GenAI Empowered Script to Storyboard Generator

Govind A*[1] , Ahad Anzar*[2] , Aiswarya Anil Nair*[3] , Rohith Syam*[4]
ORCID: [1]0009-0002-2573-4003 ,[2]0009-0003-0723-3608 , [3]0009-0006-5232-7728 , [4]0009-0008-9410-0710
*Department of Computer Science Engineering, Sree Chitra Thirunal College of Engineering, Trivandrum, India

***Abstract*—The research presents an innovative solution for automating the generation of storyboards from screenplays through the integration of advanced AI technologies. By taking a screenplay as input, the system utilizes cutting-edge neural networks to recognize characters and objects, enhancing scene comprehension. A refined Bi LSTM model is employed to extract the nuanced emotional tones embedded within in dialogues in each scene, providing valuable insights into character dynamics and narrative depth. Through the application of regular expressions, key scene attributes such as time, place, and location are extracted to establish contextual relevance.A Facebook BART-large-CNN model is then employed to generate concise summaries of each scene, enhancing comprehension efficiency. Through script summarization and tag extraction from the Movie Plot Synopses with Tags dataset, a smooth transition between scenes is enabled. These extracted features are structured into a prompt using a rule based approach, facilitating seamless integration into the subsequent creative phase. Finally, a stable diffusion model is employed to generate scene-by-scene coherent storyboard, incorporating all extracted elements to streamline the visual storytelling process where coherency is achieved with the help of cosine similarity between prompts. This comprehensive approach not only automates tedious tasks but also enhances creativity and efficiency in storyboard creation.**

***Index Terms*—Generative AI, storyboard generation, prompt generation, scene summarising, script summarising, Generative Coherence**

I. Introduction

In the realm of cinematic storytelling, the process of transitioning from a written screenplay to visual representation is a pivotal yet intricate endeavor. Traditionally, this transition has been a labor-intensive task, requiring meticulous attention to detail from storyboard artists and filmmakers. However, recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies have paved the way for innovative solutions aimed at automating and streamlining this process.

A study was done on comprehending an imagination into screen play and it sheds light on the divergent attitudes towards the screenplay in American independent cinema, emphasizing its dual role as both a creative cornerstone and a point of departure for innovation. The influence of institutions such as the Sundance Institute and American Playhouse underscores the screenplay's enduring significance in shaping cinematic narratives and fostering artistic expression [1].

In the context of animation, the symbiotic relationship between screenplay and storyboard emerges as a central theme. Despite animation's visual nature, the screenplay remains indispensable for providing narrative structure and character development. Storyboards, on the other hand, serve as visual blueprints, translating the screenplay into meticulously planned sequences. This symbiosis between screenplay and storyboard not only elevates the quality of animated visuals but also underscores the nuanced interplay between written narrative and visual representation, as elucidated by [2].

In addition, examination of the collaborative nature of screenwriting in the creation of the iconic film "Citizen Kane" is done. It explores the contributions of Orson Welles and Herman J. Mankiewicz, revealing the blurred lines of individual authorship in cinematic storytelling. By analyzing historical records and differing perspectives, the chapter challenges conventional notions of singular authorship in filmmaking, emphasizing the complex dynamics involved in crafting cinematic masterpieces [3].

Furthermore, a groundbreaking NLP approach is introduced which is designed to semi-automatically convert screenplays into animated films, thereby revolutionizing the filmmaking process. This approach, built upon the Unity platform, breaks down screenplays into scenes, identifies characters and actions, and determines shot types based on the screenplay's content. By automating tasks such as character placement and scene labeling, this system not only enhances efficiency but also democratizes the filmmaking process, making it more accessible to a broader spectrum of creators [4].

This research aims to merge traditional storytelling with AI-driven technologies. By combining neural networks and transformers with screenplay analysis, a solution for automating storyboard generation from screenplays is proposed. This promises to streamline filmmaking, boost creativity, and redefine visual storytelling while maintaining generative coherence for scene continuity.

In the subsequent sections of this paper, a robust methodology detailing the integration of AI technologies, the extraction of key narrative elements, and the generation of scene-by-scene storyboards is demonstrated. Through empirical evaluations and case studies, the study aims to demonstrate the efficacy and potential in revolutionizing the art of cinematic storytelling.

