RSET

RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)

Project Report On

Story Quest

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Niya George N (U2003219)

Riya Thomas (U2003169)

Navya Rony A (U2003146)

Shikha Mariam Joseph (U2003196)

Under the guidance of

Dr. Uma Narayanan

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
April 2024

CERTIFICATE

This is to certify that the project report entitled ***"Story Quest"*** *is a bonafide record of the work done by* ***Niya George N (U2003219)*** *,* ***Riya Thomas (U2003169)*** *,* ***Navya Rony A (U2003146)*** *,* ***Shikha Mariam Joseph (U2003196)*** *, submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Dr. Uma Narayanan
Project Guide
Asst. Professor
Dept. of CSE
RSET

Ms. Anita John
Project Coordinator
Asst. Professor
Dept. of CSE
RSET

Dr. Preetha K.G
Professor & HOD
Dept. of CSE
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Dr. P.S Sreejith**, Professor, Principal of RSET, and **Dr. Preetha K.G.** , Professor, Head of the Department of Computer Science, for providing us with the opportunity to undertake our project, "Story Quest".

We are highly indebted to our project coordinators, **Ms. Anita John**, Assistant Professor, Department of Computer Science, and **Mr. Sajanraj T.D.** , Assistant Professor, Department of Computer Science, for their valuable support.

It is indeed our pleasure and a moment of satisfaction for us to express our sincere gratitude to our project guide **Dr. Uma Narayanan** for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Niya George N

Riya Thomas

Navya Rony A

Shikha Mariam Joseph

Abstract

In the digital age, where storytelling meets technological innovation, we present "Story Quest," a novel project that seamlessly combines Natural Language Processing (NLP) for text story processing and Stable Diffusion techniques for the creation of captivating picture books from textual narratives. Leveraging state-of-the-art NLP models, our system adeptly summarizes and strategically splits textual stories to distill the essence and structure for further creative exploration. The integration of Stable Diffusion, a powerful generative method, transforms these processed narratives into visually stunning and contextually rich illustrations, ultimately crafting immersive picture book experiences.

This project aims to redefine storytelling in the digital realm, providing a dynamic platform for the synthesis of language and art, fostering creativity and engagement for users of diverse backgrounds. The use of NLP and Stable Diffusion in Story Quest holds the promise of revolutionizing the way we perceive and interact with narrative content.

Contents

List of Abbreviations

NLP: Natural Language Processing

CLIP: Contrastive Language-Image Pre-training

BERT: Bidirectional Encoder Representations from Transformers

MSCOCO: Microsoft Common Objects in Context

AI: Artificial Intelligence

DCGAN: Deep Convolutional Generative Adversarial Network

GLIDE: Guided Language to Image Diffusion for Editing

DDIM: Doubly Deterministic Implicit Models

GPT: Generative Pre-trained Transformer

AttnGAN: Attention Generative Adversarial Network

StackGAN: Stacked Generative Adversarial Network

NLTK: Natural Language Toolkit

T5: Text-To-Text Transfer Transformer

LED: Longformer Encoder Decoder

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In an increasingly digital world, storytelling has evolved beyond traditional text-based narratives. Today, the fusion of text-to-image generation and text summarization stands at the forefront of innovation, revolutionizing how stories are conveyed and consumed.

The Current Scenarios and its Importance are:

Communication Challenges: Traditional text-based storytelling faces limitations in engaging modern audiences. Attention spans have shortened, demanding more captivating and concise storytelling methods. Integrating text-to-image and summarization enables compelling and immersive storytelling. It caters to varied learning styles, capturing attention and fostering deeper engagement.

Visual Emphasis: The contemporary landscape favors visual content. There's a surge in demand for immersive and interactive storytelling experiences that blend visuals and concise textual information. Beyond entertainment, this innovation finds relevance in education, marketing, and various industries. It empowers educators, marketers, and content creators to deliver impactful messages through captivating narratives.

Technological Advancements: Text-to-image and summarization technologies have made significant strides, offering opportunities to bridge the gap between textual narratives and visual representation. In a world where multimedia content reigns supreme, the project addresses the preferences of modern audiences, catering to their visual-centric consumption habits.

Hence the significance include:

Digital Transformation: The project aligns with the ongoing shift toward multimedia consumption. It capitalizes on the opportunities presented by emerging digital mediums, offering a novel and engaging way to tell stories.

Ethical and Inclusivity Focus: By acknowledging the ethical considerations in generating visual content and working to mitigate biases, the project emphasizes inclusivity and fair representation in storytelling.

Therefore, the story visualization project represents a pivotal step in storytelling evolution. By combining text-to-image generation and summarization, it aims to create immersive, engaging, and ethical storytelling experiences, addressing the demands and preferences of modern audiences across various industries.

1.2 Motivation

The "StoryQuest" a machine learning picture book generating project aims to address traditional picture book development difficulties, such as time-consuming and expensive illustration procedures. The project intends to break down barriers to creativity and diversity by using machine learning to automate the creation of visually appealing graphics that accompany written narratives seamlessly. It aims to simplify production workflows, giving authors an efficient and cost-effective alternative for creating personalised, culturally varied, and interactive picture books.

1.3 Objectives

- Create Text-to-Image Technology:

  Build a system that accurately converts written descriptions into corresponding visuals. Achieve high accuracy in generating images that match the provided text. Evaluate performance using different methods.

- Refine Text Summarization Techniques:

  Improve algorithms to condense long text into concise and meaningful summaries. Measure accuracy by comparing generated summaries with human-made ones. Ensure the algorithm efficiently captures vital information while staying coherent.

- Integrate Text and Visuals for Storytelling:

  Combine text-to-image generation and text summarization for storytelling. Develop a system merging visual content and summarized text for coherent storytelling. Check how this fusion improves understanding and engagement.

- Create User-Friendly Interface:

  Design an easy-to-use interface for interacting with text-to-image and summarization features. Enable users to input text, view images, and access summarized stories. Get feedback to enhance usability.

- Evaluate Real-World Use:

  Test the developed system's practical application in different industries. Assess performance through case studies or user surveys. Determine its usefulness and potential improvements based on user needs.

- Optimize Performance and Scalability:

  Improve efficiency and scalability of text-to-image and text summarization models. Optimize algorithms for faster processing while maintaining quality. Evaluate handling larger data volumes.

- Address Ethical Concerns and Bias:

  Identify and mitigate biases in generated images or summaries for fairness. Implement techniques to detect and fix biases. Conduct ethical evaluations to minimize unintended biases.

1.4 Purpose and Need

- **Enhanced Storytelling Experience:** The primary objective of the "Story Quest" project is to enhance storytelling through the deft transformation of written narra-

tives into visually appealing picture books. Giving customers a more engaging and immersive experience while reading stories is the aim of this modification.

- **Interactive and Dynamic Content Creation:** Story Quest allows users to have an active part in the creative process on a platform. The interface allows users to enter their stories, and it will automatically produce the appropriate graphic content. Users' creativity and sense of ownership are encouraged by this involvement.

- **Education Tool:** The initiative offers a novel approach to presenting and understanding narratives, making it a useful teaching tool. Learning can be made more fun and approachable by using visual aids to help with literacy development, especially for younger audiences.

- **Bridge Between Text and Imagery:** Using solely text, conventional storytelling frequently limits the visual experience. Story Quest enhances the storytelling experience overall by offering a seamless transition between textual narratives and vibrant images.

- **Innovative Technological Integration:** Creative projects that make use of these skills are needed as sophisticated natural language processing and image production technology become more prevalent. This need is satisfied by Story Quest, which combines NLP and stable diffusion for a novel and imaginative application.

- **Addressing Digital Literacy:** With more people becoming digitally literate, Story Quest fills the demand for original and insightful digital content. The project, which targets a tech-savvy audience, is in line with the changing ways that people engage with and absorb storytelling.

1.5 Scope

The project's scope involves seamlessly combining text-to-image creation and text summarization methods to develop an advanced system. This includes implementing sophisticated algorithms for accurate text-to-image translation and concise yet comprehensive summarization. Designing an intuitive interface for easy text input and content exploration is crucial. Evaluation and refinement will focus on improving accuracy, efficiency,

and scalability, exploring the project's relevance in education, marketing, and diverse industries. Ethical considerations, such as reducing bias in generated content, are integral. Gathering user input will guide improvements in usability to meet various user needs. Comprehensive documentation of the process and outcomes supports future research, while plans for adaptability ensure readiness for technological advancements and changing preferences. Collaboration and knowledge-sharing efforts aim to benefit the wider community through publications and collaborative initiatives.

1.6 Challenges

- **Data Quality and Diversity:** It is difficult to obtain a high-quality and varied collection of text tales because different genres, writing styles, and themes must be sufficiently represented for the model to be resilient.
- **Fine-tuning NLP Models**: Understanding the subtleties of storytelling is essential to fine-tuning NLP models for tale summary. It is a non-trivial challenge to extract important content while maintaining simplicity.
- **Computational Resources:** Stable Diffusion models and NLP may both require a large amount of processing power. It can be difficult to manage these resources effectively, particularly when processing and generating on a large scale and taking possible financial limits into account.
- **Semantic Coherence:** It is difficult to maintain semantic coherence when the system is processing text since it must comprehend and retain the main ideas of the story while reducing the amount of data.

1.7 Assumptions

- Quality of Textual Input:

  The quality of the textual input significantly impacts the accuracy and richness of the generated visualizations. A well-structured, coherent narrative text ensures that the summarization process accurately captures the story's essence. Consistent language and storytelling style across the text aids in maintaining coherence in the

summarization and subsequent visualization. Moreover, detailed and descriptive text enhances the ability of the stable diffusion model to create vivid and immersive images, providing a more nuanced representation of the story's elements.

- Availability of Diverse Training Data:

  Access to a diverse range of stories spanning various genres, themes, and cultural backgrounds enriches the model's understanding of storytelling nuances. Diversity in the training data allows the model to create representative visualizations, ensuring inclusivity and sensitivity to different perspectives. A balanced representation of story structures, lengths, and complexities in the training data aids the model in generalizing well and accurately visualizing diverse narratives.

- Stable Diffusion Model Suitability:

  The suitability of the stable diffusion model lies in its capability to accurately translate textual descriptions into coherent and meaningful images. Evaluating its performance on similar text-to-image tasks and its adaptability to varying lengths and complexities of input text ensures its effectiveness in generating diverse visualizations. Scalability and flexibility are crucial aspects to consider for accommodating different storytelling styles and structures.

- Adequate Computational Resources:

  Sufficient computational resources, including processing power and memory, are essential for training the models involved in text summarization, NLTK processing, and image generation using the stable diffusion model. Adequate resources ensure timely processing and generation of visualizations, particularly when dealing with large volumes of text or complex storytelling structures, meeting efficiency and time constraints.

