Narrative Alchemy: Integrating Human Imagination and Generative AI for Dynamic Story Crafting

Ms. Smita. S. Wagh[1], Hema Sachin Bahl[2], Sakshi Annaso Sutar[3] and Bhagyashri Prakash Yerawar[4]
[1]Professor , Jayawantrao Sawant College Of Engineering Pune.
[2,3,4] Student, Jayawantrao Sawant College Of Engineering Pune.

Abstract-Rapid advancements in AI have enabled the seamless generation of both text and images, paving the way for multimodal storytelling. However, challenges like maintaining coherence, ensuring text-image alignment, and preserving character consistency remain. Story Teller, a Multimodal Large Language Model (MLLM), addresses these issues by predicting both text and visual tokens. It uses a visual de-tokenizer to generate consistent images and a multimodal attention sink mechanism to extend story length beyond training limits. To support this, StoryStream, a high-resolution dataset, is introduced for training and evaluation. This work advances AI-driven storytelling by improving coherence and expanding story length.

Index Terms- Detokenizer, Diffusion Model, MLLM, Story-Teller, SD-XL, Tokenizer, Vit

I. INTRODUCTION

Interleaved image-text content is widely used online, and recent advancements in AI have made generating such content more effective.[1][2] **Multimodal Story Generation** creates immersive narratives by blending text and images, but it faces challenges like maintaining coherence, ensuring character consistency, and generating high-quality visuals. To address these issues, Story Teller is introduced—a model built on Multimodal Large Language Models (MLLMs) that generates both text and images seamlessly.[3][4] It uses pre-trained image tokenizers and de-tokenizers (SD XL, ViT) to produce realistic and consistent visuals. Additionally, a Multimodal Attention Sink Mechanism helps manage longer sequences efficiently, enabling the generation of extended multimodal stories. To further support training and evaluation, the StoryStream dataset is introduced, offering high-quality image-text sequences sourced from animated videos, with improved resolution and narrative depth. New evaluation metrics are also proposed to measure image consistency, narrative engagement, and text-image coherence. Results show that Story Teller outperforms existing models in generating rich, coherent, and visually engaging multimodal stories.[5]

II. PROCEDURE FOR PAPER SUBMISSION

A. Review Stage

Submit your manuscript electronically for review. prepare it in two-column format, including figures and tables(untill it don't fit properly and data is not visible).

B. Final Stage

After your paper has been accepted. The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

C. Figures

As said, to insert images in Word, position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with ―Float over text‖ unchecked).

III. MATH

The project requires a total effort of 34.9 person-months based on its complexity. Since there are 3 developers working on the project, the effort per developer is calculated as 34.9 ÷ 3 ≈ 11.63 months per developer. Using the COCOMO time estimation formula, the estimated project duration is T = 2.5 × (34.9)^0.35 ≈ 8.9 months, which is well within the 12-month deadline. The cost estimation is based on a monthly salary of ₹5,000 per developer. With 3 developers working for 12 months, the total project cost amounts to ₹1,80,000.[6][7]

IV. UNITS

In this paper, all numerical values are represented using the International System of Units (SI) to maintain consistency and clarity. However, where applicable, Centimeter-Gram-Second (CGS) units are also mentioned for comparative analysis. Conversion factors between SI and CGS units are provided where necessary to facilitate ease of understanding.[8]

V. HELPFUL HINTS

A. Algorithms

There are numerous complex heuristic methods used in the Story Teller project to compose meaningful narratives and generate content that can fit the aesthetic of an image.

1) Vision Transformer (ViT): Vision Transformer (ViT), is an innovative design that incorporates transformer based architecture with visual data. Compared to classical convolutional neural networks, ViT incorporates images by partitioning them into patches and then using these patches as sequences as in natural language processing. This is in the recent work by Dosovitskiy et al., 2020, because this approach enables the model to learn long-range interactions and contexts within an image. The use of ViT in the framework of the Story Teller project enriches exactly the comprehension of visual aspects and hence the common positive correlation between text and images in the same story. [9][10]

2) Tokenizer: Tokenizer is one of the most importantly preprocessing tools to transform text data into easily processable format, Word or Subword for models. The need for tokenization in Story Teller project can essentially be tried to theidentification of how text in the form of a story will be processed by the transformer models. Through this process of converting text to tokens, the tokenizer is able to feed into the model, data that will help influence its learning while still retaining meaning. This process is in line with the procedures described by Sennrich et al. , 2016 about tokenization in natural language processing.[11][12]

3) Detokenizer : Unlike the Tokenizer which helps segment the text for training, the Detokenizer performs the reverse work by putting together tokens into human intelligible text. For the case of Story Teller project, the detokenizer makes sure that the outputted story is properly formed so as to make better sense to the end user. Thus, tokenization and detokenization are two essential steps for the generation process, which helps to keep the structural integrity of the generated content and helps to shift between them with less difficulty as it followed from the study identified by Kudo & Richardson (2018). [13[14]