II. Literature Survey

Reference [5] discusses effective prompt generation methods for efficient transfer learning in NLP tasks. Key findings include optimizing prompt position (0 for single-sentence tasks, 1 for sentence-pair tasks), ensuring prompt length stability across models, and highlighting potential risks in

979-8-3503-9121-3/24/$31.00 ©2024 IEEE
DOI 10.1109/FMLDS63805.2024.00085

applying prompt-based methods in high-stakes contexts due to vulnerabilities in online stored backbone models

Reference [6] surveys recent advancements in text-to-image generation, including methods like zero-shot generation, scene-based generation, hierarchical text-conditional image generation, and high-resolution synthesis. It explores techniques such as training-free guidance, compositional visual generation, attention-based semantic guidance, and conditional control.

Reference [7] highlights the multifaceted nature of creativity, as defined by scholars like Csikszentmihalyi and Boden. It explores the emergence of generative AI tools, echoing historical shifts in creative paradigms. Studies emphasize the need for comprehensive evaluation frameworks, ethical considerations, and effective human-AI interaction strategies. Major takeaways include the importance of interdisciplinary perspectives, robust evaluation methodologies, and ethical awareness in the development of AI-driven creative systems.

Reference [8] discusses about the ROUGE toolkit aids in text summarization by providing a set of metrics, such as precision, recall, and n-gram statistics, to evaluate the similarity between system-generated summaries and human-generated ones. This allows for a quantitative assessment of the effectiveness of text summarization systems.

Reference [9] discusses about how Emotion detection benefits from abundant self-expression text in social media, enabling fine-grained affective information for decision-making. Advancements in classification algorithms like neural networks and ensemble methods enhance accuracy. Exploring methodologies beyond Bag-of-Words, such as considering language flow, is crucial.

III. Materials

Kaggle served as a primary platform for accessing datasets while Hugging Face was a major source for pre-trained models.

A. Datasets

TABLE I: Datasets used and its purpose

Dataset	Purpose
emotion-detection-from-text	This dataset containing 40,000 tweets annotated with 13 different emotions was utilized for a multiclass classification model for emotion detection in text.
mpst-movie-plot-synopses-with-tags	This dataset containing around 14000 movie names along with their plot synopsis collected from IMDb and Wikipedia and tags was utilized to extract tags from the screenplay.

Note: The above mentioned datasets were obtained from Kaggle in CSV format

TABLE II: Overview of models used

Models Used	Fine Tuned or Not	Purpose	Reason
Bi LSTM	Yes	Emotion detection of scenes.	Utilized for its effectiveness in capturing nuanced emotional expressions within scenes. [10]
facebook/bart-large-cnn	No	Scene and script summarization.	Chosen due to its capability to generate concise summaries of scenes and scripts.
Jean-Baptiste/roberta-large-ner-english	No	Extracting character information from scenes.	Large NER model excels in extracting character information from scenes due to its optimized architecture, enabling effective language understanding, especially in handling long sequences like those found in financial documents. [21]
bert-base-uncased	Yes	Extracting tags for screenplays.	BERT-base excels in multiclass classification with its deep bidirectional Transformer architecture and diverse pre-training tasks like MLM and NSP, surpassing other models. [13]
stabilityai/stable-diffusion-2	No	Converting prompts to images for each scene.	Stable diffusion2 outperforms other models for image generation by ensuring stability during training and effectively handling diverse datasets, resulting in high-quality output images with minimal artifacts. [22]

Note: The models listed are utilized for various purposes related to scene analysis and processing.

B. Models Used

IV. Methodology

In this section, the detailed methodology used is explained for the screenplay to prompt workflow as shown in Fig. 1. This section gives an overview of the modules and models used.

