Diffusion-Based Content Generation for Augmented Reality

Marwan Mashra* Dávid Maruscsák† Christian Sandor‡

Université Paris-Saclay / CNRS
Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)

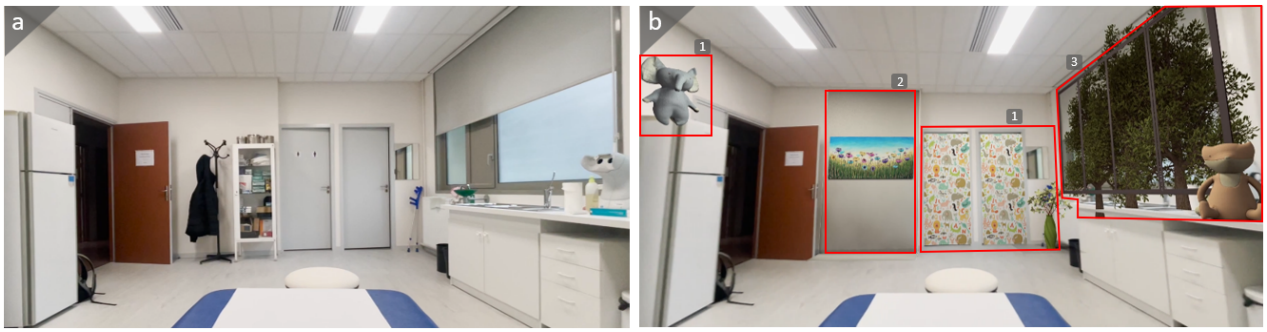

Figure 1: Two frames from a video taken in a hospital. (a) is a frame from the original recorded video. (b) is a frame after we applied our approach to generate (1), erase/hide (2), and replace (3) multiple visual elements.

Abstract

Text-conditional image synthesis has seen remarkable progress as of the year 2022, notably with the introduction of diffusion models like DALL·E-2, and Stable Diffusion. In this paper, we investigate the possibility of applying these models to Augmented Reality and discuss how to overcome their limitations. We propose an approach that leverages text-conditional generative models to allow content generation for AR and demonstrate an application in which we can generate, replace, and erase objects from the surrounding environment. This approach could be used in different scenarios, such as making the surrounding environment more friendly in hospitals, or other environments.

Index Terms: I.3.3 [Computer Graphics]: Picture/Image Generation—Image Generation for Augmented Reality; H.5.m [Information Interfaces and Presentation]: Miscellaneous

1 Introduction

Text-to-image generative models, or text-conditional image generative models, are AI models capable of generating images based on text input. Diffusion models [2] [3] [8] represent currently the state-of-the-art in that field [1]. Since early 2022, a lot of progress has been made in the field of text-to-image synthesis, and different generative models with fascinating results have seen the light. Starting with the release of DALL-E 2 [5] by OpenAI, followed by *Imagen* [6] by Google, and finally, the open source model *Stable Diffusion* released by Stability.ai at the end of August.

The fast progress in this field, along with the realistic and highly detailed images that are generated, have succeeded at attracting public attention to these AI models and their capabilities. We started seeing professionals from different fields, such as artists, graphic designers and photographers, making use of these AI models in various tasks that require creativity. We believe that these AI models could be utilized in even more challenging fields, such as AR.

Text-to-image generative models can offer many benefits for the field of AR. For instance, an unlimited number of unique 2D assets can be generated from text input, to be then used in AR. This allows end users to directly create, with minimal effort, new assets with the specific description and exact desired style. Furthermore, by leveraging text-conditional image inpainting techniques, which consists of editing a specific area in an image given a text input, we can not only generate new elements, but also erase existing elements, or even replace them with new generated elements.

Despite all these significant advantages, the application of diffusion-based generative models to Augmented Reality represent numerous limitations that we investigated during this work. The main limitations are the followings :

1) These diffusion models are currently limited to generating 2D images. Although there is some research showing promising results in 3D point cloud synthesis using diffusion models [10], all currently available state-of-the-art models are limited to 2D image generation.
2) The average inference time of these diffusion models ranges from 2s to more than 30s, depending on the model, settings, and available resources. Therefore, applying them to each frame is not an option. Furthermore, distributing the computation over several GPUs to accelerate the sampling process is very hard, because a diffusion model uses a Markov process during sampling. A common and promising method to reduce inference time in AI models is through Model Distillation [7]. As of the time of writing this paper, no distilled versions of these generative models have been announced.

In this paper, we explore the possibility of applying diffusion-based generative models to AR. Our contributions are:

- We investigate the limitations of this new type of generative models when applied to Augmented Reality.
- We design an approach to overcome these limitations, allowing us to add, erase, or replace elements from the environments.
- We illustrate a use case in which we can make the surrounding environment more friendly in hospitals (Figure 1).

*e-mail: marwan.mashra@universite-paris-saclay.fr
†e-mail: david.maruscsak@universite-paris-saclay.fr
‡e-mail: christian.sandor@universite-paris-saclay.fr

2 EXPERIMENT

We designed and implemented an approach that allows to apply diffusion-based text-to-image generative models to Augmented Reality, and demonstrated its ability to make the environment more friendly in hospitals.

