Unrolling Virtual Worlds for Immersive Experiences

Anonymous Author(s)
Affiliation
Address
email

Abstract

This research pioneers a method for generating immersive worlds, drawing inspiration from elements of vintage adventure games like Myst and employing modern text-to-image models. We explore the intricate conversion of 2D panoramas into 3D scenes using equirectangular projections, addressing the distortions in perception that occur as observers navigate within the encompassing sphere. Our approach employs a technique similar to "inpainting" to rectify distorted projections, enabling the smooth construction of locally coherent worlds. This provides extensive insight into the interrelation of technology, perception, and experiential reality within human-computer interaction.

1 Motivation

In the field of human-computer interaction, the concept of immersive technologies has a rich history, dating back to the 1830s with the creation of the first stereoscopes during the early days of photography. We will understand the term immerse as "Technology that blurs the line between the physical, virtual, and simulated worlds, thereby creating a sense of immersion"[6].

These technologies have evolved to serve many purposes, acting as mediums for education, psychotherapy, physiotherapy, interactive simulations, and entertainment [8]. The phenomenon of immersion is also well-established in modern philosophy, which only increases researchers' interest in this topic [5, 7]. One of the key and most advanced areas of immersive technologies is virtual environment creation. The advent of virtual reality headsets has enabled users to attain remarkable levels of immersion in virtual worlds, even fostering a market area for startups focusing specifically on world creation [1].

Inspired by vintage games like Myst [3], where immersion into the world was achieved through interconnected scenes creating a coherent world, we propose a novel method for developing consistent environments. This approach combines contemporary text-to-image models with stereometric transformations to innovate environmental generation, presenting a sophisticated strategy for crafting immersive spaces.

2 Approach

Despite their primary function being text-to-image conversion, modern text-to-image models often feature additional modes, such as text+image -> image. We utilized the fine-tuned StableDiffusion v1.5 model [2], capable of generating panoramas in equirectangular projection. This projection, a 2x1 rectangle, can be seamlessly converted into a spherical panorama by using open-source tools [4].

Thus, we can generate our initial scene using a prompt with an environment description. This spherical panorama already imparts a sense of being inside the world. The phenomenon of immersion within the panorama is well-known and heavily utilized by various entities, including game developers and museums, as a means to situate the viewer within a specific context and environment.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

(a) Initial panorama (b) Panorama mapped on sphere (c) View inside the sphere

Figure 1: Generating scene from the panorama in equirectangular projection.

This immersive magic begins to unravel as we move in any direction away from the center. If we do so within this static spherical panorama, we will observe visual distortion, altering our perception of the panorama.

We have derived a formula to obtain this distorted image following the viewer's movement (can be found in the Appendix). In Figure 2, the blue sphere represents the current scene, and the red sphere represents the new expected panorama after movement from the old center k to the new center l.

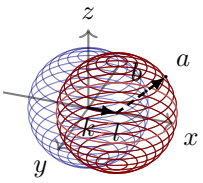


Figure 2: Movement within the scene generates a new expected panorama to maintain the feeling of immersion.

For every point "a" on the surface of the new sphere, we can find an intersection point "b" on the old sphere, determined by the radius from the new sphere center to point "a". Since point "a" represents a pixel in a new 2D panorama, we will use the pixel representing point "b" in the previous panorama to determine the color of this pixel. This is how we obtain the distorted panorama image.

After obtaining the distorted projection, we can remove this distortion using the model from the initial scene generation step, along with the initial prompt. Modern neural networks tend to "reconstruct" inputs, thereby enhancing their likelihood. We attribute this phenomenon to the networks' ability to learn the manifold of plausible objects, grounding various noisy or distorted objects onto this manifold and projecting them to the nearest suitable region. This can be interpreted as a form of spontaneous denoising.

In this methodology, the distorted projection is passed through the network utilizing the same text prompt but coupled with the reconstructed input image from the preceding step. This technique is essentially similar to "inpainting"; however, its application in this unique form for amending distorted projections is a novel exploration in our study. Pairs of the distorted and restored images can be viewed in Figures 4 and 5 in the Appendix.

