section corresponds to a different functionality. The survey includes comparison of 4 different input images represented by different rows of the table, where as the columns are outputs of different models. All outputs for the respective input image were generated with the same input prompts and number of inference steps.

B.1. Background Blurring

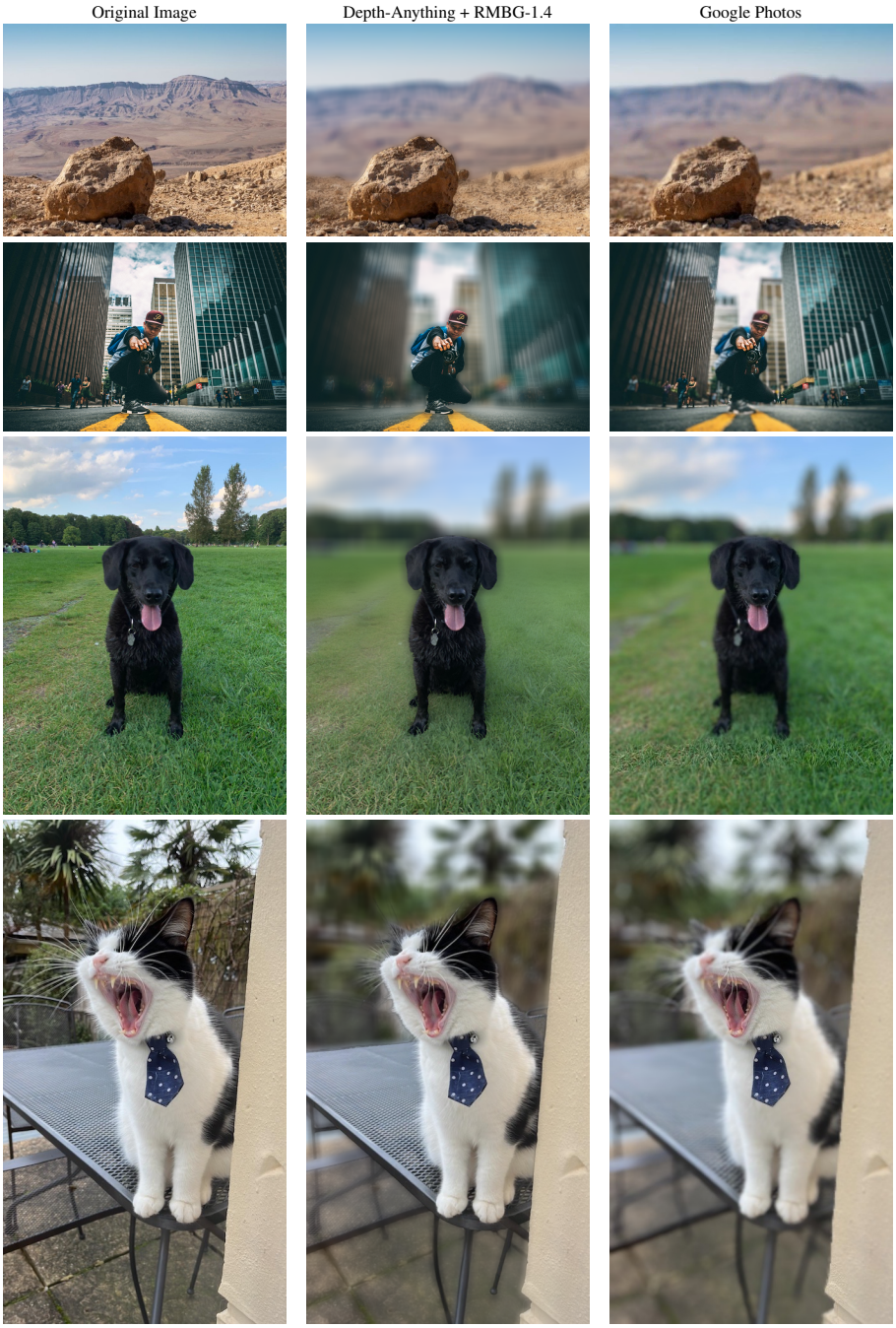


Table 2. Background Blurring Evaluation