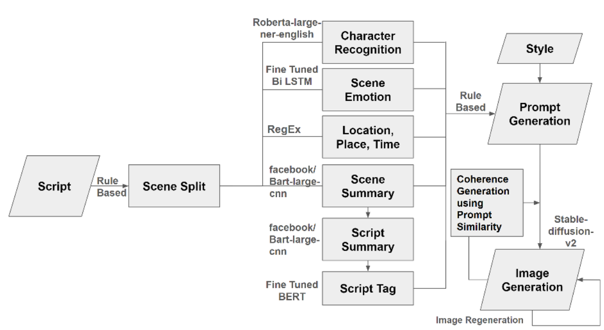

Fig. 1: Screenplay to Prompt Workflow

A. Script to Scene Split

Scene changes are detected by locating the words in shot headings. The script is then split according to a rule-based approach by detecting shot headings *INT/EXT/INTERIOR/EXTERIOR*. The content starting from the initial shot heading until encountering the next heading is stored as scenes and the process continues until all the scenes are stored.

B. Character Identification using NER

The paper [11] investigates the efficacy of four pre-training models—BERT, ERNIE, ERNIE2.0-tiny, and RoBERTa—in Named Entity Recognition (NER) tasks. It discusses the architectures and pre-training tasks of these models before employing them through fine-tuning. During fine-tuning, a fully connected layer and a Conditional Random Field (CRF) layer are incorporated after the output of the pre-training models. Results indicate a significant improvement in recognition performance with the use of these pre-trained models. Particularly, the study highlights RoBERTa as well-suited for NER tasks, potentially due to its modified masking strategy and removal of the Next Sentence Prediction (NSP) task. RoBERTa shares similarities with BERT but employs a dynamic masking strategy. The paper suggests further exploration of model structures tailored to different downstream tasks for future research.

The model was fine-tuned from roberta-large on conll2003 dataset and was validated on emails or chat data and outperformed other models on this type of data specifically. In particular, the model seems to work better on entity that don't start with an upper case. *PER* entities helps in the extraction of characters in each scene.

C. Scene Emotion Identification

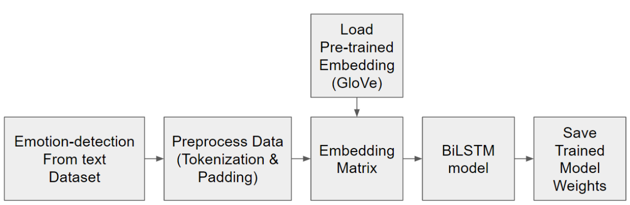

Fig. 2: Emotion Extraction Pipeline

The process begins with loading and exploring the dataset, focusing on tweets and their associated emotions. Textual data undergoes preprocessing steps like tokenization and padding before being fed into a Bidirectional LSTM model, enriched with pre-trained GloVe embeddings for better understanding of word semantics. During training, the model is compiled with suitable loss and optimizer functions, with model checkpointing to preserve optimal iterations. Evaluation metrics such as accuracy and confusion matrices assess model performance on a validation set, while the trained model is saved for future use.

A prediction function enables emotion classification in user-provided text, tailored for fine-tuning sentiment analysis models. Emotions included are 'empty', 'sadness', 'enthusiasm', 'neutral', 'worry', 'surprise', 'love', 'fun', 'hate', 'happiness', 'boredom', 'relief' and 'anger'. The dialogues are removed from the scenes using regular expression as only emotion is considered since token limit needs to be clipped.

BiLSTM excels in text classification, but faces challenges due to text data's high dimensionality and sparsity, as well as the complexity of natural language semantics. AC-BiLSTM, integrating attention mechanism and convolutional layers, addresses these challenges. It extracts higher-level phrase representations, accesses both preceding and succeeding context information, and enhances focus on relevant details, ensuring accurate classification. Experimental results across datasets confirm its superiority, capturing both local and global semantic features [10].

D. Location, Place & Time Extraction

From the shot headings in the screenplay, the location, place and time of the scene is extracted with the help of regular expressions. A RegEx, or Regular Expression, is a sequence of characters that forms a search pattern. RegEx can be used to check if a string contains the specified search pattern. Since all the screenplays contain the same format of headings, pattern matching is done to extract the required information.

E. Scene Summarising

After the splitting of the screenplay into scenes and extraction of relevant information from each scene, it is further summarized to undergo prompt generation. For this purpose, a Facebook BART-large-CNN model is utilized. The model, initialized from BART-large, shares the same architecture, enabling it to process up to 16000 tokens by leveraging BART-large's position embedding matrix. It aims to generalize well and be effective in summarizing lengthy text. The maximum length of the summary is set to 55 tokens to accommodate prompt generation for storyboard creation.