- Ethical Considerations:

  Ethical considerations encompass various aspects, including the representation and avoidance of biases or stereotypes in the generated visualizations. Regular evaluations and bias detection mechanisms help in ensuring fair and unbiased representations. Respecting intellectual property rights by obtaining permissions and

providing appropriate attribution for the textual content used in the project is essential. Moreover, prioritizing user privacy and employing ethical data handling practices maintains trust and integrity in the project's execution.

1.8 Societal / Industrial Relevance

'StoryQuest' has numerous applications in a variety of industries. Through interactive textbooks, it customises materials, enhances language acquisition, and promotes literacy in education. It streamlines production in the publishing industry, lowering costs and expediting the creation of visually appealing and diversified picture books. Furthermore, it is important in entertainment, technological innovation, inclusive literature for visually impaired people, cross-cultural exchange programmes, digital archives, employment development in technology-driven creative industries, and the demand for personalised content platforms. Its potential ranges from societal impact initiatives to educational institutions that contribute to cultural appreciation and understanding.

1.9 Organization of the Report

- Introduction: Sets the stage for the study, providing background, motivation, objectives, purpose, scope, challenges, assumptions, societal/industrial relevance, and an overview of the report's organization.
- Literature Survey: Reviews existing research on webtoon generation, summarizing key studies and identifying gaps in the literature.
- Requirements: Details the hardware and software requirements necessary for the project, including both technical specifications and functional necessities.
- System Architecture: Describes the overall structure of the system, including an overview, architectural design, module division, work breakdown, responsibilities, and a Gantt chart illustrating the project's timeline.
- System Implementation: Discusses the proposed methodology, user interface design, database design, and implementation strategies for realizing the system.

- Results and Discussions: Presents the results of testing, provides an overview of the findings, and engages in discussions about the implications and potential limitations of the results.
- Conclusions & Future Scope: Summarizes the key conclusions drawn from the study and outlines potential future directions for research and development in the field.

Chapter 2

Literature Survey

2.1 A Study on Webtoon Generation Using CLIP and Diffusion Models [1]

The method of generating webtoons using CLIP and the diffusion model involves a multi-step process that leverages the capabilities of both models:

Training the CLIP Model: The process begins with training the CLIP model on the treatment webtoon dataset. The CLIP model leverages pretrained multilingual BERT and ViT encoders for text and images, respectively, with an added projection layer of 512 dimensions in the final output of each encoder. During training, all layers of the encoders are optimized without freezing any layers. The contrastive loss function is employed by placing the text and image vectors in the same-dimensional projection layer, measuring the cosine similarity between the vectors, and computing the loss through cross-entropy.

Identifying Similar Images: To generate webtoons, the first step involves identifying the image most similar to the desired treatment within the dataset. This is achieved by computing the cosine similarity between the CLIP embedding vectors of all dataset images and the embedding vector of the given text. The index of the embedding vector with the highest similarity score corresponds to the most similar image.

Depth-to-Image Model of Stable Diffusion: Once the most similar image is identified, it is input, along with text, into the depth-to-image model of stable diffusion to generate a webtoon. The depth-to-image model uses the depth information of the input image as an initial image and progressively removes noise through an inverse diffusion process to create an RGB image. This approach is viable because the model is trained using paired RGB and depth images. During training, the model predicted the color in-

formation associated with the depth information and used it to determine the RGB values of the individual pixels.

Generating Cartoon-Style Images: The depth-to-image model excels in producing realistic images, and by inputting the model with the keyword "webtoon" as a query, the desired cartoon-style images are generated. This process allows for the creation of webtoons based on the input text and the identified similar image, resulting in the generation of visually appealing and contextually relevant webtoon content.

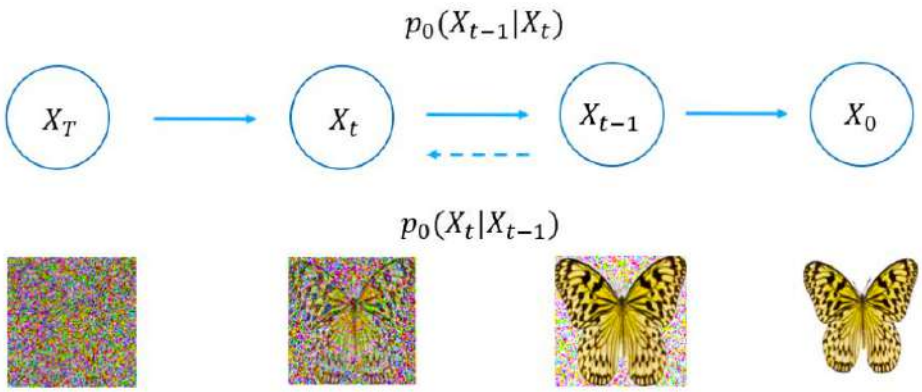


Figure 2.1: Training process of Diffusion Model [1]

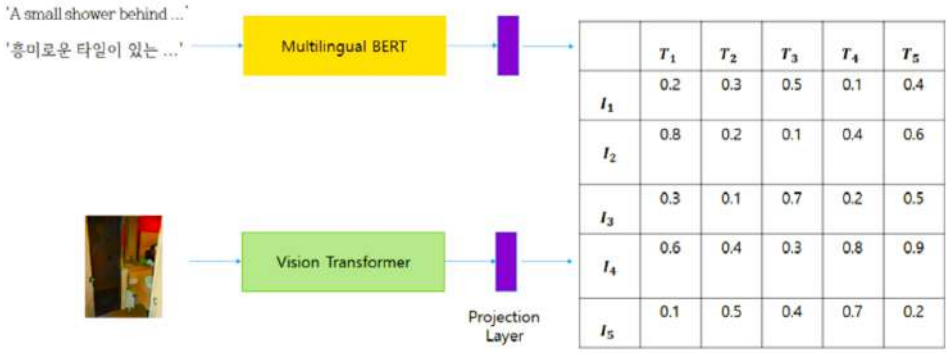


Figure 2.2: Structure of CLIP [1]

2.2 Generating Webtoons Using Multilingual Text-to-Image Models [2]

The study's methodology include creating a multilingual text-to-image model that can process text inputs in both Korean and English to produce webtoon visuals. The procedure is divided into multiple phases, such as creating the dataset construction, feature extraction, model training and evaluation.

Dataset Construction: The authors construct a webtoon dataset consisting of English and Korean text and image data. They use the MSCOCO dataset released by Microsoft and the Korean MSCOCO dataset translated and released by the AI Hub. These datasets are transformed into cartoons using CartoonGAN, resulting in a multilingual treatment-webtoon dataset.

Feature Extraction: The study utilizes a pre-trained multilingual BERT model to extract feature vectors from the multilingual text inputs. The feature vectors are obtained using the "cls" token in the BERT model as the sentence vector. This step is crucial as the quality of the generated images is heavily influenced by the quality of the learned features.

Model Training: The authors train a GAN-based text-to-image model using the extracted text features and image data. They use the DCGAN (Deep Convolutional Generative Adversarial Network) model for this purpose. The training process involves combining the extracted text features with noise and training the model to generate webtoon images based on the input text.

Evaluation: The performance of the proposed multilingual text-to-image model is evaluated using the inception score and Frechet inception distance (FID) scores. These scores are used to assess the quality and diversity of the generated images and compare the mean and covariance values of the feature values of the real and generated images.

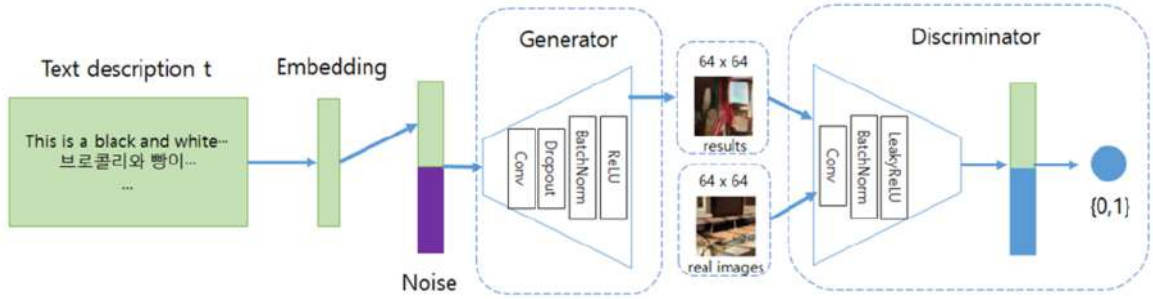


Figure 2.3: BERT Architecture [2]

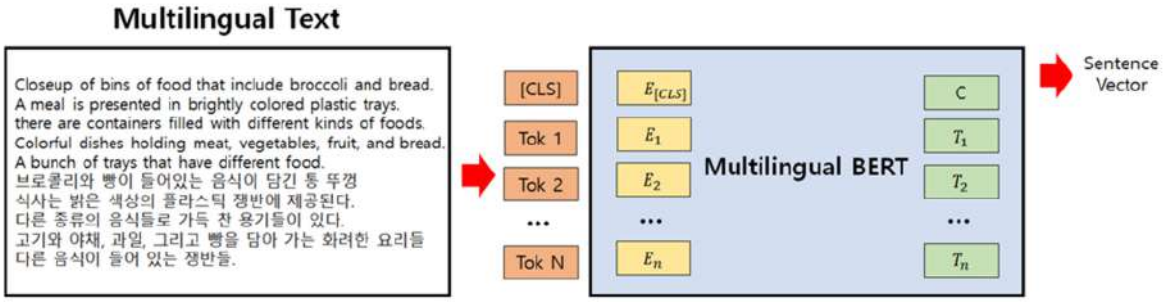


Figure 2.4: DCGAN Architecture [2]

2.3 Hierarchical Text-Conditional Image Generation with CLIP Latents [3]

The study proposes a two-stage model for image generation using CLIP representations: a prior that generates a CLIP image embedding given a text caption,capturing both semantics and style and a decoder that generates an image conditioned on the image embedding producing variations of an image.

The model explicitly generates image representations, which improves image diversity without significant loss in photorealism and caption similarity.

The joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion.

The study experiments with autoregressive and diffusion models for the prior.The diffusion prior directly models the continuous vector of the CLIP image embedding using a Gaussian diffusion model conditioned on the text caption. Diffusion models are found to be more computationally efficient and produce higher-quality samples.

The authors also mentions the use of diffusion upsampler models to generate high-resolution

images. The upsamplers are trained to upsample images from 64x64 to 256x256 and further to 1024x1024 resolution.