B.2. Background Replacement

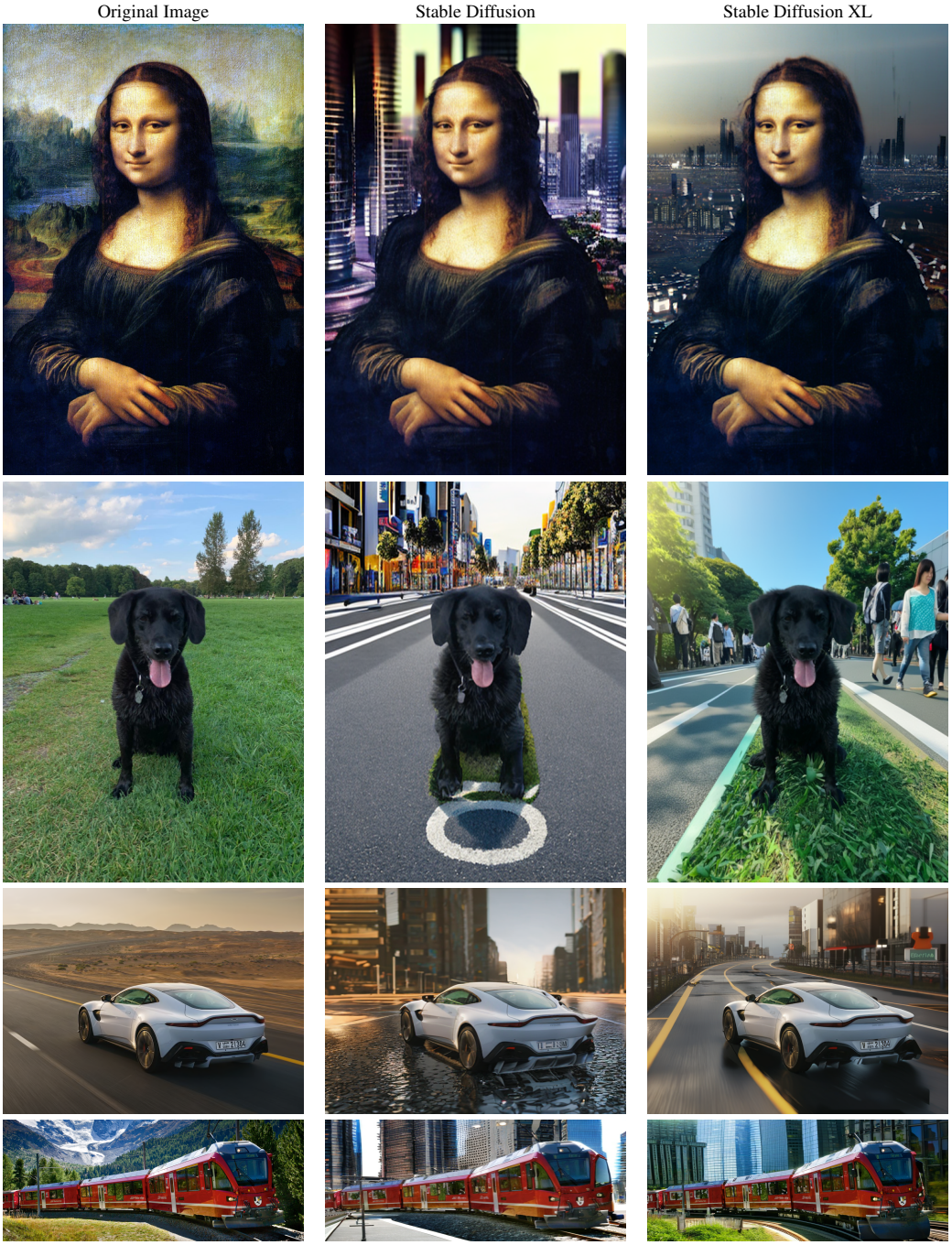


Table 3. Background Replacement Evaluation

B.3. Object Removal

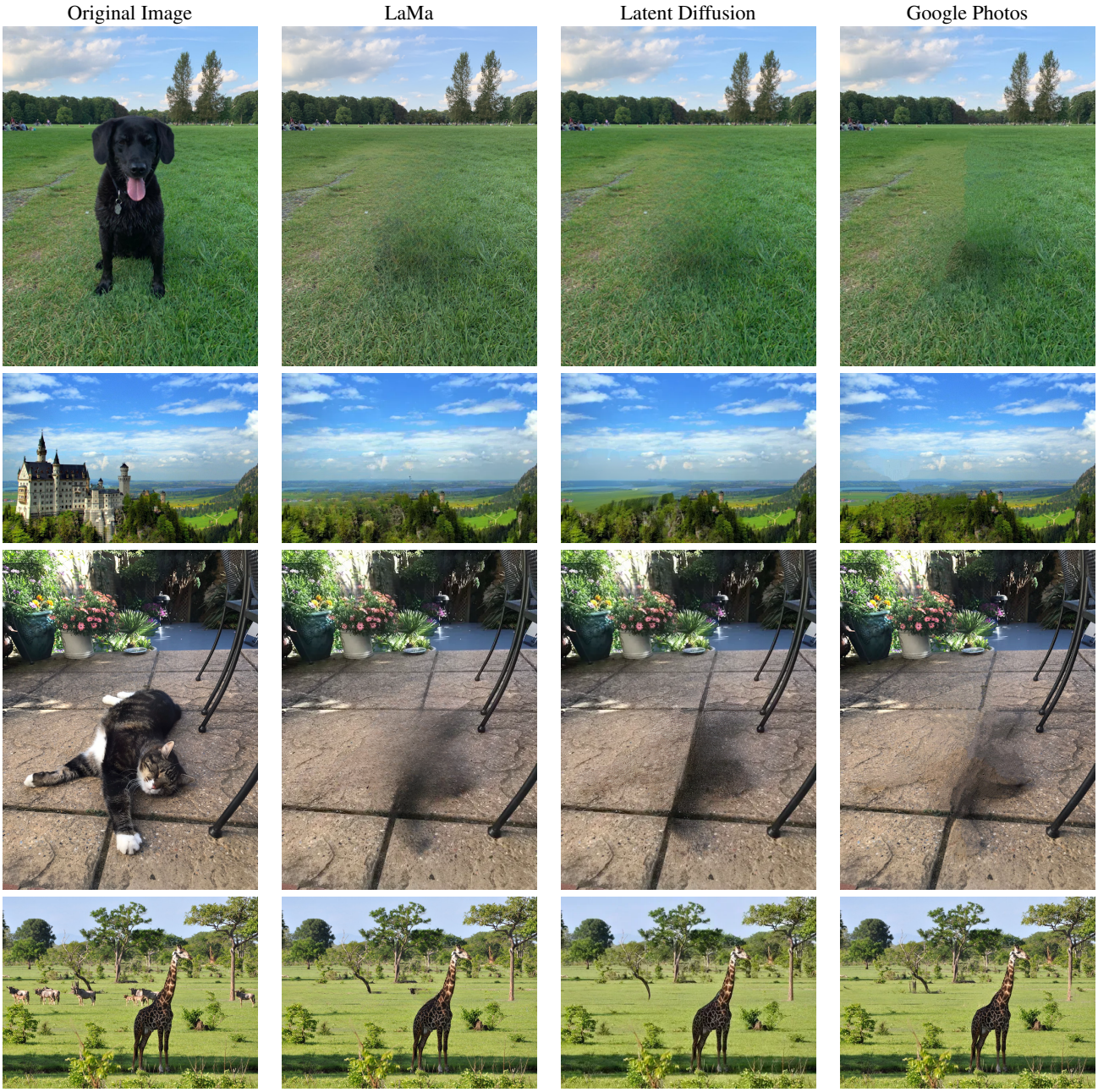


Table 4. Object Removal Evaluation

B.4. Inpainting

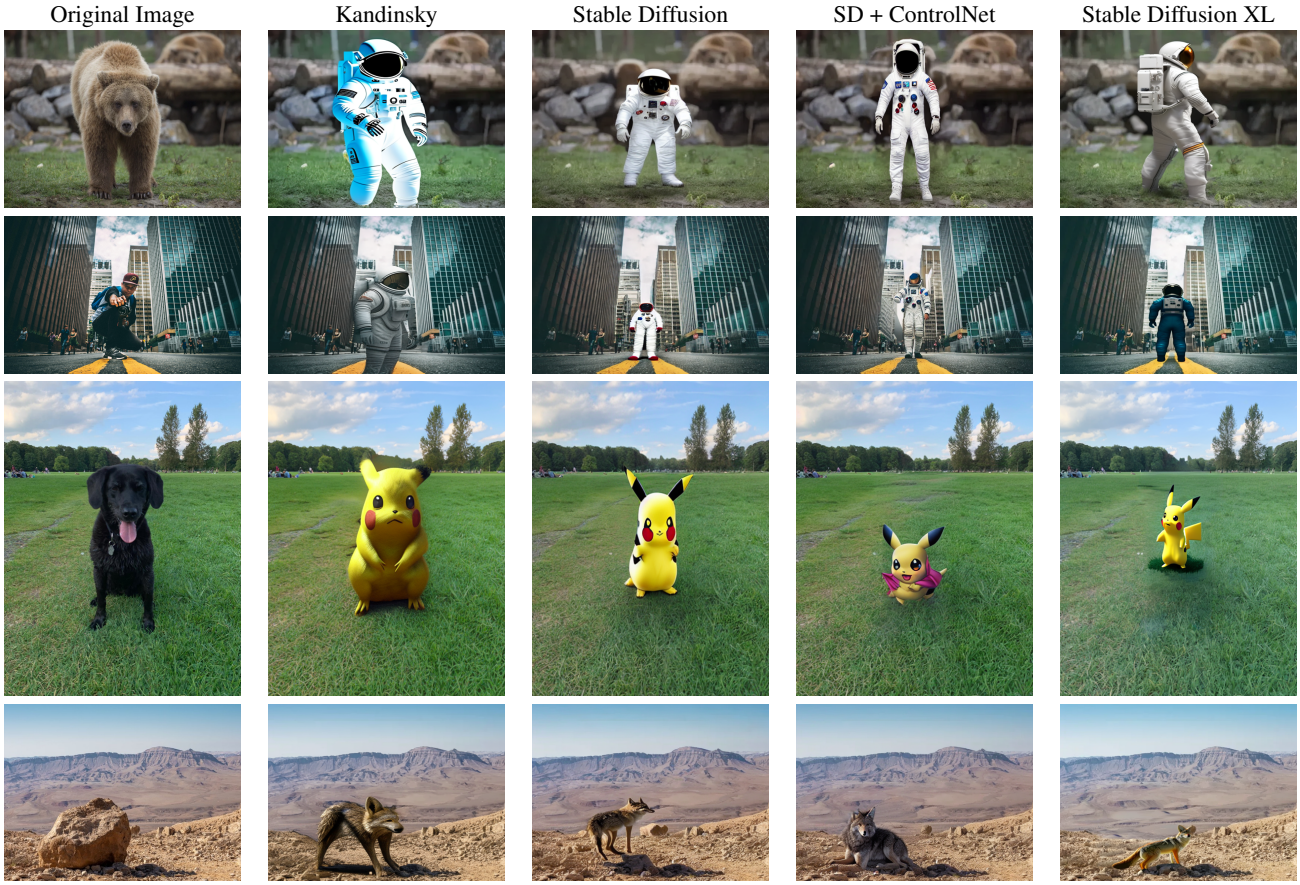


Table 5. Inpainting Evaluation

B.5. Restyling

Table 6. Restyling Evaluation

B.6. Outpainting

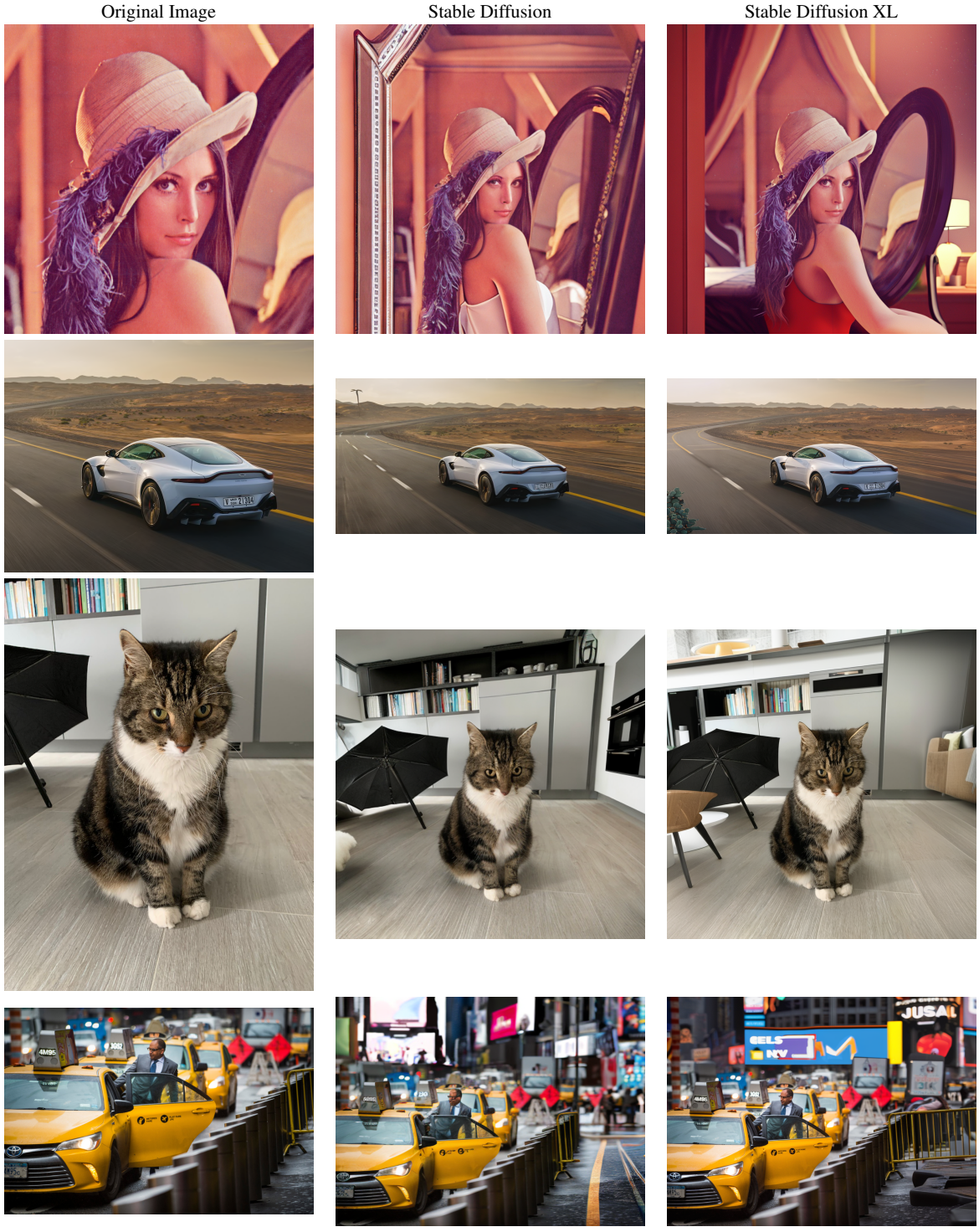