4) SDXL SDXL is a novel model architecture that has been proposed for high return generation tasks such as image and text synthesis. Refining the frameworks of the prior generative models, [15] SDXL introduces multi-scale features and advanced attentions to improve the coherence of the reported stories and generated visuals. Further, its architecture to process contextual information is enhanced, making it ideal for story generation. Other research by Stability AI, in 2023, on SDXL shows that SDXL holds promise to achieving state-of-the-art performance in multimodal applications, in agreement with the objectives of Story Teller proposal. [16][17]

5) LoRA (Low-Rank Adaptation) LoRA (Low-Rank Adaptation) is a method that was developed with the purpose of making subsequent learning more effective on a large model. LoRA has integrated low-rank decomposition in the adaptation process, making it easy the computational cost and memory requirements needed to train large models of data. Such a process, as described by Hu et al. (2021), can be completed more effectively and efficiently to adapt the models for a particular task without having to train again.In the Story Teller project, LoRA allows.[18]

6) Transformer based algorithm Story generation is built upon transformer architectures including GPT-4 and BERT. They are particularly good in identifying the context and consistency of different texts in long stories. Two papers by Vaswani et al. (2017) and Radford et al. (2019) prove the potential of transformers to generate textual content laconically imitating humans and thus fit for narrative purposes.[19]

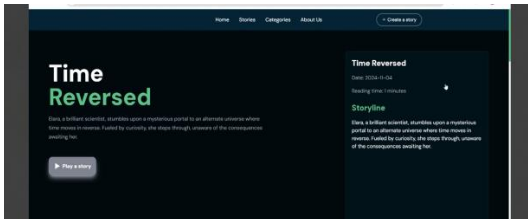
7) Attention Mechanisms Special attention to certain features of the narrative is provided by self- and cross attention. The self- attention technique of the Transformer makes it possible for the model to isolate critical points and characters, as demonstrated by author. [20]

VI. PUBLICATIONPRINCIPLES

The AI
storytelling principles aim to optimize narrative

coherence, text-image matching,and multimodal user interaction. Our system incorporates Multimodal Large Language Models (MLLMs) to create engaging stories with coherent characters, rational narratives, and thematic consistency. By applying the multimodal attention sink technique among others, our system provides perfect unification of textual and visual components, delivering richer storytelling. Moreover, the project focuses on flexibility, allowing customization to fit various storytelling styles and genres.

VII. RESULTS



Time Reversed

Elara, a brilliant scientist, stumbles upon a mysterious portal to an alternate universe where time moves in reverse. Fueled by curiosity, she steps through, unaware of the consequences awaiting her.

Time Reversed

Date: 2024-11-04

Reading time: 1 minutes

Storyline

Elara, a brilliant scientist, stumbles upon a mysterious portal to an alternate universe where time moves in reverse. Fueled by curiosity, she steps through, unaware of the consequences awaiting her.

The home page of our AI-powered storytelling platform is designed to provide users with an engaging and intuitive experience. It features a sleek, dark-themed user interface with a structured layout to ensure readability and ease of navigation.

Navigation Bar :

Located at the top, the navigation bar includes links to essential sections such as "Home," "Stories," "Categories," and "About Us."

A "Create a Story" button is prominently placed, allowing users to initiate their storytelling process effortlessly.

Story Display Section:

The home page showcases featured or recently added stories.

Each story has a bold title and a concise description, drawing readers into the narrative.

Story Details Panel:

On the right side, users can see metadata such as the story's date of creation and estimated reading time.The storyline is summarized, providing a quick preview of the content.

Call-to-Action Button:

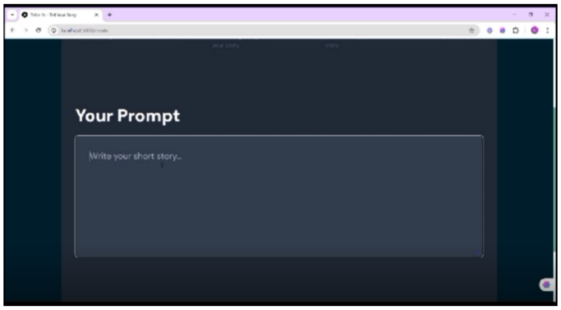
A "Play a Story" button encourages user interaction, guiding them to explore the full story.

User Experience & Design Considerations:

The interface follows a minimalistic design with a dark theme to enhance readability.

The typography and color choices (green highlights on black background) create a visually appealing experience.The structured layout ensures users can quickly scan and navigate through stories.

This home page serves as the foundation for user engagement, ensuring seamless access to AI-generated stories while maintaining an aesthetically pleasing and functional design.



Your Prompt

Write your short story...

The Story Prompt Input Page is designed to allow users to create custom stories by providing an initial text prompt. This step is crucial in defining the storyline, as it guides the AI in generating a coherent and engaging narrative.

Users can enter a custom short story or prompt in the provided text box.

This input serves as the starting point for AI-generated storytelling.