The CLIP embedding used in the model does not explicitly bind attributes to objects, leading to difficulties in accurately representing and reconstructing objects and attributes. The models were trained using Adam with corrected weight decay and momentum.Adam combines the benefits of two other optimization algorithms, namely AdaGrad and RMSProp, to provide efficient and effective optimization.Percentile 50 was found to be optimal in all experiments.The embeddings maintains a learning rate for each parameter and adapts the learning rate based on the first and second moments of the gradients.

Dummy captions or no captions were also experimented with and yielded good results. The study includes random samples from the production model for some prompts, showing less than 1 average mean-squared error in reconstructing the image representations

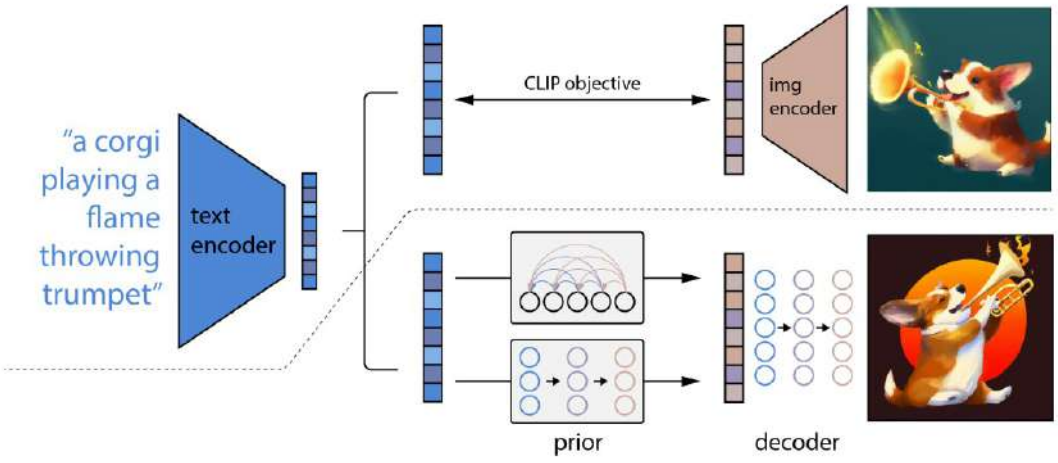


Figure 2.5: A high-level overview of unCLIP [3]

In the figure, above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image.

2.4 GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models [4]

GLIDE (Guided Language to picture Diffusion for Generation and Editing) is a text-guided diffusion model designed for photorealistic picture generation and editing. The model is based on diffusion models, which have been proved to produce high-quality synthetic images when combined with a balancing strategy for diversity and fidelity. The model's primary components include diffusion models, classifier-free guidance, and CLIP guidance.

The data distribution is used to generate a Markov chain of samples using diffusion models based on Gaussian diffusion. In a series of stages, the model gradually noises high-resolution images, and an approach for learning the noise scale allows the model to create high-quality samples with fewer diffusion steps. Diffusion models have also been used to super-resolve images, exhibiting its adaptability in image synthesis applications.

Text prompts are used to assist the model, and two distinct guidance procedures are compared: CLIP guidance and classifier-free guidance. Human assessors favour the classifier-free guide technique for both photorealism and caption similarity. The model can render a wide range of text prompts in zero-shot mode, as well as photorealistic graphics with shadows, reflections, and high-quality textures. Furthermore, the model can generate creative renderings of unique thoughts and realistic changes to existing photos using natural language inputs.

The proposed method also includes a smaller filtered model, referred to as GLIDE (filtered), which is fine-tuned to perform image inpainting. The model is able to match the style and lighting of the surrounding context to produce realistic image completions. The resulting model significantly reduces the effort required to produce convincing disinformation or Deepfakes, and to safeguard against these use cases, a smaller diffusion model and a noised CLIP model trained on filtered datasets are released.

To compare the performance of the models, human evaluations are performed, and sam-

pling hyperparameters are optimised to achieve good sample quality. To measure the capabilities acquired from training a large model on a diverse dataset, the model is compared to smaller models. In terms of attaching qualities to objects, composing tasks, and merging unexpected concepts, the larger model surpasses the smaller models.

The proposed method also includes a people filter, which employs CLIP to detect the presence of people in photos, and a classifier to eliminate images of violent objects and hate symbols from the dataset. These elements contribute to ethical considerations and appropriate model use.

Overall, the GLIDE model is capable of producing photorealistic images, text-driven image editing, and addressing ethical concerns in image synthesis. Human evaluations and comparisons with smaller models are used to assess the model's performance, demonstrating its effectiveness in creating high-quality images driven by natural language cues.

2.5 DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation [5]

The paper introduces a novel technique named DiffusionCLIP, which utilizes diffusion models and CLIP loss for manipulating images based on textual prompts. Diffusion models, a class of generative models, generate high-quality images by progressively eliminating noise from a random input. CLIP loss measures the similarity between an image and text within a shared feature space. The paper illustrates that DiffusionCLIP effectively performs faithful and robust image manipulation for both familiar and unfamiliar images. This means it can maintain crucial details and the identity of the input image while altering desired attributes in accordance with the provided text prompt. Moreover, the paper showcases various innovative applications of DiffusionCLIP, including image translation across previously unseen domains, synthesis of images conditioned on strokes, and multi-attribute transfer.

The paper introduces a novel approach termed DiffusionCLIP, designed for text-guided image manipulation. This method combines two fundamental components: diffusion mod-

els and CLIP loss. Diffusion models, known for their generative capabilities, involve a forward process that introduces noise to data and a reverse process that denoises the data using a learned score function. This unique combination enables the gradual removal of noise from a random input, contributing to the synthesis of high-quality images.

CLIP loss, a pivotal element in the DiffusionCLIP method, operates as a metric to measure the similarity between an image and a textual description in a shared embedding space. Leveraging a pre-trained CLIP model, this loss function plays a crucial role in aligning images with the provided text prompts during the manipulation process.

The paper goes further to refine the performance of the reverse diffusion process by introducing a directional CLIP loss. This fine-tuning process aims to align the direction between image embeddings with the direction between reference and target texts. By incorporating this directional CLIP loss, the DiffusionCLIP method enhances the accuracy and fidelity of the image manipulation process.

In addition, the paper adopts a deterministic forward and reverse process based on DDIM (doubly deterministic implicit models). This choice in methodology facilitates full inversion of latent variables and expedites the sampling process. The deterministic nature of the forward and reverse processes contributes to the overall efficiency of image manipulation using DiffusionCLIP.

The paper demonstrates the practical applications of DiffusionCLIP in various scenarios. It showcases the method's ability to perform faithful image manipulation for both familiar and unfamiliar images, preserving essential details and identity. Additionally, DiffusionCLIP proves its versatility by successfully translating images between domains that were not encountered during the training phase. The method also introduces a novel application—stroke-conditioned image synthesis, where images are generated based on stroke inputs. Furthermore, the paper illustrates the capability of DiffusionCLIP in multi-attribute transfer, showcasing its effectiveness in simultaneously manipulating multiple attributes in images. In summary, DiffusionCLIP emerges as a robust method for text-guided image manipulation, offering a range of applications and demonstrating its

potential for diverse and innovative use cases in the field.

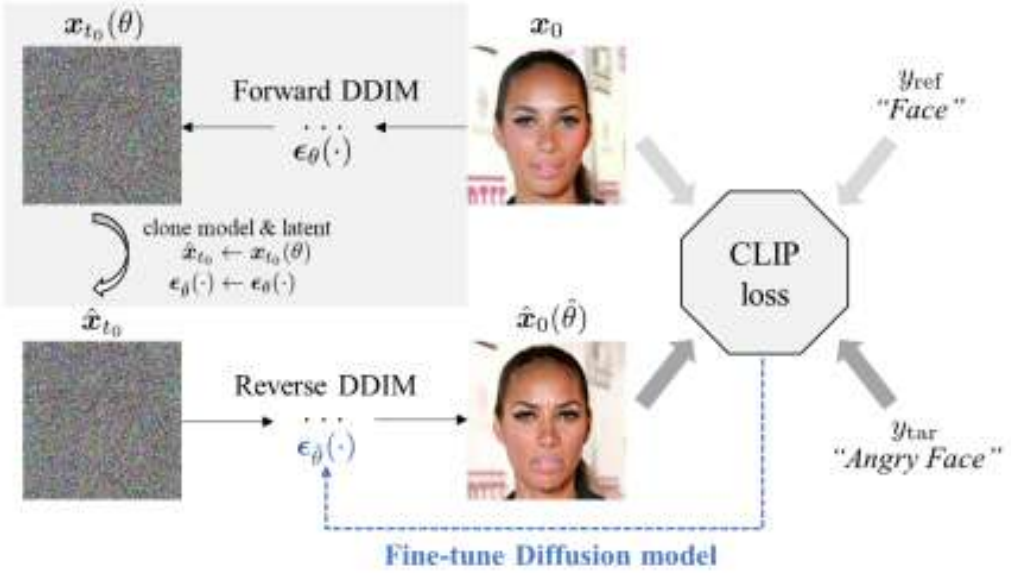


Figure 2.6: Overview of DiffusionCLIP [5]

The input image is first converted to the latent via diffusion models. Then, guided by directional CLIP loss, the diffusion model is fine-tuned, and the updated sample is generated during reverse diffusion.

2.6 Existing Methods

Existing methods for the conversion of a textual story to picture book is limited.
Most common method is the usage of text-to-image models that convert text prompts to image. The user has to then manually combine the images to create the desired picture book.

Commonly used text-to-image models :

Stable Diffusion: It is a generative model for image synthesis that employs controlled random transformations, ensuring stability throughout the process. It gradually enhances the complexity of an initial image, resulting in high-quality and diverse outputs.

DALL-E: Developed by OpenAI, DALL-E is a text-to-image model that generates images from textual descriptions. It uses a variant of the GPT architecture and is trained on a diverse dataset to understand and create images based on textual prompts.

AttnGAN: It is a model that incorporates attention mechanisms into the traditional Generative Adversarial Network (GAN) framework. It pays selective attention to different parts of the input text to generate more realistic and detailed images.

StackGAN: It is another GAN-based model that generates high-resolution images from textual descriptions. It works in a two-stage process, where the first stage generates a low-resolution image, and the second stage refines it to a higher resolution, resulting in more detailed and realistic images.

Midjourney: It employs a process that carefully evolves an initial image to a final state. This approach ensures the generated images maintain coherence and quality throughout the synthesis process, making Midjourney effective for producing visually appealing results.

2.7 Summary and Gaps Identified

Paper Title	**Advantages**	**Disadvantages**
A Study on Webtoon Generation Using CLIP and Diffusion Models [1]	• CLIP's ability to identify the most similar image to a given text. • The use of the depth-to-image model within Stable Diffusion excels in producing realistic images.	• Limited Control over Style and Content. • The multi-step process involving training the CLIP model and utilizing the depth-to-image model adds complexity to the overall method.
Generating Webtoons Using Multilingual Text-to-Image Models [2]	• Multilingual Capabilities for Diverse Audiences. • BERT model used for feature extraction, ensuring that the generated images are influenced by high-quality and cross-lingual learned features.	• Accuracy of DCGAN when trained with webtoon dataset was significantly low. • The conversion process using CartoonGAN may not fully capture the intricate details of webtoons.

