Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2024)
IEEE Xplore Part Number: CFP24CV1-ART; ISBN: 979-8-3503-2753-3

2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 979-8-3503-2753-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDCIOT59759.2024.10467441

Voice - Driven Panoramic Imagery: Real-Time Generative AI for Immersive Experiences

Nirmala Venkatachalam
Department of Artificial Intelligence and Data Science
Easwari Engineering College
Chennai, Tamil Nadu
nirmala.research17@gmail.com

Mukul Rayana
Department of Artificial Intelligence and Data Science
Easwari Engineering College
Chennai, Tamil Nadu
aimukulrayana@gmail.com

Bala Vignesh S
Department of Artificial Intelligence and Data Science
Easwari Engineering College
Chennai, Tamil Nadu
balviky1310@gmail.com

Prathamesh S
Department of Artificial Intelligence and Data Science
Easwari Engineering College
Chennai, Tamil Nadu
prathamesh661274@gmail.com

***Abstract* — This research study introduces an innovative system that aims to synthesize 360-degree panoramic images in Realtime based on vocal prompts from the user, leveraging state-of-the-art Generative AI with a combination of advanced NLP models. The primary objective of this system is to transform spoken descriptions into immersive and interactive visual scenes, specifically designed to provide users with first-person field views. This cutting-edge technology has the potential to revolutionize the realm of virtual reality (VR) experiences, enabling users to effortlessly create and navigate through personalized environments. The fundamental goal of this system is to enable the generation of real-time images that are seamlessly compatible with VR headsets, offering a truly immersive and adaptive visual experience. Beyond its technological advancements, this research also highlights its significant potential for creating a positive social impact. One notable application lies in psychological interventions, particularly in the context of phobia treatment and therapeutic settings. Here, patients can safely confront and work through their fears within these synthesized environments, potentially offering new avenues for therapy. Furthermore, the system serves educational and entertainment purposes by bringing users' imaginations to life, providing an unparalleled platform for exploring the boundaries of virtual experiences. Overall, this research represents a promising stride towards a more immersive and adaptable future in VR technology, with the potential to enhance various aspects of human lives, from mental health treatment to entertainment and education.**

Keywords* — *Virtual Reality (VR), Panoramic Images, Voice commands, prompts, Generative Artificial Intelligence (AI), Natural Language Processing (NLP), Immersive Visual Scenes, Psychological Interventions, Phobia Treatment, Real-Time Synthesis, Adaptive Visual Experience, User Interaction, Personalized Environments, Virtual Experiences, VR Headsets, First-Person view.

I. INTRODUCTION

In the rapidly evolving technological landscape, research emerges as a conduit bridging the gap between science fiction and present-day realities, extending influence into human connection and collaborative endeavors. This research delves into voice-activated panoramic image synthesis, envisioning a future where individuals, regardless of geographical separation, collaborate seamlessly within a virtual domain that transcends physical boundaries. Imagine educators and students from different corners of the world exploring history through spoken words that bring immersive visuals to life. Architects, designers, and artists converge in a virtual space, overcoming cultural and linguistic barriers to create together. This research harnesses the potential of AI, NLP and VR [3],[5] to make these visionary ideas a reality, fostering global collaborations, artistic expression, and improved education. It promotes interaction, empathy, and knowledge exchange across cultures, addressing complex global challenges.

In therapeutic settings, it could revolutionize mental health treatment, and in education, students can explore historical sites and scientific frontiers through immersive narration. This research aims to unite a more empathetic global society, bound by immersive experiences. While navigating this transformative technology, the power of combining human voice and immersive visuals [5] to elevate understanding and connection. It's a frontier where imagination meets reality, promising boundless possibilities. In solidarity with the global research community, embarking on this journey to shape a world where the spoken word vitalizes the canvas of virtual reality, fostering thriving creativity, knowledge, and human connection to carve out a transformative tomorrow where immersive experiences transcend the digital realm, imprinting an enduring essence on the fabric radiates a sense of promise, foretelling a future enriched by understanding and unity.