The importance of text summarization examined in today's information-rich environment, where search engines like Google, Yahoo, and Bing play a central role. It evaluates three pre-trained models—google/pegasus-cnn-dailymail, T5-base, and facebook/bart-large-cnn—across three datasets (CNN-dailymail, SAMSum, and BillSum), each with 2000 examples. Using metrics like ROUGH and BLEU, the study compares the models' performance in generating concise summaries for different types of text sources [12].

F. Script Summarising

A summary of the script is obtained with the help of a Transformer based model, led-large-book-summary. Due to the limitation on the number of tokens that the summarizer can take as input, the entire script cannot be passed to obtain the plot of the movie. Instead, the summaries of the scenes are concatenated and passed to the model and the movie synopsis is generated. Furthermore, [13] explores active learning strategies with BERT for text summarization, highlighting performance variations across datasets. It demonstrates that BERT-based active learning can reduce labeling efforts while maintaining summarization quality. However, no single strategy emerges as definitively superior, with metrics like diversity and runtime efficiency offering insights into strategy effectiveness. Overall, the study provides valuable insights into the challenges and effectiveness of using BERT-based active learning for script summarization tasks.

G. Script Tagging

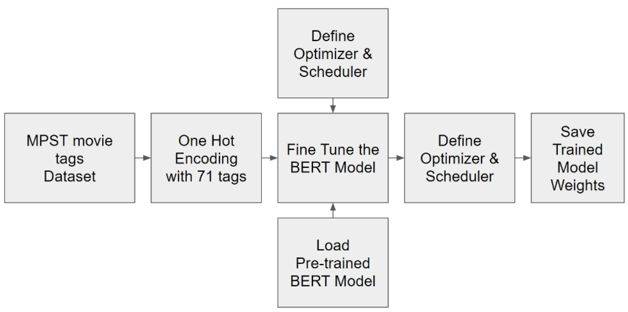

Fig. 3: Script Tagging Pipeline

Fine-tuning was performed using a dataset containing plot synopses and various associated tags for multitag classification. The dataset underwent preprocessing to handle missing values and split the tags accordingly. A one-hot encoded DataFrame was then generated to represent the data. Utilizing the BERT tokenizer, plot synopses were encoded, and a custom dataset class was created to prepare the data for training. A pre-trained BERT model was employed to establish the architecture for sequence classification, which was fine-tuned using the multitag dataset. Optimization was carried out using the AdamW optimizer along with a linear scheduler. The model was trained over multiple epochs, involving gradient calculation, clipping, and parameter updates. Evaluation was performed on the validation set to gauge accuracy. Finally, the fine-tuned model was saved and loaded for inference as shown in Fig 3. Additionally, a function was provided to predict tags for given plot synopses using the fine-tuned model and BERT tokenizer. Script tags encompass categories like 'cult', 'horror', 'gothic', 'murder', 'atmospheric' or 'violence'.

H. Prompt Generation

To craft the prompt for the image generation model, a rule-based approach is utilized. Additionally, users can also specify the desired style, such as 'pencil drawing', 'sketch', 'cartoon', 'animation' or 'photoreal'. These predetermined options guide the AI in tailoring the image generation process along with emotional tone and narrative context, as depicted in Fig. 4.

Within the prompt structure, the desired style of the generated image can be specified as:

"Generate" [DRAWING TYPE] *"for"* [SCRIPT TAG] *"movie set in"*[LOCATION,PLACE,TIME] *"conveying"* [EMOTION] *"emotion and"* [NUMBER OF CHARACTERS] *"main characters for scene:"* [SCENE SUMMARY]

Furthermore, [14] explores text-to-image prompt generation, emphasizing elements like subject, medium, technique, genre, mood, tone, lighting, and resolution. This structured understanding offers valuable insights for educators and students. The study aims to validate its approach empirically, enabling better integration of text-to-image prompts in education and creativity. It also discusses ethical and legal aspects of AI-generated artwork, including copyright and fair compensation.