Paper Title	**Advantages**	**Disadvantages**
Hierarchical Text-Conditional Image Generation with CLIP Latents [3]	• Improved Image Diversity without Loss in Realism. • Efficient Language-Guided Manipulations.	• The use of CLIP embeddings may result in a lack of fine-grained control over attribute representation in generated images • Computational Complexity of Diffusion Models.
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models [4]	• Photorealistic Image Generation with Guided Diffusion. • It uses CLIP guidance and classifier-free guidance to assist in the image generation process.	• Computational Complexity and Resource Requirements. • Filtering training data for safety can limit dataset diversity and representation.
DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Ma- nipulation [5]	• The diffusion model can achieve high-quality image synthesis. • Flexibility and broad utility of DiffusionCLIP in creative image synthesis tasks	• Fine-tuning the diffusion model for specific texts is time-consuming and computationally expensive. • Dependency on Pre-trained CLIP Model.

Table 2.1: Gaps Identified

Chapter 3

Requirements

3.1 Hardware and Software Requirements

3.1.1 Hardware Requirements

1. Processor-Intel i5:

   It should have an Intel i5 processor or an equivalent which is suitable for a variety of computing tasks, offering a balance between performance and cost.

2. RAM-8GB or more: The system should have a minimum of 8 gigabytes of RAM which is essential for storing and quickly accessing data that the processor needs during operation, contributing to overall system performance.

3. Hard Disk Space-256GB:

   The system should have at least 256 gigabytes of storage space which is crucial for holding the operating system, applications, and user data.

3.1.2 Software Requirements

1. Python 3.10

   Requires the use of Python 3.9 as the programming language.

2. Libraries such as NLTK (Natural Language Toolkit) and spaCy

   Incorporation of external libraries, specifically NLTK and spaCy, for natural language processing tasks.

3. PyTorch v2.1.0

   Mandates the integration of PyTorch, an open-source machine learning library, for the implementation of deep learning algorithms.

4. TensorFlow v2.14.0

   Requires the inclusion of TensorFlow, an open-source machine learning framework, that is essential for the implementation of various machine learning and deep learning models.

5. Stable Diffusion XL (Text-to-Image Model)

   Requires the implementation of a stable diffusion technique for generating images from text descriptions.

6. Visual Studio Code 17.5

   It is designated as the preferred integrated development environment (IDE) for the project, providing features such as debugging, syntax highlighting, and Git integration.

Chapter 4

System Architecture

4.1 System Overview

Story Quest aims to automate the creation of picture books, offering users the convenience of inputting their own stories or selecting a classic story that is available on the website. By leveraging advanced natural language processing (NLP) and image processing techniques, the system generates visually engaging picture books with minimal user intervention.The project is executed as a Flask web application, with HTML, CSS, and JavaScript utilized for the frontend. The system follows a multi-step process to generate picture books:

- Input Handling: Users input their own stories or select one of the classic tales availble in the website.
- Text Processing: If the word count exceeds 200, the story undergoes text summarization using a LED model trained on BookSum data.
- Co-reference Resolution: Resolution is done using Fastcoref helps in maintaining character consistency for the generate images.
- Image Generation: The processed text is passed to a stable diffusion model, which generates corresponding images for each sentence of the story. The text is overlaid onto the images using the Python Pillow library.
- Displaying Picture-book: The resulting images are displayed in a flip book format according to the narrative flow of the story on the web application, allowing users to navigate through the story.
- Video and Generation: Audio narration is generated for each image using the Google Text-to-Speech (gtts) library which is then combined with the generated images using MoviePy python library to generate video.

- Output Generation: Users have the option to download the picture book as a PDF and also to view video with audio narration.

4.1.1 LED Architecture

For processing lengthy documents and text production activities, a key breakthrough is the Longformer encoder-decoder (LED). LED's two-part architecture solves the problem of processing lengthy sequences effectively. Using a novel attention mechanism, the Longformer encoder utilizes a unique attention mechanism that prioritizes nearby words for most interactions but can also consider distant connections when necessary, striking a balance between local and global attention. In the meantime, the decoder creates precise and logical output sequences by using the encoder's detailed summary and knowledge of the words it generates.

4.1.2 Stable Diffusion Architecture

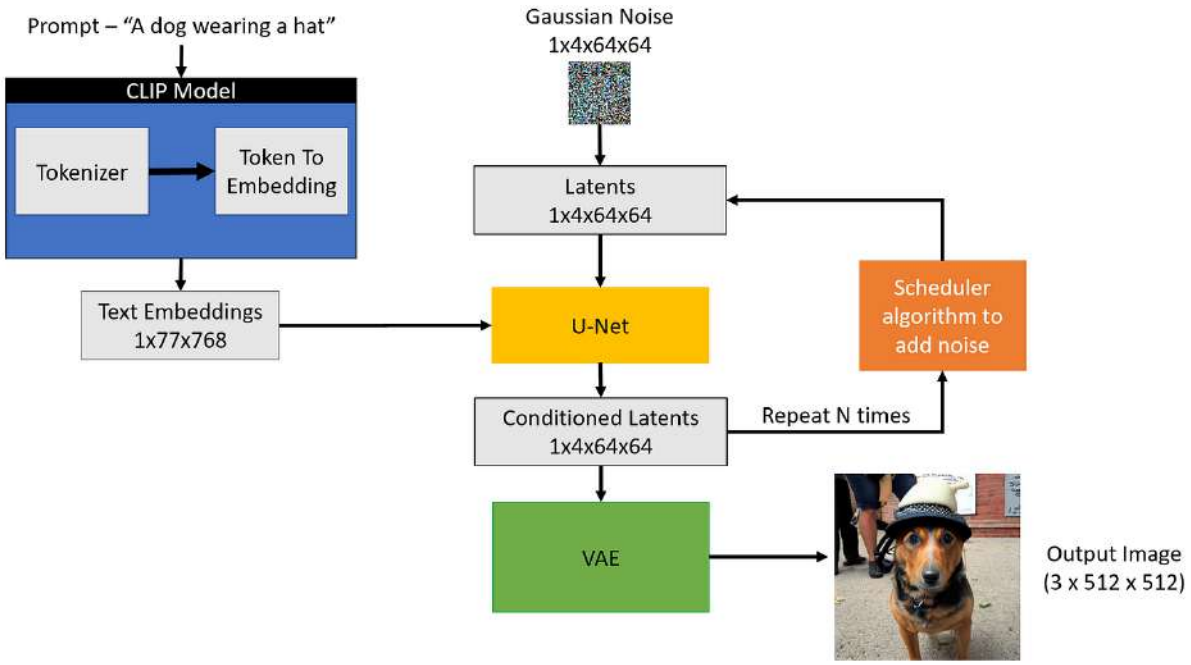


Figure 4.1: Stable Diffusion Architecture [11]

The stable diffusion model combines text processing with image processing to create high-quality images from text prompts. The text input is converted to text embeddings by CLIP model to capture text in the latent space. The seed passed is used to produce

Gaussian noise to form the initial image latent representation. This initial image latent representation and the text embeddings obtained from the CLIP model serve as an input for the U-net architecture. The U-net iteratively fine-tunes the noise latent representation by subtracting noise in several steps. At every step, the U-net predicts residual noise which is then subtracted from latent space to get closer to de-noised state. This process is repeated multiple times until the noise is minimal. Finally, the latent image representation is decoded by Variational Auto Encoder (VAE) to generate the final output image.