Transitions between neighboring scenes occur smoothly and seamlessly, despite a noticeable pattern of accumulating errors and hallucinations over iterations. To immerse in a fully realized world, drawing inspiration from old adventure games, a grid of scene-image-panorama generations can be created. Examples of such worlds can be found in the demo [1] and in Figure 6 in the Appendix.

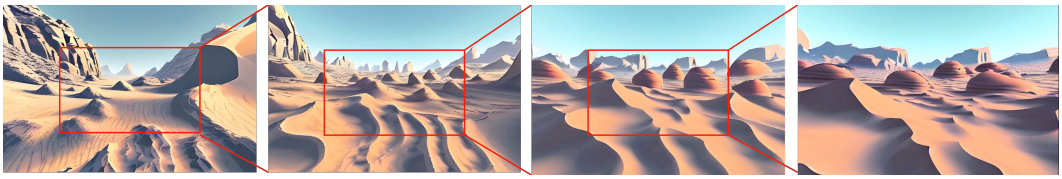


Figure 3: Example of a grid of scenes with forward movement between them. When combined, such scenes create an immersive experience of movement in the virtual world.

[1] Omitted to preserve anonymity

3 Ethical Implications

Ethical considerations in the realm of image generation hold paramount importance, particularly given the substantial implications of this technology. The ethical dimensions encompass a broad spectrum, particularly focusing on the potential utilization of such technologies to accentuate societal inequalities and proliferate visual representations entrenched in stereotypes and biases. The ramifications of deploying algorithms without careful consideration of their inherent biases and potentially harmful impacts can be profound, perpetuating existing disparities and possibly creating new ones.

References

[1] Blockade labs. URL `http://blockadelabs.com`. Accessed: 2023-09-26.

[2] Latentlabs 360. URL `https://civitai.com/models/10753/latentlabs360`. Accessed: 2023-09-26.

[3] Myst. URL `https://cyan.com/games/myst/`. Accessed: 2023-09-26.

[4] Pannellum. URL `https://pannellum.org/`. Accessed: 2023-09-26.

[5] Nick Bostrom. Are we living in a computer simulation? *Philosophical Quarterly*, 53(211): 243–255, 2003.

[6] Hyuck-Gi Lee, Sungwon Chung, and Won-Hee Lee. Presence in virtual golf simulators: The effects of presence on perceived enjoyment, perceived value, and behavioral intention. *New Media & Society*, 15:930–946, 09 2013. doi: 10.1177/1461444812464033.

[7] Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge, 1945.

[8] Ayoung Suh and Jane Prophet. The state of immersive technology research: A literature analysis. *Computers in Human Behavior*, 86, 04 2018. doi: 10.1016/j.chb.2018.04.019.

4 Appendix

Let's define our parameters as:

step The step of displacement, where 0 is no displacement, and 1 is displacement by the full radius.

direction The direction of displacement in the form of an X-axis angle when looking at the sphere from above, an angle from 0 to 359 degrees.

width / height The dimensions of the image.

To get the (x coordinate of the point "b", we'll use next formulas:

$$x_b = \left(\frac{\text{direction}}{360} \cdot \text{width} + \alpha \cdot \frac{\text{width}}{2}\right) \mod \text{width}$$

$$\alpha = \frac{\text{diff\_x\_norm} - \arcsin\left(0.5 \cdot \sin\left(\text{diff\_x\_norm}\right)\right)}{\pi}$$

$$\text{diff\_x\_norm} = \frac{2\pi\left(x - \frac{\text{direction}}{360} \cdot \text{width}\right)}{\text{width}}$$

To get the y coordinate of the point "b", we'll use next formulas:

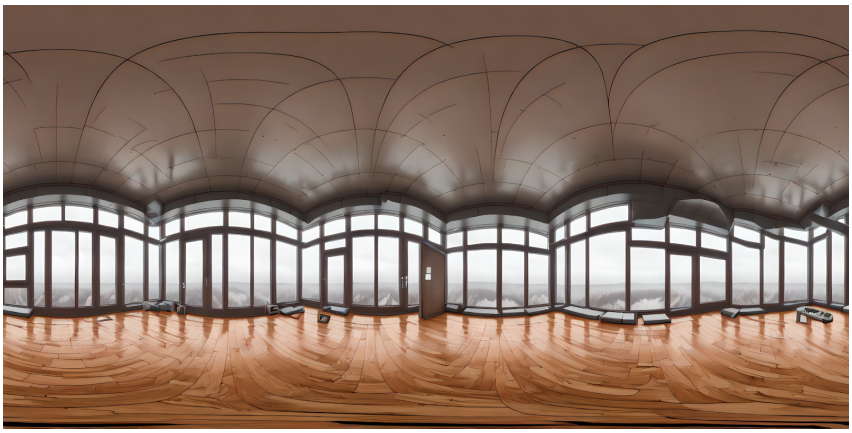
$$y_b = \frac{\text{width} \cdot \left(\frac{1}{2} + \frac{\beta}{\pi}\right)}{\pi}$$

We have two "β": one for the case when point "b" has crossed the zenith during the movement, and the second if it has crossed the nadir:

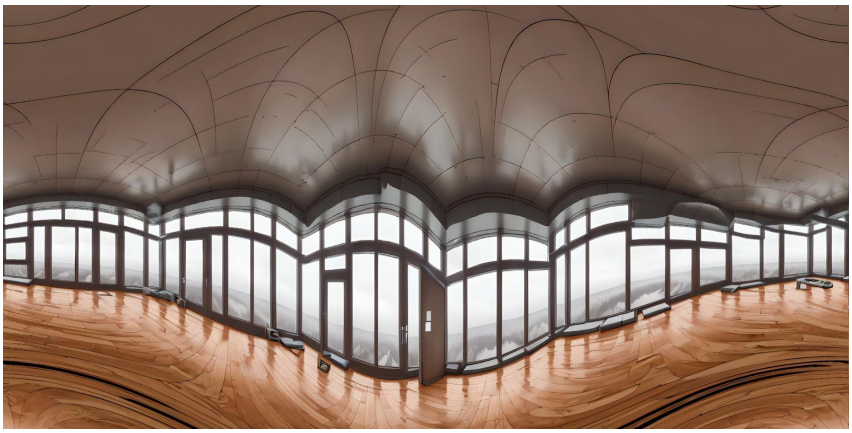
$$\beta_z = \text{sign}\left(\pi\frac{y}{\text{height}} - \frac{1}{2}\right) \cdot \left(\pi - \arccos\left(\frac{\text{step\_adjstep} \cdot \alpha - \cos\_\text{va}}{\sqrt{\text{step} \cdot \alpha^2 - 2 \cdot \text{step} \cdot \alpha \cdot \cos\_\text{va} + 1}}\right)\right)$$

$$\beta_n = \text{sign}\left(\pi\frac{y}{\text{height}} - \frac{1}{2}\right) \cdot \arccos\left(-\frac{\text{step} \cdot \alpha - \cos\left(\pi\frac{y}{\text{height}} - \frac{1}{2}\right)}{\sqrt{\text{step} \cdot \alpha^2 - 2 \cdot \text{step} \cdot \alpha \cdot \cos\left(\pi\frac{y}{\text{height}} - \frac{1}{2}\right) + 1}}\right)$$

$$\text{step\_adjusted} = \text{step} \cdot \alpha$$



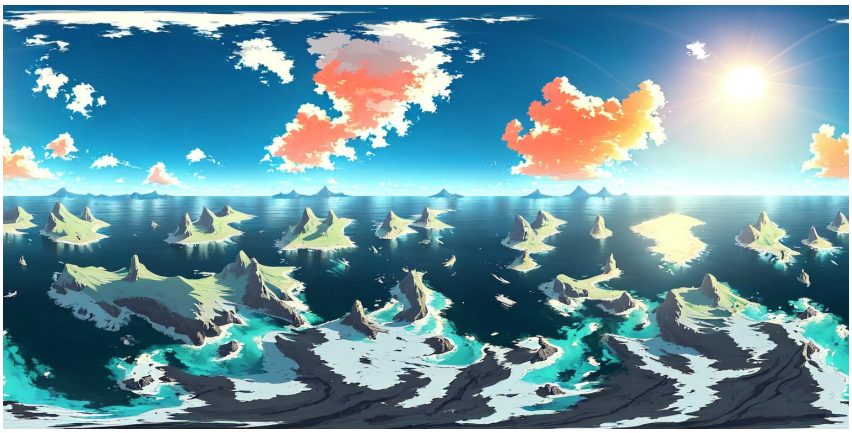
(a) Initial panorama



(b) Distorted panorama

(c) Restored panorama after distortion

Figure 4: Process of the panorama restoration.