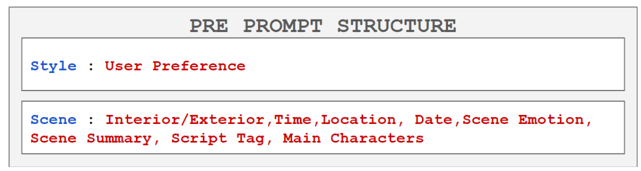

Fig. 4: Structure of the Pre Prompt used for Prompt Generation

I. Coherence Generation

A cohesive narrative thread is essential to ensure that a storybook does not appear as a collection of disjointed images. Conditional generation is leveraged to maintain this coherence. When creating an image I_i, it is conditioned not only on the corresponding narrative prompt xi but also on the most contextually congruent image I_j generated up to I_{i-1}. This contextual congruence is determined by assessing the similarity between the prompts associated with I_i and I_j, where $j<i$.

Additionally, [15] presents a novel method for extending comic narratives using Stable Diffusion models. By generating storylines and visuals, the study aims to maintain narrative coherence and art style fidelity. It introduces a new evaluation metric, the story score, which assesses the similarity of generated stories to the original plot. Results show improved character fidelity compared to existing benchmarks. The study highlights the potential of this approach for enhancing comic narrative coherence and suggests avenues for future research.

J. Image Generation

The Latent Diffusion Models (LDMs) for high-quality image generation use transformer-based architectures with global self-attention layers and MLPs, supporting conditional generation. They achieve competitive performance while significantly reducing computational resources compared to existing models. Additionally, the paper discusses adversarial training of autoencoder models and highlights regularization techniques for high-fidelity reconstructions. Evaluation metrics demonstrate the effectiveness of LDMs across tasks like super-resolution, inpainting, and semantic synthesis, making them a promising approach to image generation [16].

Thus for image generation the stabilityai/stable-diffusion-2 model was integrated to the end of the pipeline to generate storyboard images. Generative Coherence was also maintained across the image generation model by considering the immediate previous image in order to establish continuity. Additionally,To enhance the creative experience, users are also given the option to regenerate selected scenes. This enables users for a better interactive experience where they can regenerate certain scenes until they satisfy them,

V. RESULTS & DISCUSSIONS

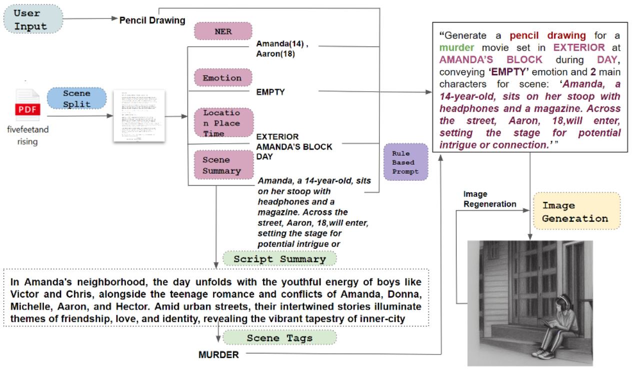

Fig. 5: Image Generation Flow for a scene from short film Five Feet High And Rising(2000)

As portrayed in the above illustration, the process begins with feeding a script, which is then segmented into individual scenes. Following this segmentation, various analyses are conducted, including Named Entity Recognition (NER), emotion detection, scene summarization, and extraction of location, place, and time details. These analyses serve to gather additional information about the scene, such as character details and sentiment. Subsequently, a script summarization step is executed based on the scene's content, resulting in the generation of descriptive scene tags like "murder.", ultimately guiding the generation of the corresponding image. The dialogues are extracted and removed using regular expression before prompting the image model as total tokens need to be clipped.

The figure 6 below demonstrates a comparison between generated images and how the directors conceived the scene from the screenplay. There is a difference between aspect ratios as stabilityai/stable-diffusion-2 can generate only square images as the model was trained on crops of size 512x512 and is a text-guided latent upscaling diffusion model.Using Residual Squeeze VGG16 for image similarity tasks offers a potent solution with its compressed model size and faster training time, facilitating efficient similarity computations [17]. The methodology used for evaluating similarity involves preprocessing input images to ensure uniformity in size, extracting high-level features using the VGG16 convolutional neural network pretrained on ImageNet, computing embeddings for each image, and subsequently quantifying similarity between these embeddings using cosine similarity. This process allows for the comparison of images based on their visual content, with higher similarity scores indicating greater resemblance between images.