4.2 Architectural Design

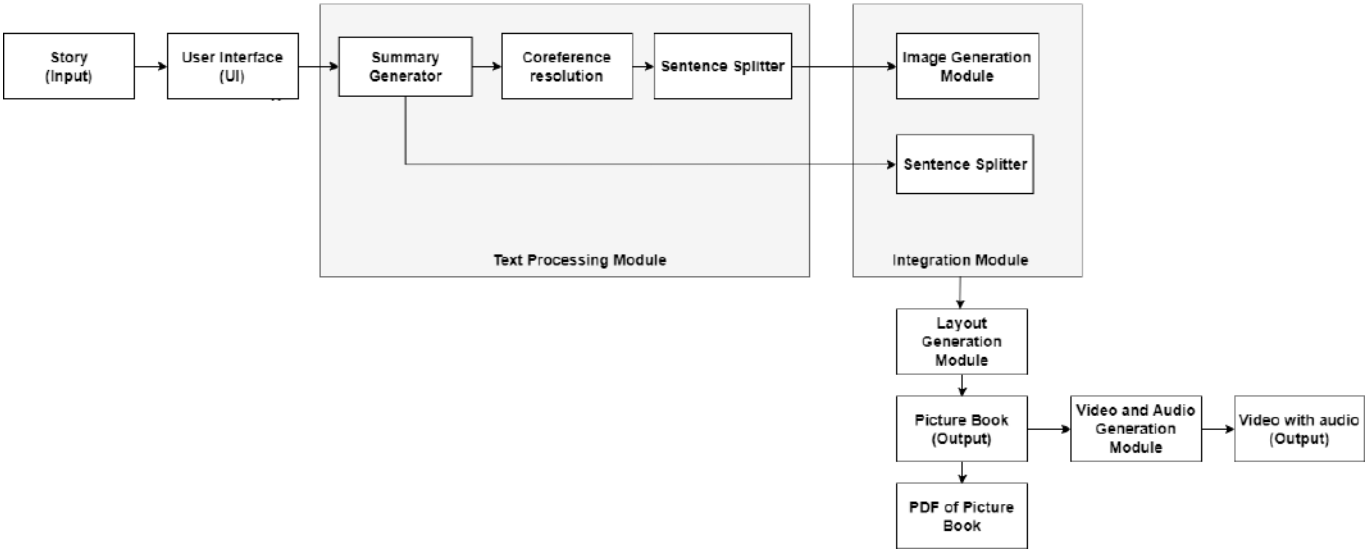


Figure 4.2: Architecture Diagram

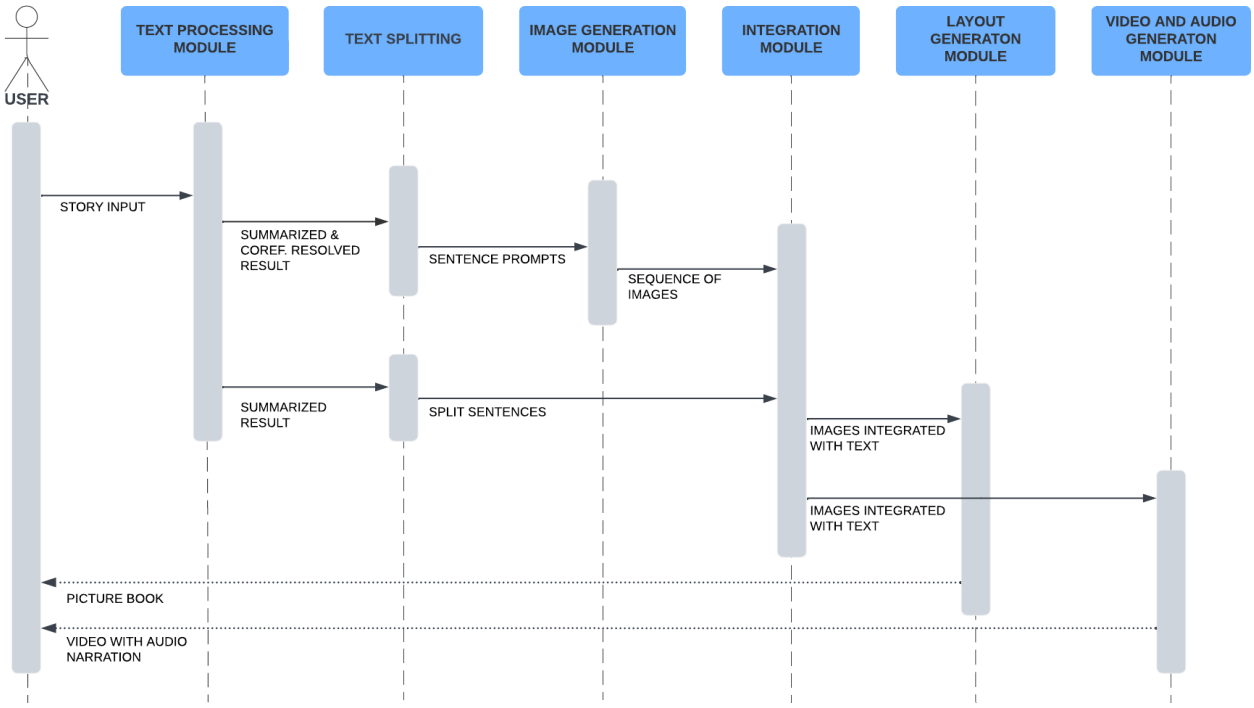


Figure 4.3: Sequence Diagram

4.3 Module Division

The system comprises of the following modules:

- User Interface (UI) Module
- Text Processing Module
- Image Generation Module
- Integration Module
- Audio and Video Generation Module
- Layout Generation Module

4.3.1 User Interface Module

The UI module is composed of an intuitive and user-friendly interface for users to interact with the system. Users can input a story or select one of the classic stories available on the website. The picture book is displayed along with its corresponding text and buttons are used to navigate through the story. The corresponding video is also displayed in the website.

4.3.2 Text Processing Module

The text processing module handles the input and manipulation of textual data, such as the story provided by the user. It employs an LED model for summarising the story if the word count is more than 200. It also uses fastcoref for co-reference resolution where the pronouns are replaced to main consistency of the characters when generating the image. The co-reference resolved text is then spilt into individual sentences to be passed as input prompts to the Image Generation Module

4.3.3 Image Generation Module

This module is responsible for generating images based on the processed textual data. Using Stable Diffusion, which is a generative model for creating high-quality images, it creates visually appealing illustrations that correspond to the different events in the story.

4.3.4 Integration Module

The integration module facilitates seamless communication between the text processing and image generation modules. It ensures that the extracted information from the story is effectively utilized to produce coherent and contextually relevant images. This maintains the narrative flow and visual consistency of the story.

4.3.5 Layout Generation Module

The layout generation module emphasis on arranging the generated images in a cohesive and visually pleasing manner to form a picture book. This include determining the placement of images, considering factors like pacing, page transitions, and overall aesthetic coherence and also placing .

4.3.6 Video and Audio Generation Module

The audio and video generation module utilizes the GTTS to convert the story text into natural-sounding audio narration, synchronized with the flip book images. MoviePy python library compiles these images into a cohesive video format, ensuring smooth transitions and synchronization with the narration. This integration of audio and video elements enhances the storytelling experience, providing users with a captivating multi-sensory journey through the narrative.

4.4 Work Breakdown and Responsibilities

- Niya George - Text Processing Module, Layout Design, Documentation
- Riya Thomas - Text Processing Module, Layout Design, Documentation
- Navya Rony A - Image Generation module, UI Design, Documentation
- Shikha Mariam Joseph - Image Generation module,UI Design, Documentation

4.5 Work Schedule - Gantt Chart

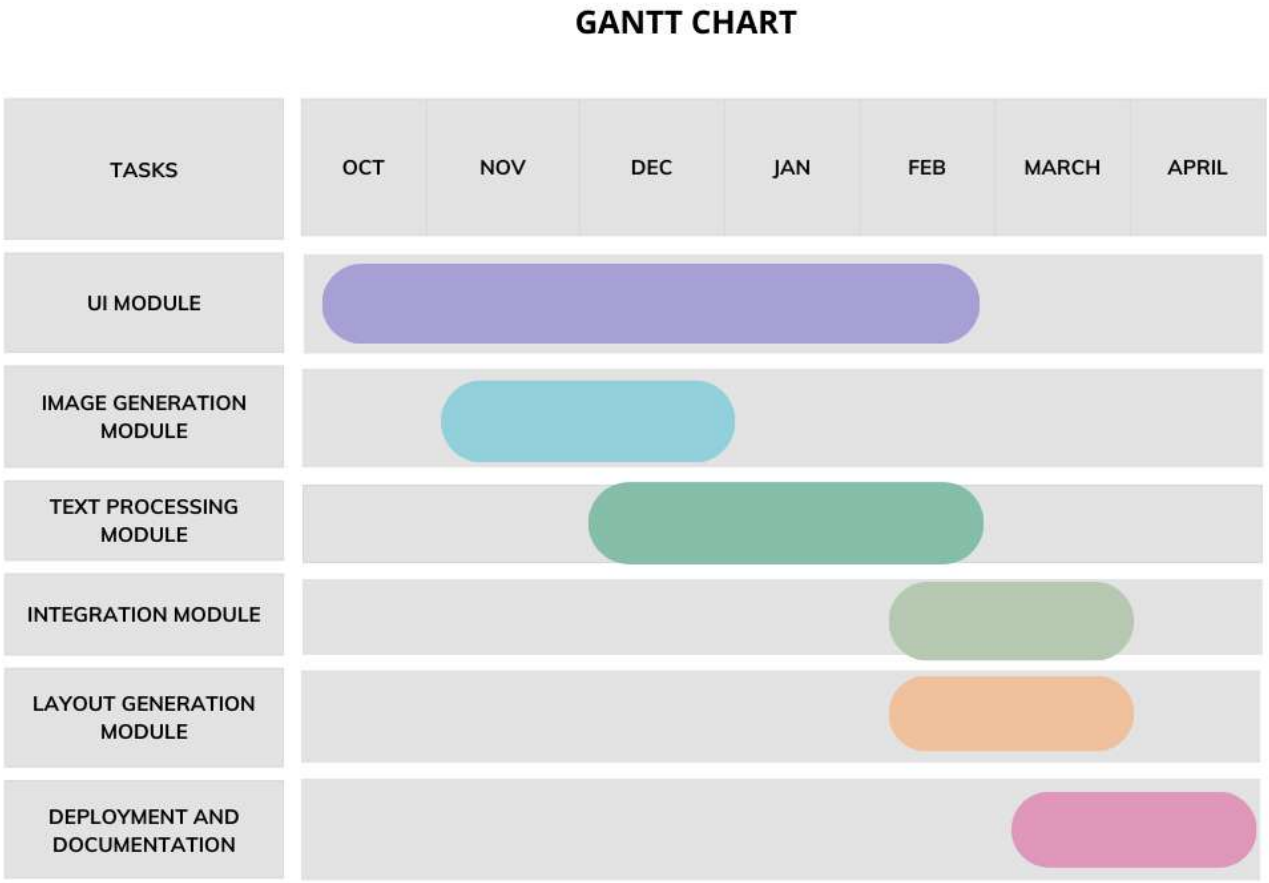


Figure 4.4: Gantt Chart

Chapter 5

System Implementation

5.1 Proposed Methodology

Story Submission: Users submit written stories to the platform.

Length Check: Check if the submitted story is longer than 200 words. If it is, pass it to the LED summarizer. If not, the story is passed for coreference resolution.

LED Summarization: Pass the story through an LED (Longformer Encoder-Decoder) summarization model. This model can effectively condense longer texts while retaining the essential information. The summarized version of the story is passed for coreference resolution.

Coreference Resolution: Perform coreference resolution on the summarized story using fastcoref. This step ensures that pronouns and other referring expressions in the text are correctly linked to their corresponding entities, enhancing the overall coherence and readability of the story.

Sentence Splitting: Split the story into individual sentences. This step is crucial for feeding the sentences into the text-to-image model separately.

Text-to-Image Generation: Utilize the Stable Diffusion XL Text-to-Image model to generate images corresponding to each sentence of the story. This model can produce high-quality images based on textual descriptions.

Image Compilation: Combine all the generated images into a cohesive sequence, preserving the narrative flow of the story.

Video Generation: Provide an option for users to watch the compiled images as a video with accompanying audio. Use GTTS (Google Text-to-Speech) to convert the text of the story into audio narration. Sync the audio with the sequence of images to create a multimedia experience.

PDF Generation: Offer users the option to download the compiled story as a PDF document. The PDF can include both the textual content and the corresponding images, providing a convenient format for offline viewing and sharing.

5.2 User Interface Design

- Sign In/Sign Out Page: The sign-in page features input fields for the user's email address and password. A "Sign In" button allows users to submit their credentials for authentication. If authentication is successful, users are redirected to the home page. Otherwise, an error message is displayed.

- Home Page: Upon successful sign-in, users are directed to the home page. The home page includes a prompt box where users can write their stories. After writing the story, users can initiate the process of generating a flip book by clicking a button or selecting an option.

- Browse Page: The browse page presents users with a collection of classic children's stories. Each story is displayed as a clickable item, featuring its title and book cover images. When a user clicks on a story, its corresponding flip book is dynamically generated, providing an interactive experience.

5.3 Database Design

The detailed database design and its schema is expected in this section. The database used in the work can be mentioned here. The reason for choosing the database can be substantiated in this section.