II. Literature Review

The contribution of feature-based image stitching algorithms, Alomran, Murtadha Et Al [1] in panoramic photography, primarily lies in their ability to create seamless and high-quality wide-angle panoramic images. These algorithms have become integral to the functionality of 360-degree cameras and virtual reality applications, where the facility to enable the merging of multiple images into a single panoramic view. The key to their success is the efficient overlapping of images with common scenes, ensuring both speed and accuracy in processing. The technological advancements in these algorithms also allow for the mapping of panoramic images onto different projective layouts, such as spherical or stereographic, which significantly enhances the perception of depth and space in the resulting images. This capability is particularly important in creating more realistic and immersive 3D reconstructions and virtual experiences.

In recent advancements in 360° panoramic image processing, the SURF (Speeded-Up Robust Features) algorithm has undergone significant optimization to enhance feature detection and matching capabilities, Y.Yan Et Al [3]. Key issues addressed include uneven feature point extraction and low matching rates. Integration with efficient descriptors, such as BRIEF (Binary Robust Independent Elementary Features), has played a crucial role in improving performance. These enhancements have led to the development of more sophisticated image stitching methods, adaptable to a variety of conditions, including changes in rotation, zoom, and illumination. Such adaptability is essential in maintaining the high quality of stitched images in dynamic environments. The advancements ensure stability and integrity in panoramic images, broadening the application of this technology in various photographic and video graphic contexts, including virtual reality and detailed photographic surveys.

The transition in natural language processing for image description from traditional feature recognition methods to advanced deep learning models marks a significant evolution in the field. Deep learning, particularly through the use of CNNs and RNNs like LSTM networks, has transformed the capability of systems to extract detailed features from images and generate coherent natural language descriptions, Maurya, Himanshu Et Al [5]. This progression addresses the challenges previously faced with large datasets and complex image contexts. The current standards in prompt generation, influenced by these deep learning advancements, reflect a more nuanced and accurate representation of images in textual form. This evolution not only enhances the quality of image descriptions but also broadens the scope of applications in areas like assistive technology, content creation, and AI-driven analytics.

III. Methodology

A. Environment Setup

- Initialize either a local or cloud-based development environment for using Jupyter notebook instances for efficient Python code execution.
- For optimal performance in intensive tasks, configure notebook instances with high VRAM GPU support. Install the latest GPU drivers and Nvidia CUDA Toolkit (version 11 or higher) for smooth operation.

B. Installation of Dependencies

- Speech Recognition: Install SpeechRecognition and PyAudio.
- Image Processing: Use pip to get OpenCV-Python package for setting up the necessary image processing functions.
- Image Generation: Deploy Stable Diffusion Model 1.5 with required libraries.
- Image to Image Processing: Utilize StableDiffusion Img2ImgPipeline.
- VR Viewing: Incorporate JavaScript Extensions and Pannellum Plugin.

C. Speech Recognition Integration

- Use PyAudio for microphone input.
- Employ SpeechRecognition with Google Speech Recognition for accurate speech-to-text conversion.
- Correct the raw sentence with regular expressions and improve grammar and spelling using TextBlob.
- Present the corrected text for user interaction. Establish an iterative feedback loop to enhance accuracy. Incorporate adaptive correction for diverse speech patterns.

D. NLP model for prompt generation

Input Raw Text into GPT-J: Start with a readable and coherent text file from speech-to-text conversion.

- Post Process Text with Fine-Tuned GPT-J: Feed the text into a GPT-J model, specifically fine-tuned for image prompt generation.
- Generate Split Prompts Instruct GPT-J to divide the prompt into three parts for a panoramic image:
    i. Left Half of Image: Describes the left segment.
    ii. Center of Image: Details the central segment for seamless connection
    iii. Right Half of Image: Focuses on the right segment, complementing others.
- Ensure Continuity: Ensure GPT-J maintains thematic and visual continuity across the three prompts.
- Finalize Prompts Review and adjust the prompts from GPT-J for alignment and continuity in the panoramic image.

E. Image Generation with Stable Diffusion Model

- *Text to image:*

a) Utilize Stable Diffusion 1.5 Model on a CUDA-enabled GPU for fast and precise image generation from descriptive prompts.

b) A generator with a manual seed ensures image generation consistency and reproducibility. This pre-trained model converts text into visually compelling images, offering accuracy and reproducibility.