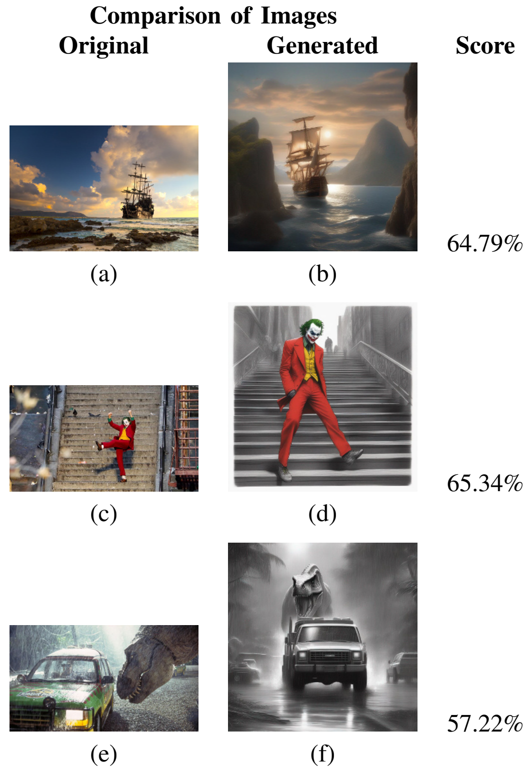

Fig. 6: Comparison of images: original scenes from (a) Pirates of the Caribbean (2003), (c) Joker (2019), (e) Jurassic Park (1993), with (b) representing a photoreal image, (d) representing a sketch drawing, and (f) representing a pencil drawing generated by the model for the respective films.

The figure 7, presents meticulously crafted storyboards, meticulously illustrating seamlessly unfolding scenes, each meticulously designed to maintain coherence throughout. Particularly noteworthy are scenes 3, 4, and 5, where great care has been taken to ensure the preservation of coherence in the architectural elements, meticulously depicted to maintain authenticity and continuity. Similarly, scenes 26, 27, and 28 place a premium on the coherence of characters, meticulously crafted to ensure their believability and consistency within the narrative context. Furthermore, utilizing the pre-trained image embedding model ('clip-ViT-B-32') from the Sentence Transformers library and the PyTorch framework. Images are loaded, encoded into feature vectors, and pairwise similarity scores are computed using cosine similarity. It concludes that scenes 3, 4, and 5 exhibit 73.8 percent coherence, while scenes 24, 25, and 26 display 79.1 percent coherence, providing insights into visual consistency among them. Thus we were able to obtain desirable results across the proposed pipeline maintaining robustness, clarity, coherence and detailing across the entire output.

Additionally, [18] and [19] highlights the intricate ethical terrain surrounding AI image generation, emphasizing the imperative for comprehensive frameworks that tackle issues

Fig. 7: Storyboards generated on screenplay of Five Feet High And Rising(2000) the proposed methodology

ranging from unrealistic portrayals to copyright and data protection concerns. It underscores the transformative impact of AI on traditional media roles and production methods, stressing the necessity for ethical guidelines and regulatory measures to navigate evolving societal norms and address challenges like censorship and shifting business models.

VI. ACKNOWLEDGEMENT

We would like to express our gratitude to the IEEE Industrial Electronics Society for organizing the IEEE IES Generative AI Hackathon. Out of 150 teams, we were honored to reach the final round and become eligible for the project grant. This opportunity significantly contributed to the development of this research.

VII. CONCLUSION

In conclusion, this research presents a comprehensive framework leveraging advanced AI technologies to automate the generation of storyboards from screenplays, with a paramount focus on coherence. By integrating cutting-edge models such as BiLSTM, transformers, and Stable Diffusion V2, the system efficiently extracts scene elements, character dynamics, emotional tones, and maintains coherence throughout the storyboard generation process. Through conditional generation techniques, the coherence of the narrative thread is meticulously preserved, ensuring that the storyboard does not appear as a collection of disjointed images. Empirical evaluations and comparisons with existing benchmarks demonstrate the efficacy of the proposed methodology in maintaining coherence, promising to revolutionize the landscape of visual storytelling. This innovative approach not only streamlines the filmmaking process but also enhances creativity and efficiency by bridging the gap between traditional storytelling methods and emerging AI-driven technologies. With the potential to democratize the filmmaking process and make it more accessible to a wider spectrum of creators, this research marks a significant advancement in the field of cinematic narrative generation, emphasizing the importance of coherence.