In our project, we've implemented two databases to facilitate different aspects of our storytelling platform. The first database serves as a repository for user authentication, featuring columns for email addresses and passwords, enabling users to sign in and out securely. Meanwhile, the second database is dedicated to housing a collection of beloved classic children's stories. This database comprises three columns: an ID for each story, the title of the story, and the story text itself. By structuring our data in this manner, we ensure that users can securely access the platform while also providing them with a rich selection of timeless tales to explore and enjoy.

Database 1: User Authentication

Table Name: Users

Columns:

- email (VARCHAR): User's email address.
- password (VARCHAR): password for user authentication.

Database 2: Classic Children Stories

Table Name: Stories

Columns:

- storyId (Primary Key, VARCHAR): Unique identifier for each story.
- storyTitle (VARCHAR): Title of the classic children's story.
- storyText (TEXT): Full text of the story.

5.4 Description of Implementation Strategies

Text Summarization and Coreference Resolution: We utilized the Hugging Face Transformers library in Python to implement the LED summarization model and the Fastcoref python package and used the transformers library for model loading, inference, and post-processing of summarized text.

Text-to-Image Generation: Implemented the Stable Diffusion XL Text-to-Image model

and preprocessed textual descriptions of sentences into a format compatible with the text-to-image model's input requirements.

Video and Audio Generation: Used the moviepy library in Python for video generation and combine the generated images into a video sequence using moviepy library. Utilized GTTS (Google Text-to-Speech) library for text-to-audio conversion and synchronize audio narration with the video frames.

PDF Generation: Used the FPDF library in Python for PDF generation and designed the PDF template to include both textual content and corresponding images and generated PDF documents programmatically by embedding images and text using FPDF.

User Interface Design: Implemented the front end using HTML, CSS, and JavaScript for web-based interfaces and design responsive layouts to ensure compatibility across different devices and screen sizes.

Database Implementation: We used SQL databases such as MySQL for storing user authentication and Microsoft Excel datasheet for story data and designed database schemas to efficiently store and retrieve user information and story content.

Chapter 6

Results and Discussions

6.1 Overview

The project successfully transforms input stories into engaging flip books, enhancing user experience with multimedia narration.The project workflow begins with users inputting a story, followed by LED model summarization to condense longer narratives while retaining their core meaning. After summarization, to ensure consistency in characters throughout the story, we utilize fastcoref. The summarized story undergoes sentence splitting to prepare for Stable Diffusion model processing. Each sentence is then fed into the Stable Diffusion model, which generates corresponding images depicting the core idea or context of the text. These images are organized sequentially to form a flip book, mimicking the narrative flow. Concurrently, GTTS (Google Text-to-Speech) converts the text into audio for narration, synchronized with the flip book images using MoviePy python library for video generation. Finally, users are provided with the option to download the entire storybook in PDF format, facilitating offline access and sharing. This process integrates advanced AI models and multimedia elements to create an immersive storytelling experience while enhancing accessibility and user engagement.