2.1 Methods

Since the inference time of these generative models exceeds several seconds, they can't be applied to each frame in real-time. Therefore, after studying different alternative solutions, we decided to generate each visual element once, and then overlay it on the environment using motion tracking. This allows computation to be more efficient and avoid regenerating the same object for different frames.

To generate visual elements, instead of using text-conditional image generation, we used a slightly different approach called text-conditional Image Inpainting [4]. Image Inpainting is a process where we mask a certain area of the image, and the AI model uses the rest of the image to complete the masked zone. In text-conditional image inpainting, both the text input and the rest of the image will be used to generate the masked area. In our case, this technique allows the generated elements to fit better in the scene in terms of size, lighting, colors saturation, and perspective. It also enables us to use the same AI model, with minor changes in the post-processing, to erase or replace existing objects in addition to generating new ones.

After the generation, we extract the generated part of the image (previously masked area), and forward in through a post-processing pipeline to be then overlaid in the environment. This pipeline consists of 2 main steps :

1) **Upscaling**: the output of generative models is usually a low resolution images. A lot of generative models, including the one we're using, already contain some upscaling techniques to increase the resolution and sharpness. Since our goal is to display these generated elements in the environment, we decided to add another upscaler to hide small imperfections in the generated elements and increase the resolution even farther, allowing for more detailed and higher resolution output.
2) **Background Removal**: we use image segmentation to remove the background of the generated part of the image, which allows us to extract the generated element. We found that removing the background and using only the extracted generated element offers several advantages in our case. It makes overlaying elements in the environment easier and more tolerant to any potential misalignment, which results in a more realistic outcome. This doesn't hold in the case of erasing existing objects from the environment, where we generate a background to project over and hide elements rather than generating a specific element. Therefore, background removal isn't applied when erasing existing elements.

In summary, this approach consists of the user selecting an area in the video and entering a text prompt. A diffusion-based text-conditional image inpainting model will then use the unmasked area of the image to complete (generate) the masked one, while respecting the given text input. Post-processing is then applied to that generated area, upscaling it to a higher resolution and removing the background. Finally, the generated element is overlaid on the chosen area of the environment, and more content can be generated.

2.2 Experimental Setup

For this experiment, we illustrate the ability of our approach to make the environment in a hospital more friendly, adding, erasing, and replacing elements. For the diffusion-based text-conditional generative model, we use the open source model Stable Diffusion Inpainting that was initialized with the weights of the Stable Diffusion v-1-2. We generate images of size 512×512, with $guidance\_scale = 7.5$ and $steps = 50$ and 16-bit floating point for fast sampling on the GPU. The inference time of the model with these settings is around 3 seconds on an Nvidia RTX 3080 GPU according to our tests. For the upscaler, we use Real-ESRGAN, an improved version of ESRGAN [9] which is a GAN-based upscaler. We use Rembg, an open source background removal tool. Finally, for overlaying the elements in videos and real-time motion tracking, we use TouchDesigner, a node based visual programming language for real time interactive multimedia content by Derivative.

2.3 Results

The results of our experiment are illustrated in Figure 1 that represents two frames taken from the video of a hospital on which we used applied generative models to generate, erase, and replace several elements. We see that certain properties of the generated elements, such as the perspective, the lighting, the colors saturation, the size, fit well in the scene thanks to the use of Image Inpainting. We can also notice that medical equipment was well hidden or replaced in the scene. The modified scene appears more colorful, and friendly.

Nevertheless, we observe that some generated elements (like the toys in the scene) are lacking details, and show some shape imperfections. Others, like the trees on the windows, are more challenging to blend well into the scene. It's clear that the quality of the results is heavily depended on, and mainly limited by, the generative model used and their capabilities.

3 CONCLUSIONS

In this paper, we investigated the limitations of applying diffusion-based text-conditional generative models to Augmented Reality, and presented our approach to overcome these limitations using Image Inpainting. In the future, as text-to-image generative models continue to improve, their applications in domains like Augmented Reality and Virtual Reality will expand. For instance, an anticipated break through in the text-conditional 3D shape generation will enable even more real-world applications for these generative models in such challenging fields. The ability to wear Augmented Reality glasses and modify your surrounding environment using only text input (or alternatively voice if combined with speech recognition) can have numerous impactful use cases in our life.

REFERENCES

[1] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis, 2021. doi: 10.48550/ARXIV.2105.05233

[2] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020. doi: 10.48550/ARXIV.2006.11239

[3] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022. doi: 10.48550/ARXIV.2207.12598

[4] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022. doi: 10.48550/ARXIV.2201.09865

[5] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022. doi: 10.48550/ARXIV.2204.06125

[6] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. doi: 10.48550/ARXIV.2205.11487

[7] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models, 2022. doi: 10.48550/ARXIV.2202.00512

[8] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2020. doi: 10.48550/ARXIV.2010.02502

[9] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. doi: 10.48550/ARXIV.1809.00219

[10] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis. Lion: Latent point diffusion models for 3d shape generation, 2022. doi: 10.48550/ARXIV.2210.06978