Table 7. Outpainting Evaluation

B.7. Superresolution

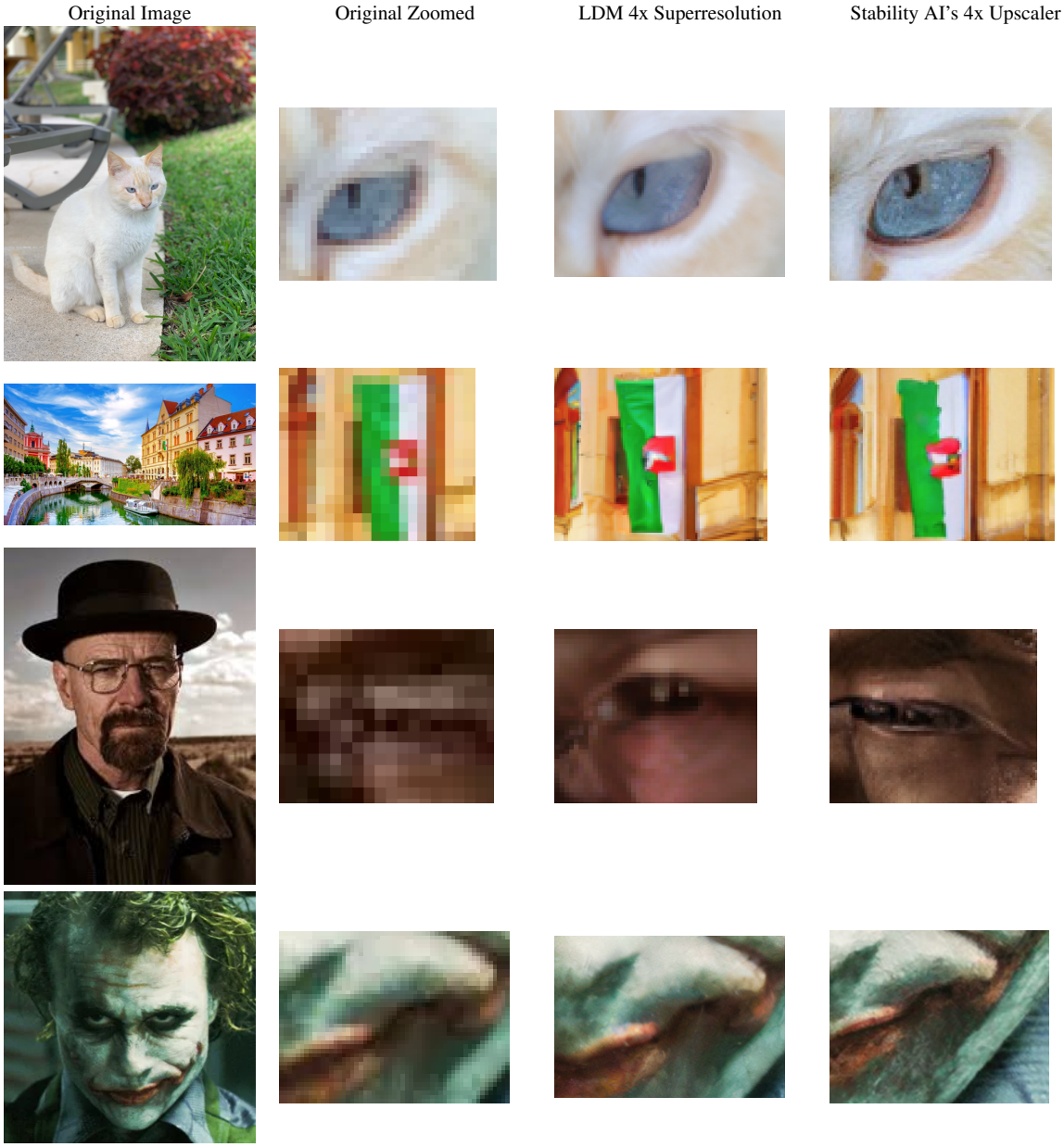


Table 8. Superresolution Evaluation

C. Generative Parameters

Table 9 presents a qualitative comparison of images generated using various guidance scale and strength parameters with Stable Diffusion v1.5. The table rows correspond to different strength values from the set $\{0.5, 0.75, 0.9\}$, while