6.2 Testing

Figure 6.1: Login Page

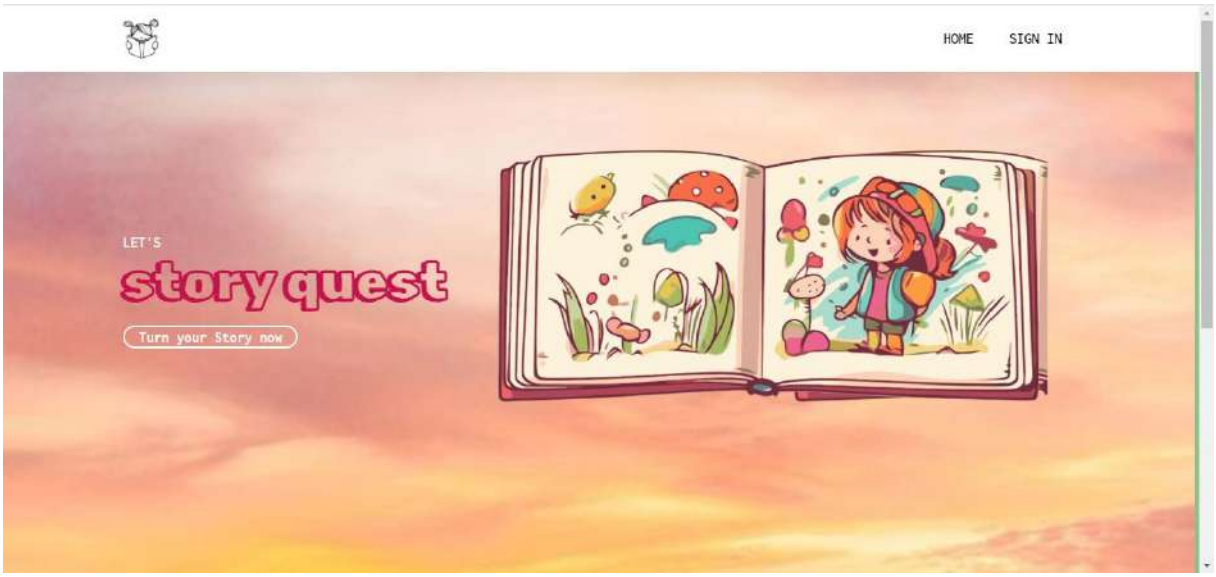


Figure 6.2: Home Page

Figure 6.3: Story input box

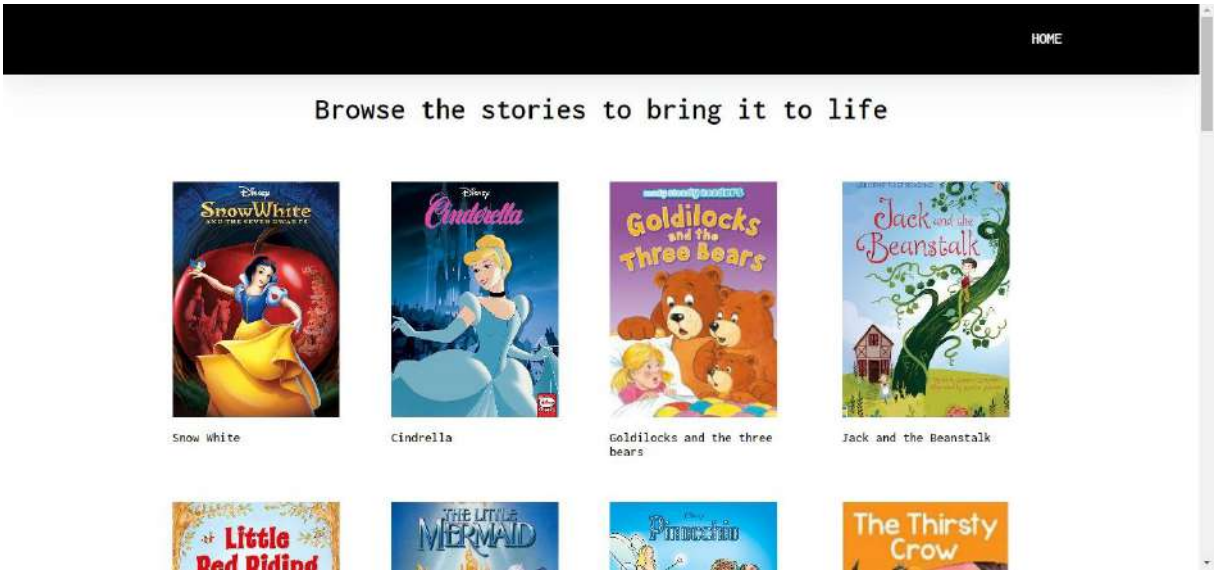


Figure 6.4: Available stories page

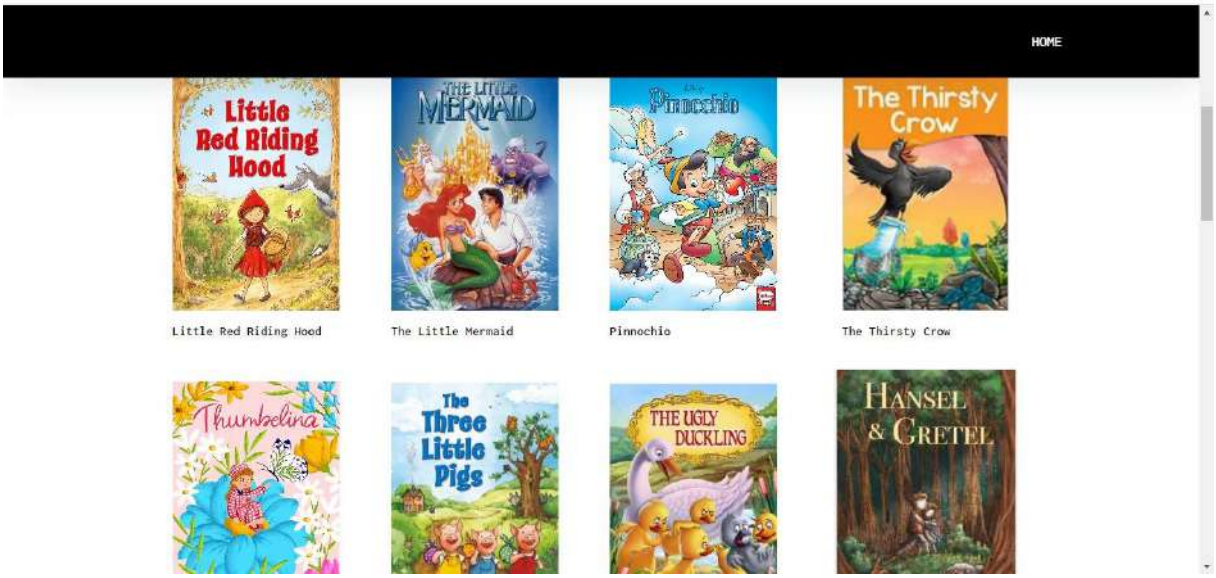


Figure 6.5: Available stories page

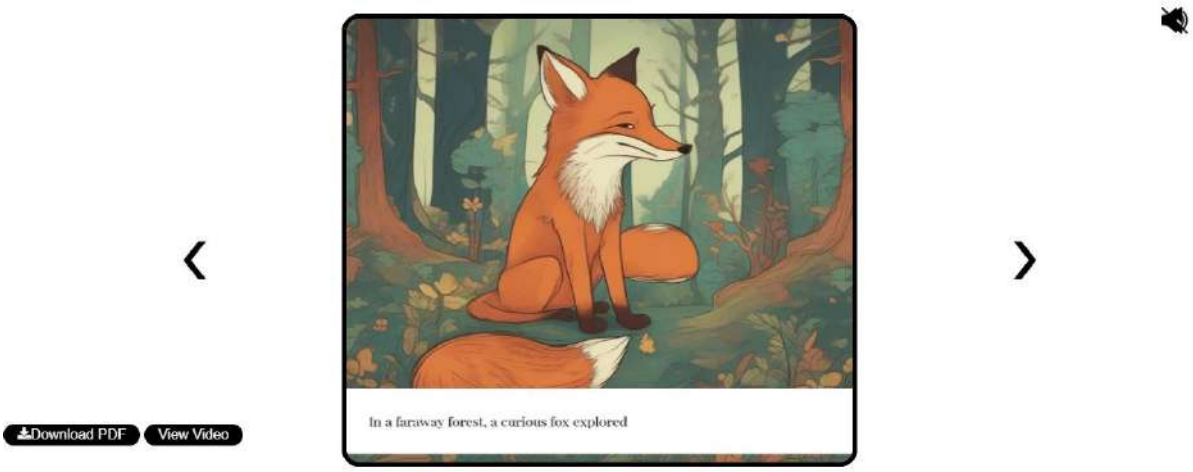


Figure 6.6: Generated picture book

Figure 6.7: Generated picture book

Figure 6.8: Generated picture book

Figure 6.9: Generated picture book

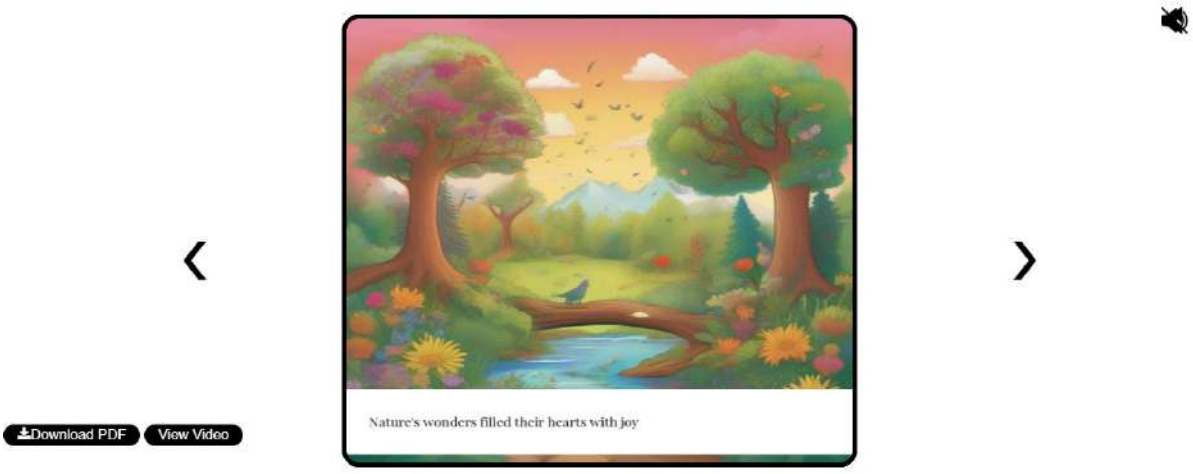


Figure 6.10: Generated picture book

Figure 6.11: Video of picture book with audio

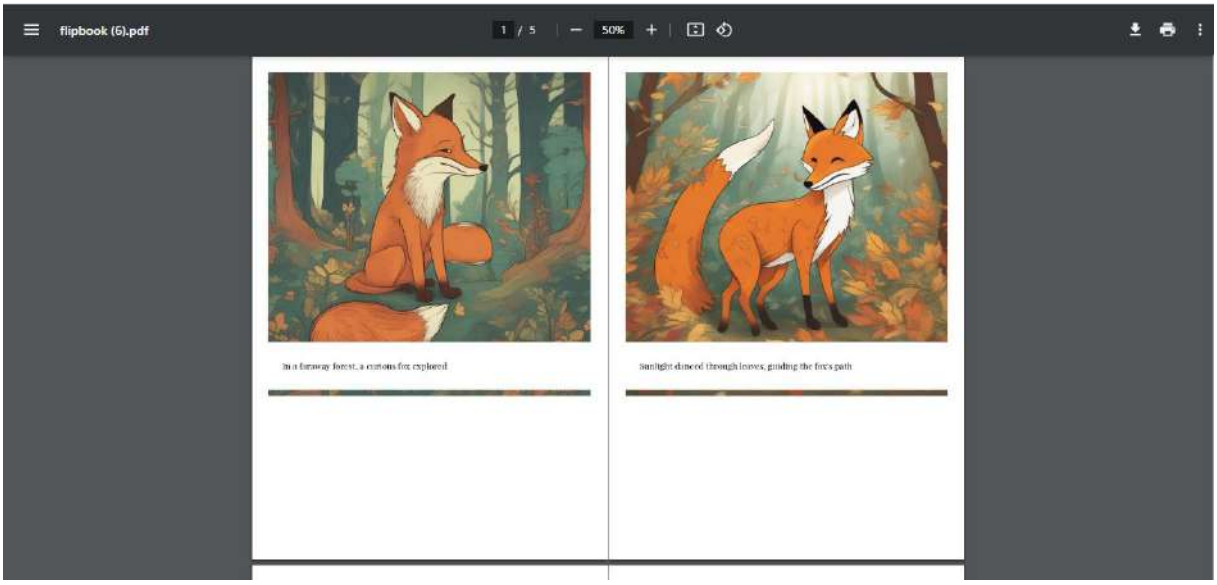


Figure 6.12: PDF file of the generated picturebook

6.3 Discussion

By implementing our project StoryQuest we found that even though coref was used to generate consistent characters, however for certain stories, the characters were not consistent. The project also takes approximately 2-6 minutes based on the length of the story to generate the final output.

Chapter 7

Conclusions & Future Scope

In conclusion, a significant advance in the field of artificial intelligence and creative content creation has been achieved by integrating stable diffusion models and summarization models into automated production of picture books from written texts. The combination of these two advanced models makes it possible to transform textual narrative into visually engaging and coherent books. The stable diffusion model is critical in preserving the visual storyline's consistency and seamless flow. Furthermore, by extracting crucial parts from the textual story, the text summarization model aids to the efficiency of the picture book generation process by effectively extracting key information for the production of visually appealing scenes that capture the essence of the story.

'StoryQuest' merges text and images providing a combination of narrative depth and visual richness. This invention accommodates to a wide range of reader tastes and opens up new options for creative storytelling expression. The project's future scope includes expanding its capabilities to include dynamic interactive features, personalised storytelling experiences, and interaction with upcoming technologies such as augmented reality, which will provide users with an even more engaging and immersive platform. Exploring prospective relationships with educational institutions and publishers may also help to broaden its impact in the sectors of education and literature.

References

[1] u, K.; Kim, H.; Kim, J.; Chun, C.; Kim, P. A Study on Webtoon Generation Using CLIP and Diffusion Models. Electronics 2023, 12, 3983. https://doi.org/10.3390/electronics12183983

[2] Yu, K.; Kim, H.; Kim, J.; Chun, C.; Kim, P. A Study on Generating Webtoons Using Multilingual Text-to-Image Models. Appl. Sci. 2023, 13, 7278.

[3] Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text-conditional image generation with clip latents. arXiv 2022, arXiv:2204.06125.

[4] Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv 2021, arXiv:2112.10741

[5] Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2426–2435

[6] H. Tan, X. Liu, B. Yin and X. Li, "Cross-Modal Semantic Matching Generative Adversarial Networks for Text-to-Image Synthesis," in IEEE Transactions on Multimedia, vol. 24, pp. 832-845, 2022, doi: 10.1109/TMM.2021.3060291.

[7] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, J. Y. Lim and A. Alqahtani, "Recent Advances in Text-to-Image Synthesis: Approaches, Datasets and Future Research Prospects," in IEEE Access, vol. 11, pp. 88099-88115, 2023, doi: 10.1109/ACCESS.2023.3306422.

[8] M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi:10 1109/TASLP.2020.3006731

[9] Zakraoui, J.; Saleh, M.; Al-Maadeed, S.; Alja'am, J.M. A Pipeline for Story Visualization from Natural Language. Appl. Sci. 2023, 13, 5107. https://doi.org/10.3390/app13085107

[10] Waseemullah, Zainab Fatima, Shehnila Zardari, Muhammad Fahim, Maria Andleeb Siddiqui, Ag. Asri Ag. Ibrahim, Kashif Nisar, and Laviza Falak Naz. 2022. "A Novel Approach for Semantic Extractive Text Summarization" Applied Sciences 12, no. 9: 4479. https://doi.org/10.3390/app12094479

[11] https://towardsdatascience.com/stable-diffusion-using-hugging-face-501d8dbdd8

Appendix A: Presentation

Story Quest

100% Presentation

Guide: Dr. Uma Narayanan

Team 12
Navya Rony A
Niya George N
Riya Thomas
Shikha Mariam Joseph

May 2, 2024

Overview

Problem Definition

- ”Story Quest” is an entertainment platform. It seeks to automate the conversion of written stories into dynamic visual representations, thus enhancing the storytelling experience.

Project objective

1. **Create a Sequence of Visuals:** Develop a system to produce a series of images that illustrate a story.

2. **Enhance Learning Experience:** The primary objective of this project is to enhance the learning experience, especially for young children.

3. **Automated Story Visualization:** To develop a system that can automatically convert text-based stories into visual images, making them more engaging and comprehensible.

Novelty of Idea and Scope of Implementation

1. **Instant Transformation:** StoryQuest stands out by instantly converting written stories into picture books with minimal effort from the user.

2. **Multimedia Options:** Unlike other apps, StoryQuest offers the unique feature of converting picture books into dynamic slideshows with audio narration, enhancing the storytelling experience.

Literature Review

A Study on Webtoon Generation Using CLIP and Diffusion Models [5]

- Advantages
    1. CLIP's ability to identify the most similar image to a given text.
    2. The use of the depth-to-image model within Stable Diffusion excels in producing realistic images.
- Disadvantages
    1. Limited Control over Style and Content.
    2. The multi-step process involving training the CLIP model and utilizing the depth-to-image model adds complexity to the overall method.

Literature Review

A Study on Generating Webtoons Using Multilingual Text-to-Image Models [2]

- Advantages
  - Multilingual Capabilities for Diverse Audiences.
  - BERT model used for feature extraction, ensuring that the generated images are influenced by high-quality and cross-lingual learned features.
- Disadvantages
  - Accuracy of DCGAN when trained with webtoon dataset was significantly low.
  - The conversion process using CartoonGAN may not fully capture the intricate details of webtoons.

Literature Review

Hierarchical Text-Conditional Image Generation with CLIP Latents [1]

- Advantages
  - Improved Image Diversity without Loss in Realism.
  - Efficient Language-Guided Manipulations.
- Disadvantages
  - The use of CLIP embeddings may result in a lack of fine-grained control over attribute representation in generated images
  - Computational Complexity of Diffusion Models.

