RAMAIAH
Institute of Technology

Department of Artificial Intelligence and Data Science
VI Semester, B.E
Deep Learning (AD62)

Text to video generation using GANs for educational purposes

BY

K Nikisha (1MS21AD023)

S Kushal (1MS21AD043)

Under guidance of

Jamuna S Murthy, M.Tech, (PhD)

Assistant Professor

Agenda

- Introduction
- Literature Review
- Research Gap Analysis
- Social Impact
- Implementation (if any)
- Results and Performance Analysis (if any)
- Conclusion
- Future Work
- References

Introduction

Learning has always been a vital part of every individual's life. Sometimes it may so happen that we fall short of resources to learn and explore about certain topics or concepts . Stable Diffusion is a deep learning architecture that focuses on gradually transforming random noise into coherent data, such as images or videos, through a series of diffusion steps producing highly realistic data. We will be using this neural network to produce different educational content on animal kingdom to make the learning easier and engaging .

Literature Review

DCGANS:

A Deep Convolutional Generative Adversarial Network (DCGAN) is an advanced type of Generative Adversarial Network (GAN) that incorporates deep convolutional layers, making it particularly effective for generating high-quality images.

Video GAN:

Compared to image GANs, video GANs require different treatment because of the data complexity. A video consists of multiple images with the additional time dimension.

Text2Video zero shot:

Text2Video-Zero is a new method that allows creating videos from text descriptions without the need for training on large video datasets. This method uses existing models designed for generating images from text and adapts them for video creation.

Stable Diffusion model:

Stable diffusion model is a image generative from the text descriptions. It uses latent diffusion technique in which the model starts with a random noisy image and iteratively improves it to create a clear, high-quality image that matches the text.

Research Gap Analysis

Limitations of current approaches :

1. They need a lot of training on large video datasets, which is costly and time-consuming.
2. Keeping the video consistent across frames is difficult, often resulting in objects changing unexpectedly.
3. The generation process can be complex and resource-intensive.
4. These models often rely too much on specific reference videos, limiting their ability to create new content.
5. Adapting image generation models to create videos isn't easy, leading to issues with maintaining coherence.
6. Their application scope is limited due to these challenges.

Social Impact

It makes learning more engaging and accessible by turning text-based information into dynamic visual content. This can help students better understand complex concepts, especially in subjects like science and history. Additionally, it can support personalized learning experiences, cater to different learning styles, and provide educational content in multiple languages.

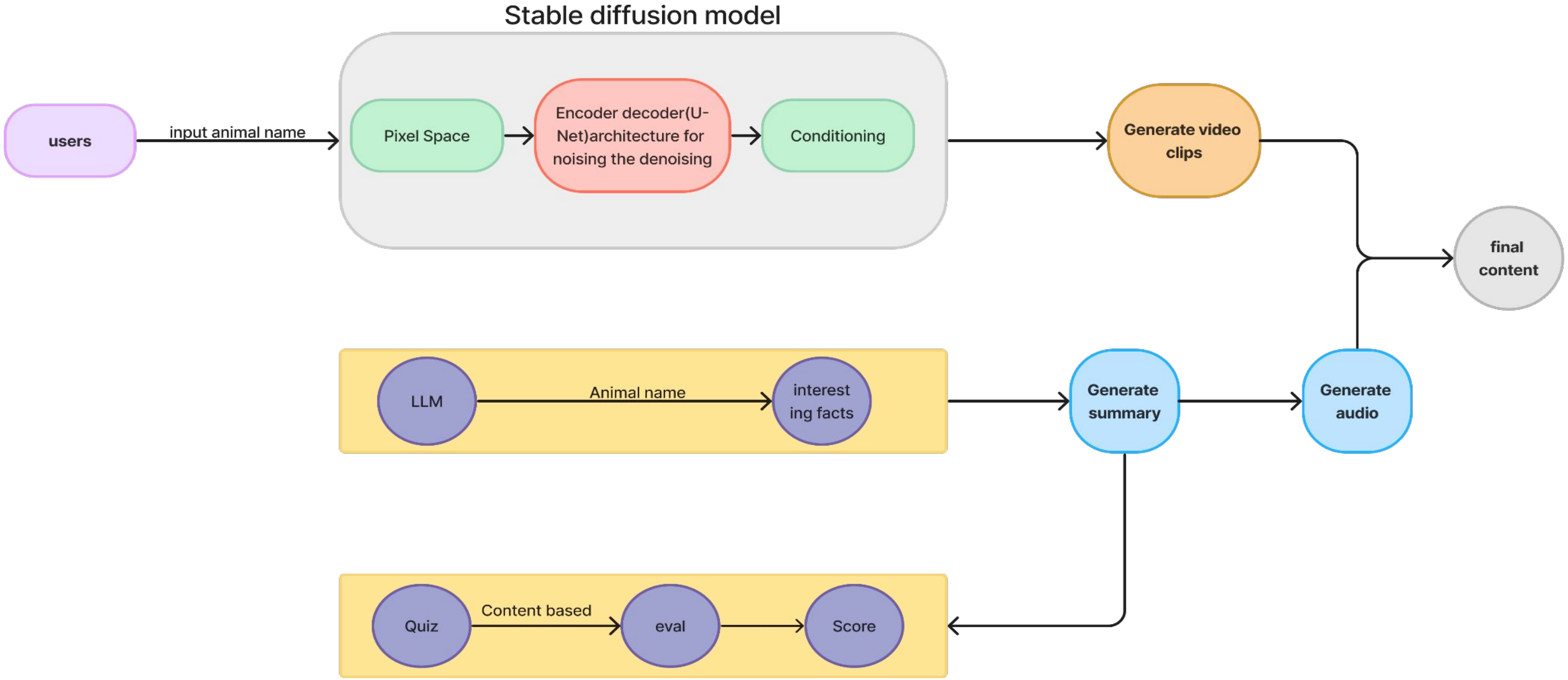
Implementation (if any)

We are using the LLM for generation of text concept for providing an description about the animal input given by the user . the content generated is later being converted into an audio file using google text-to-speech library . this will give us the output 1 .

For video generation text-to-video-zero-shot is given some standard prompts along with the input animal to generate a clip of that specific activity. This results in output 2

Later the above two outputs are combined together to generate the final video with audio and generated content.

Architecture



Results and Performance Analysis

Conclusion

The project employs advanced language and video generation techniques to produce concise animal descriptions and corresponding visual representations. By utilizing a Language Model (LLM) and Text-to-Video-Zero-Shot technology, we create informative videos combining narration and dynamic visuals. This approach enhances accessibility and engagement, showcasing the potential of AI-driven multimedia content creation.

Future Work

The future will include the following features :

1. Expanding to domain
2. Generating longer and optimised videos
3. Reduction of noise in the video
4. Integration with quiz app

References

- Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15954-15964).
- https://freedium.cfd/https://towardsdatascience.com/state-of-video-generation-76595bf75f46
- Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Video generative adversarial networks: a review. ACM Computing Surveys (CSUR), 55(2), 1-25.
- https://colab.research.google.com/github/tensorflow/docs/blob/master/site/en/tutorials/generative/dcgan.ipynb
- https://www.simplilearn.com/generative-adversarial-networks-applications-article
  https://freedium.cfd/https://medium.com/age-of-awareness/how-to-create-text-to-video-using-ai-a85f84d21252