F. Panoramic Image Synthesis

- Crop and Resize Strategy: Utilizing computer vision, enhance Stable Diffusion-generated Image1 for panoramic continuity. Apply a 40% right-side crop for optimization, followed by automatic resizing to 512x512 dimensions (Higher resolution requires more GPU VRAM > 24GiB) .
- Optimized Input for Image-to-Image Model: Tailored through careful cropping and resizing, this input enhances visual coherence, enabling a seamless and pleasing panoramic sequence.
- This integration enhances the immersive nature of the panoramic experience, making it responsive to the user's real-time interactions and environmental changes.

G. Panoramic Image-to-Image Transformation

- A Sophisticated Pipeline for Multi-Perspective Image Generation. In this advanced pipeline, the output from the Stable Diffusion text-to-image process is harnessed. Progressive prompts from the NLP model guide this pipeline to create a rich array of images with various view perspectives.
- Including alternative perspectives like left and right views. This pre-trained model excels at turning textual descriptions into captivating images, now with the added dimension of varied viewpoints, expanding the possibilities of visual story.
- Displaying flexibility, the system ensures images adjust seamlessly to the unfolding narrative, enhancing the flow and consistency of the visual story. Intelligent adjustments based on contextual cues elevate storytelling possibilities, providing a more immersive and personalized visual journey for users.

H. Image Stitching Process

- Utilize OpenCV's image stitching algorithms to merge the series of images produced ensuring a seamless panoramic view.
- Implement feature detection and matching techniques to identify overlapping areas between consecutive images for accurate alignment and stitching.

I. Fit to Panorama

- Leveraging the capabilities of the Pillow library in Python, images are converted into equirectangular format, a crucial step for panoramic viewing. This involves stretching the image to cover a 360-degree view, ensuring compatibility with immersive experiences such as virtual reality or panoramic image viewers.
- The resulting equirectangular images contribute to a visually expansive and seamless exploration of digital environments. To showcase the transformed image, it can be viewed in a Pannellum viewer, allowing for an interactive and immersive visual experience.

J. *Image Refinement*

- Introduce SMAA (Subpixel Morphological Antialiasing) or FXAA (Fast Approximate Antialiasing) techniques to the pipeline to smooth out textures and reduce the appearance of sharp edges in the stitched panoramic image.

K. *VR Headset Compatibility*

- Addition of a web based visual display for testing, a VR headset can directly support web GUIs for Virtual Simulation
- Adapt the final panoramic image for VR headset display, ensuring that the image format and resolution are compatible with common VR hardware.
- Test the immersive experience with different VR headsets to validate the quality and responsiveness of the panoramic visualization in a real-time setting.