Literature Review

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models [3]

- Advantages
  - Photorealistic Image Generation with Guided Diffusion.
  - It uses CLIP guidance and classifier-free guidance to assist in the image generation process.
- Disadvantages
  - Computational Complexity and Resource Requirements.
  - Filtering training data for safety can limit dataset diversity and representation.

Literature Review

DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation [4]

- Advantages
  - The diffusion model can achieve high-quality image synthesis.
  - Flexibility and broad utility of DiffusionCLIP in creative image synthesis tasks
- Disadvantages
  - Fine-tuning the diffusion model for specific texts is time-consuming and computationally expensive.
  - Dependency on Pre-trained CLIP Model.

Methodology

1. Story Submission: Users submit written stories to the platform.

2. Length Check: Check if the submitted story is longer than 200 words. If it is, pass it to the LED summarizer. If not, the story is passed for coreference resolution.

3. LED Summarization: Pass the story through an LED (Longformer Encoder-Decoder) summarization model. This model can effectively condense longer texts while retaining the essential information. The summarized version of the story is passed for coreference resolution.

Methodology

4. Coreference Resolution: Perform coreference resolution on the summarized story using fastcoref. This step ensures that pronouns and other referring expressions in the text are correctly linked to their corresponding entities, enhancing the overall coherence and readability of the story.

5. Sentence Splitting: Split the story into individual sentences. This step is crucial for feeding the sentences into the text-to-image model separately.

6. Text-to-Image Generation: Utilize the Stable Diffusion XL Text-to-Image model to generate images corresponding to each sentence of the story. This model can produce high-quality images based on textual descriptions.

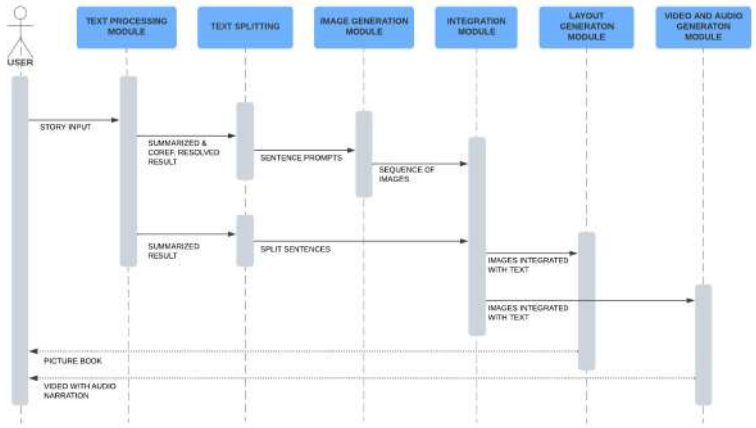
Methodology

7. Image Compilation: Combine all the generated images into a cohesive sequence, pre- serving the narrative flow of the story.

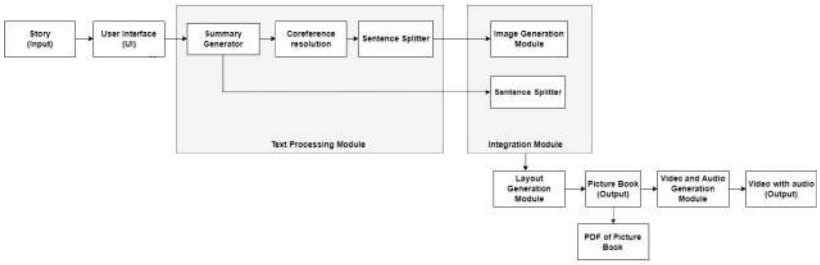
8. Video Generation: Provide an option for users to watch the compiled images as a video with accompanying audio. Use GTTS (Google Text-to-Speech) to convert the text of the story into audio narration. Sync the audio with the sequence of images to create a multimedia experience.

9. PDF Generation: Offer users the option to download the compiled story as a PDF document. The PDF can include both the textual content and the corresponding images, providing a convenient format for offline viewing and sharing.

Sequence Diagram



Architecture Diagram



Results

Figure: User login page

Results

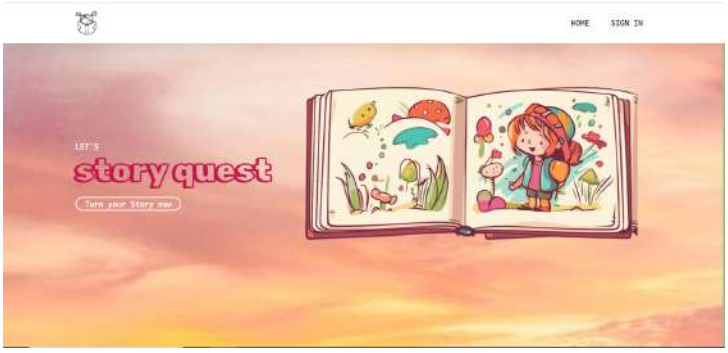


Figure: Home Page

Results

Figure: Input story box

Results

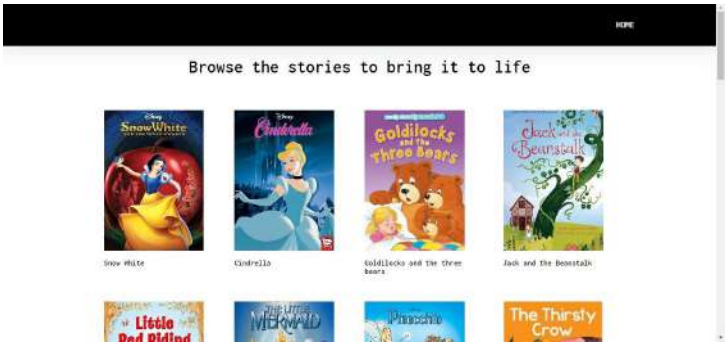


Figure: Available stories page

Results

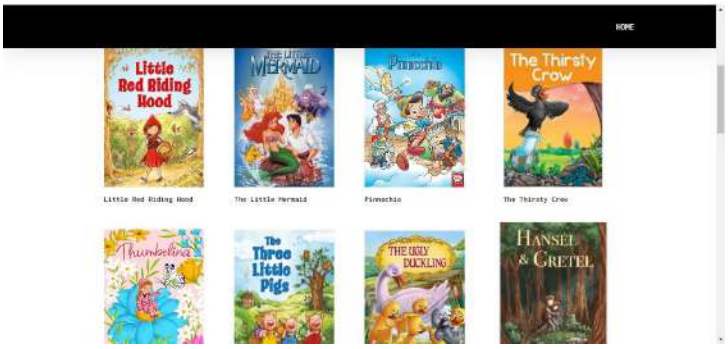


Figure: Available stories page

Results

Figure: Generated picture book

Results

Figure: Generated picture book

Results

Figure: Generated picture book

Results

Figure: Generated picture book

Results

Figure: Generated picture book

Results

Figure: Video of picture book with audio

Results

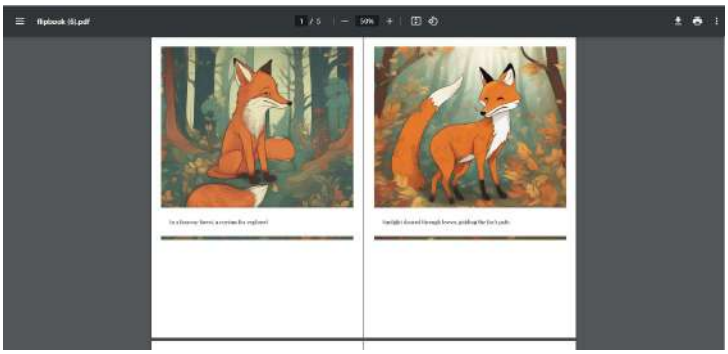


Figure: PDF file of the generated picture book

Task Distribution

1. Niya George - Text Processing Module, Layout Design
2. Riya Thomas - Text Processing Module, Layout Design
3. Navya Rony A - Image Generation module, UI Design
4. Shikha Mariam Joseph - Image Generation module, UI Design

Future Scope

1. Localization: Support multiple languages and cultural preferences to cater to a diverse user base globally.
2. Mobile Application: Develop a mobile application version of the project to reach a wider audience and provide on-the-go access to multimedia content generation.

Conclusion

This project involving the generation of digital picture books as a slideshow is an innovative undertaking that has the potential to greatly enhance the way information is conveyed and presentations are made.

References

1. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical text‐conditional image generation with clip latents. arXiv2022, arXiv:2204.06125.
2. Yu, K.; Kim, H.; Kim, J.; Chun, C.; Kim, P. A Study on Generating Webtoons Using Multilingual Text‐to‐Image Models. Appl. Sci. 2023, 13, 7278. https://doi.org/10.3390/app13127278
3. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Chen, M. Glide: Towards photorealistic image generation and editing with text‐guided diffusion models. arXiv 2021, arXiv:2112.10741
4. Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text‐guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp.2426–2435

References

5. Yu K, Kim H, Kim J, Chun C, Kim P. A Study on Webtoon Generation Using CLIP and Diffusion Models. Electronics. 2023; 12(18):3983. https://doi.org/10.3390/electronics12183983
6. H. Tan, X. Liu, B. Yin and X. Li, "Cross-Modal Semantic Matching Generative Adversarial Networks for Text-to-Image Synthesis," in IEEE Transactions on Multimedia, vol. 24, pp. 832-845, 2022, doi: 10.1109/TMM.2021.3060291.
7. Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, J. Y. Lim and A. Alqahtani, "Recent Advances in Text-to-Image Synthesis: Approaches, Datasets and Future Research Prospects," in IEEE Access, vol. 11, pp. 88099-88115, 2023, doi:10.1109/ACCESS.2023.3306422.
8. M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi:101109/TASLP.2020.3006731

References

9. Zakraoui, J.; Saleh, M.; Al-Maadeed, S.; Alja'am, J.M. A Pipeline for Story Visualization from Natural Language. Appl. Sci. 2023, 13, 5107. https://doi.org/10.3390/app13085107
10. Waseemullah, Zainab Fatima, Shehnila Zardari, Muhammad Fahim, Maria Andleeb Siddiqui, Ag. Asri Ag. Ibrahim, Kashif Nisar, and Laviza Falak Naz. 2022. "A Novel Approach for Semantic Extractive Text Summarization" Applied Sciences 12, no. 9:4479. https://doi.org/10.3390/app12094479

Thank you

Guide: Dr. Uma Narayanan

Team 12
Navya Rony A
Niya George N
Riya Thomas
Shikha Mariam Joseph

May 2, 2024

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)
A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems

in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes(CO)

SNO	DESCRIPTION
CO1	Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).
CO2	Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).
CO3	Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).
CO4	Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).
CO5	Identify technology/research gaps and propose innovative/cre- ative solutions (Cognitive knowledge level: Analyze).
CO6	Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).