In future we can expand the model into video stream cartoon generation as demonstrated in research. Incorporating CORVIS to this can enable use for easy generation [20].

REFERENCES

[1] Zhang, Yeyao, et al. "Generating animations from screenplays." arXiv preprint arXiv:1904.05440 (2019).

[2] Bell, Aaron. "The Symbiotic Relationship of Screenplay and Storyboard: Arguing For the Crucial Role of Screenplay in Animation." (2023).

[3] Buckland, Warren. "Screenplays: Words on the Page." Who Wrote Citizen Kane? Statistical Analysis of Disputed Co-Authorship. Cham: Springer Nature Switzerland, 2023. 29-41

[4] K. Jorgensen, H. Wang and M. Wang, "From Screenplay to Screen: A Natural Language Processing Approach to Animated Film Making," *2023 International Conference on Computing, Networking and Communications (ICNC)*, Honolulu, HI, USA, 2023, pp. 484-490, doi: 10.1109/ICNC57223.2023.10074526.

[5] Wu, Zhuofeng, et al. "IDPG: An instance-dependent prompt generation method." *arXiv preprint arXiv:2204.04497* (2022).

[6] Ding, Yuxuan, et al. "The CLIP Model is Secretly an Image-to-Prompt Converter." Advances in Neural Information Processing Systems 36 (2024).

[7] YILDIRIM, Erdem. "TEXT-TO-IMAGE GENERATION AI IN ARCHITECTURE." Art and Architecture: Theory, Practice and Experience (2022): 97.

[8] Tas, Oguzhan, and Farzad Kiyani. "A survey automatic text summarization." PressAcademia Procedia 5.1 (2007): 205-213.

[9] Seyeditabari, Armin, Narges Tabari, and Wlodek Zadrozny. "Emotion detection in text: a review." arXiv preprint arXiv:1806.00674 (2018).

[10] Liu, Gang, and Jiabao Guo. "Bidirectional LSTM with attention mechanism and convolutional layer for text classification." Neurocomputing 337 (2019): 325-338.

[11] Wang, Yu, et al. "Application of pre-training models in named entity recognition." 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). Vol. 1. IEEE, 2020.

[12] Rehman, Tohida, et al. "An analysis of abstractive text summarization using pre-trained models." Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2021. Singapore: Springer Nature Singapore, 2022.

[13] Prabhu, Sumanth, Moosa Mohamed, and Hemant Misra. "Multi-class text classification using BERT-based active learning." arXiv preprint arXiv:2104.14289 (2021).

[14] Dehouche, Nassim, and Kullathida Dehouche. "What's in a text-to-image prompt? The potential of stable diffusion in visual arts education." Heliyon 9.6 (2023).

[15] Jin, Ze, and Zorina Song. "Generating coherent comic with rich story using ChatGPT and Stable Diffusion." arXiv preprint arXiv:2305.11067 (2023).

[16] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[17] Selimović, Alena, et al. "Analysis of content-aware image compression with VGG16." *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. IEEE, 2018.

[18] Bendel, Oliver. "Image synthesis from an ethical perspective." AI SOCIETY (2023): 1-10.

[19] Romele, Alberto. "Images of Artificial Intelligence: a Blind Spot in AI Ethics." Philosophy Technology 35.1 (2022): 4.

[20] Hwang, Won-Il, et al. "Cinema comics: Cartoon generation from video stream." GRAPP 6 (2006): 299-304.

[21] Khanna, Urvashi, Samira Ghodratnama, and Amin Beheshti. "Transformer-based models for long document summarisation in financial domain." Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022. 2022..

[22] Anderson, Charles, Thomas Taylor, and Christopher Moore. "Advancing AI Image Generation: Unveiling the Mechanisms and Innovations of Stable Diffusion Technology."