IV. ARCHITECTURE

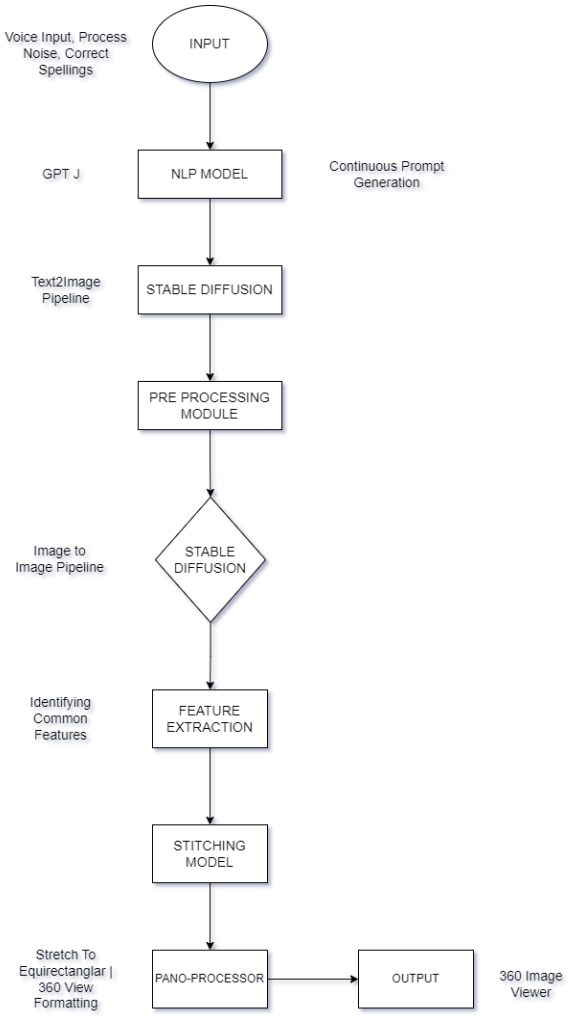


Fig 1: System Design

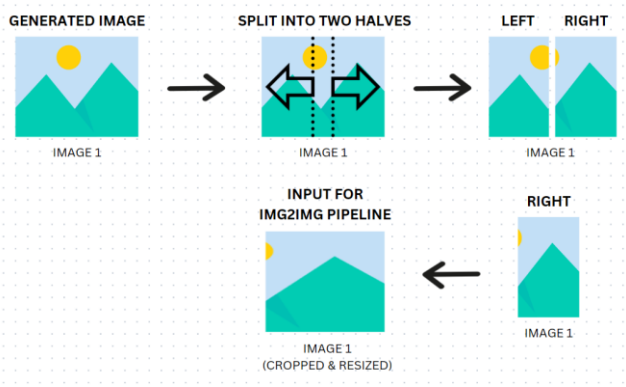


Fig 2: Img2Img Pre-processing Module

V. LIMITATIONS

1. *GPU and VRAM Requirements:*

Requiring significant computational resources, a GPU with 24GB VRAM or higher is essential for optimal performance in upscaling tasks and enhancing Image resolution. The current 6GB GPU, while functional, lacks capacity for tasks demanding higher resolution and clarity.

2. *Iterative Generation:*

The iterative process of image generation has certain challenges, given the current learning capacity of generative AI models. Achieving a seamless and continuous image generation involves manual adjustments of weights through trial and error. This highlights the necessity for advancements in model comprehension and automation to streamline the iterative generation process.

VI. RESULTS AND DISCUSSION

Recent research developments reveal significant progress in image processing. Commencing with the integration of real-time voice input, the system seamlessly transcribes prompts while correcting errors on the fly for a smooth communication interface.

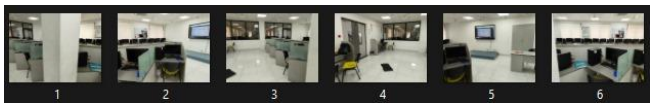
Expanding capabilities, the framework introduces thematic prompt transcription, translating raw inputs into immersive scenarios. For example, in therapeutic applications, assuming a case with a person having hydrophobia - a straightforward prompt like "Looking at a waterfall" dynamically generates multidirectional prompts, situating the waterfall in the center, a forest to the right, and hills to the left.

Venturing further, the research explores image stitching through feature extraction, enabling the creation of continuous images for panoramic scenes. This meticulous stitching process combines diverse angles captured through a mobile camera to form a unified and expansive panorama. A significant milestone includes the deployment of Pannellum, a 360-image viewer tool seamlessly presenting panoramic images. Beyond its visual achievements, this

viewer effortlessly integrates with VR devices, offering users an immersive experience that transcends conventional boundaries. This enables a paradigm shift in visual communication and immersive experiences. The convergence of advanced technologies and practical methodologies reshapes the landscape and opens new horizons in virtual reality and therapeutic environments. The precision and efficacy demonstrated in the image processing framework underscore its potential to craft immersive visual narratives and enhance the quality of virtual tours, representing a transformative chapter in visual communication.

Tested using Camera:

Input:



Output:

Fig 3: Given Input and Final Output Images

Tested using Generative AI:

- Input prompt - “Inside of a Forest Near a Water Fall”
- NLP Response:

  i. A panoramic view of a waterfall's left segment, surrounded by lush greenery and tall trees. Water cascades down rocky outcrops, creating mist. Dappled sunlight filters through foliage.

  ii. A Panoramic right segment of a waterfall flowing into a river with a beautiful forest having pitch dense and green trees.

1. Generated Image using base prompt as the left half of Fig 4.

2. Generated the continuous scene using a progressive prompt – with set strength value as 0.85. (Strength value was adjusted based on trial and error as shown on the right half of Fig 5.

Fig 4: Generated Image samples

An interactive virtual display to view the panoramic image using a 360-image viewer pannellum as shown in Fig 6.3.