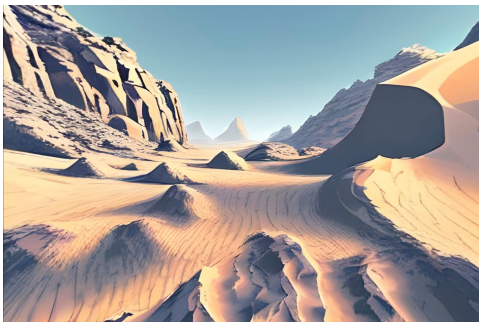


(a) Initial panorama

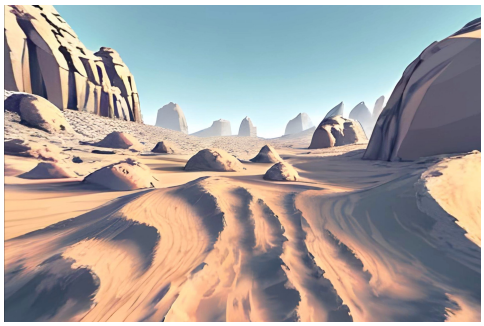
(b) Distorted panorama

(c) Restored panorama after distortion

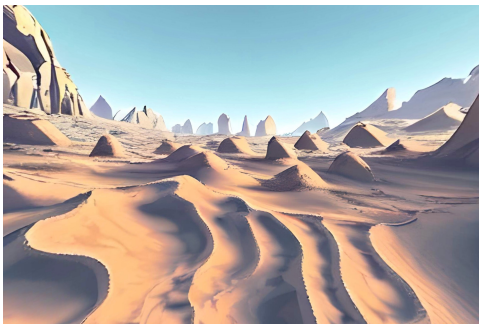
Figure 5: Process of the panorama restoration.



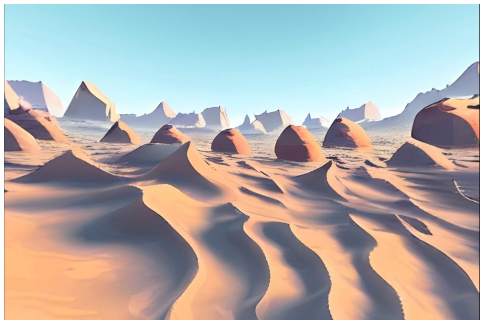
(a) Scene 1



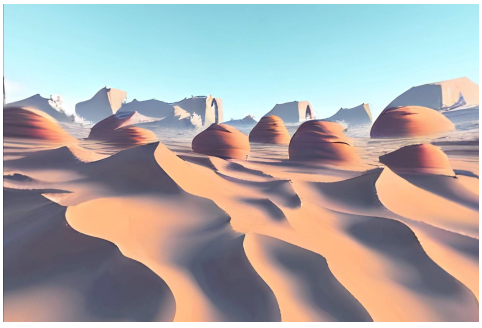
(b) Scene 2



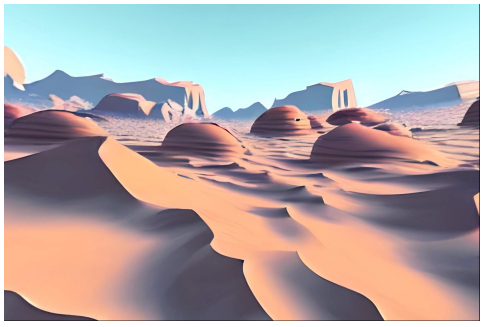
(c) Scene 3



(d) Scene 4



(e) Scene 5



(f) Scene 6

Figure 6: A simple desert world, which consists of 6 scenes with the structure 1->2->3->4->5->6.