the columns correspond to values of the guidance scale from $\{2.5, 7.5, 15, 50\}$. All images were generated with the same (negative) prompts, inference steps, and a fixed seed generator to eliminate randomness, allowing us to attribute differences solely to the variations in guidance scale and strength parameters. The artistic style specified in the prompt was "comic strip", which becomes more pronounced as the parameter values increase.

Upon examining the images, we observe how these parameters influence the output. Higher values of strength and guidance scale result in outputs that deviate more creatively from the original image, displaying a more pronounced artistic effect. Conversely, lower values produce outputs that closely resemble the input image, seen on Figure 2. Specifically, the image generated with a guidance scale of 2.5 and a strength of 0.5 is nearly identical to the original, demonstrating minimal artistic alteration.

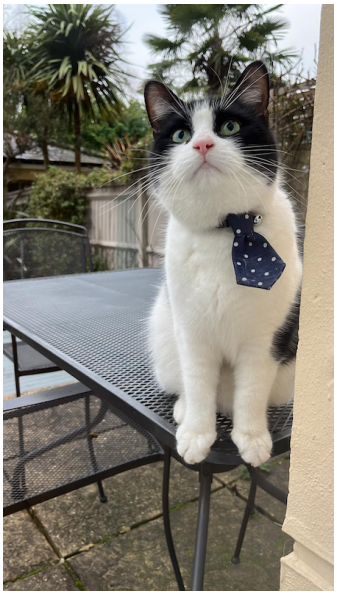


Figure 2. Original Image for Generative Parameters Study

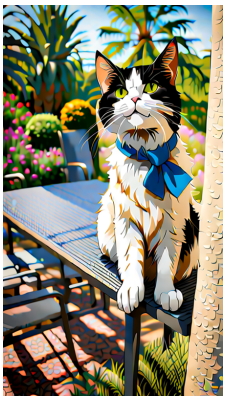
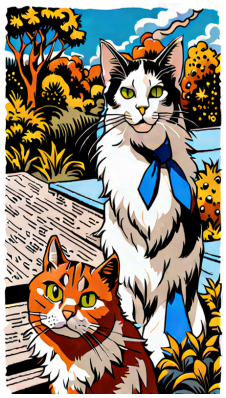
Guidance Scale\Strength	0.5	0.75	0.9
2.5			
7.5			
15			
50			

Table 9. Generative Parameters Evaluation