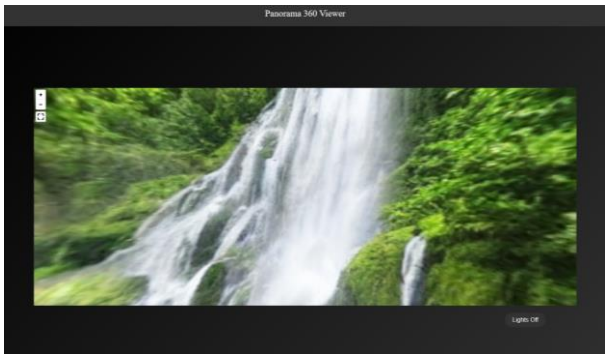


Fig 5: Panoramic View

VII. CONCLUSION

In conclusion, this research study pioneers the synthesis of real-time 360-degree panoramic images through vocal prompts, employing cutting-edge Generative AI and advanced NLP models. The system seamlessly transforms spoken descriptions into personalized visual scenes, offering first-person field views. With compatibility for VR headsets, this technology promises an adaptive and truly immersive visual experience. Beyond technological advancements, the system holds societal promise, particularly in therapeutic applications like phobia treatment. Patients can navigate and confront fears within synthesized environments, opening new avenues for therapy. Additionally, the system serves educational and entertainment purposes, providing a unique platform for exploring virtual experiences. This promising stride towards a more immersive future in VR technology has the potential to enhance mental health treatment, entertainment, and education. The research marks a significant step forward in shaping a future where immersive experiences seamlessly integrate into daily lives.

VIII. FUTURE ENHANCEMENT

1. VR Simulation Integration:

Elevate the immersive experience by integrating with multiple VR devices for comprehensive compatibility. While the current Pannellum plugin showcases promising work, expanding support to various VR platforms ensures a more inclusive and user-friendly VR optimization. Users can explore panoramic environments seamlessly, making the system adaptable to a wide array of virtual reality hardware such as Oculus quest, Apple Vision Pro, etc.

2. Depth Addition for Immersive Exploration:

Enhance the panoramic image processing by incorporating depth. Introducing depth to the images provides a more immersive experience, allowing viewers to analyze the virtual field with a heightened sense of perspective. This addition of depth enriches the overall visual perception, creating a more engaging and realistic virtual environment.

3. Generative Image-to-Video Model:

Introduce a generative image-to-video model to the system, enabling the creation of short video clips from a given image. This enhancement goes beyond static images, providing a dynamic and immersive feel to the viewer. By generating motion for virtual objects and environments, users can now experience a live and interactive virtual space, adding another layer of engagement to the panoramic exploration.

IX. REFERENCES

[1] Alomran, Murtadha, and Douglas Chai. "Feature-based panoramic image stitching." In 2016 14th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 1-6. IEEE, 2016.

[2] Ha, You-Jin, and Hyun-Deok Kang. "Evaluation of feature based image stitching algorithm using OpenCV." In 2017 10th International Conference on Human System Interactions (HSI), pp. 224-229. IEEE, 2017.

[3] Gunkel, Simon NB, Marleen DW Dohmen, Hans Stokking, and Omar Niamut. "360-degree photo-realistic VR conferencing." In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 946-947. IEEE, 2019.

[4] Y. Yan, Y. Ren and X. Zhou, "Research and Application of 360. Surrounding View Stitching Technology," 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 2020, pp. 661-664, doi: 10.1109/IFEEA51475.2020.00141.

[5] Xu, Mai, Chen Li, Shanyi Zhang, and Patrick Le Callet. "State-of-the-art in 360 video/image processing: Perception, assessment and compression." IEEE Journal of Selected Topics in Signal Processing 14, no. 1 (2020): 5-26.

[6] Maurya, Himanshu, Kamal Rohilla, Manoi Kumar, and Kumar Sundaram. "Image Description Generator: An Overview." In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1796-1799. IEEE, 2022.

[7] Zhang, Kuan, Yan Guo, Muhan Guo, and Younghwan Pan. "Using User Behavior Models and Visual Immersion to Optimize Digital Space Content." In 2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 59-63. IEEE, 2023.

[8] Meskó, Bertalan, and Eric J. Topol. "The imperative for regulatory oversight of large language models (or generative AI) in healthcare." npj Digital Medicine 6, no. 1 (2023): 120.

[9] Zhang, Chenshuang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. "Text-to-image diffusion model in generative ai: A survey." arXiv preprint arXiv:2303.07909 (2023).

[10] Zhang, Lvmin, 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, pp. 3836-3847. 2023.