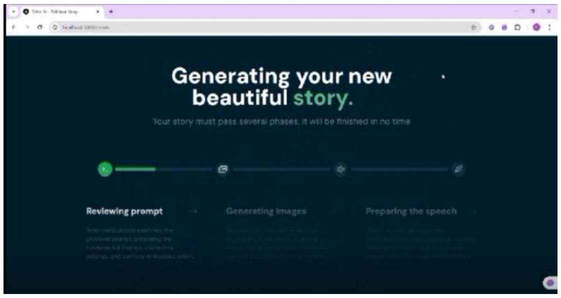
A large text area is provided for ease of writing.

Placeholder text “Write your short story...” encourages user input.

Encourages creativity and personalization in storytelling.

Provides a seamless writing experience with a clean, distraction-free design.

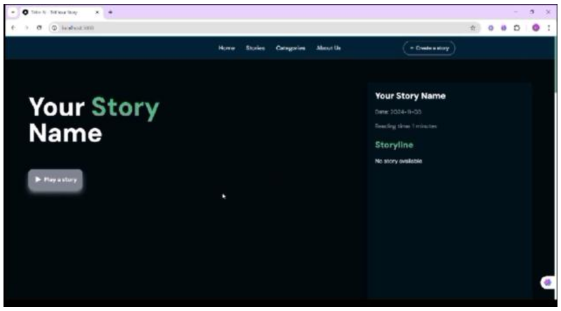
Ensures flexibility, allowing both short prompts and longer user-written stories.



Generating your new beautiful story.

Your story must pass several phases, it will be finished in no time

The Story Generation Progress Page provides users with real-time updates on the AI-driven story generation process. It ensures an interactive and engaging experience by visualizing the different stages involved in creating a multimodal story.



Your Story Name

Your Story Name

Reading time: 1 minutes

Storyline

No story available

The Story Playback Page serves as the final stage of the storytelling process, allowing users to interact with the generated story through text and audio narration.

The story name is displayed prominently.

Date of creation is included Estimated reading time is given. Displays the generated text-based storyline.

VIII. CONCLUSION

Advancements in AI have greatly improved the ability to generate text and images together in a smooth and meaningful way. Models like Story Teller, which use technologies such as Vision Transformers (ViT), Tokenizers, SDXL, and LoRA, help create visuals that match the story while keeping the narrative consistent across multiple scenes. These models solve key challenges like maintaining character consistency, ensuring smooth plot progression, and generating high-quality images that align with the text. Attention mechanisms, such as self-attention and cross-attention, improve the connection between text and images, while fine-tuning methods like LoRA make the process more efficient. New techniques, including diffusion-based models and autoregressive language models, further enhance storytelling by managing character consistency and scene transitions effectively. Additionally, user customization and personalization are becoming more important, leading to the development of interactive and dynamic story generation systems. These improvements mark a major step forward in automated storytelling.

APPENDIX

This appendix contains extra information that aids the research outcomes and technical realization.

Sample Input-Output for Story Generation

Input: Brief prompt that illustrates a medieval knight going on a mystical adventure.

Generated Output: In-depth story coupled with AI-created images, ensuring consistency in plot development andcharacter traits.

Algorithm Overview

Pre-processing the input for context extraction.

Using MLLMs to generate story text.

Implementing multimodal attention mechanisms for image creation.

Establishing coherence through iterative feedback loops.

Dataset Breakdown (StoryStream)

Includes organized story components, character profiles, and related AI-produced images.

Used for fine-tuning and improving multimodal story generation models.

ACKNOWLEDGMENT

We express our sincere gratitude to Jayawantrao Sawant College of Engineering for providing resources and support throughout this research. Special thanks to our mentor, Ms. Smita S. Wagh, for her invaluable guidance and encouragement. We also appreciate the contributions of open-source AI communities, particularly Stability AI, for advancing research in AI-generated storytelling. Additionally, we acknowledge our peers and reviewers who provided constructive feedback to refine our work.

REFERENCES

[1] Jiquan Ngiam1, Aditya Khosla1, Mingyu Kim1, Juhan Nam2, Honglak Lee3, Andrew Y. Ng1.: Multimodal Deep Learning .ResearchGate (2024).

[2] Jiayang Wu1, Wensheng Gan1,2, Zefeng Chen1, Shicheng Wan3, Philip S. Yu4.: Multimodal Large ResearchGate (2023).

[3] Yingzi Huo ,Kai Jin ,Jiahong Cai : Vision Transformer (ViT)-based Applications in Image Classification . ResearchGate (2023).

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge: QWEN TECHNICAL REPORT . ResearchGate (2022).

[5] Yitong Li 1, Zhe Gan2, Yelong Shen4, Jingjing Liu2, Yu Cheng2, Yuexin Wu5 : StoryGAN: A Sequential Conditional GAN for Story Visualization . ResearchGate (2022).

[6] Edward Hu, Yelong Shen, Phillip Wallis ,Zeyuan Allen-Zhu, Yuanzhi Li ,Shean Wang, Lu Wang, Weizhu Chen : LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODEL. arXiv:2106.09685v2 [cs.CL] 16 Oct 2021.

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Mülle, Joe Penna, Robin Rombach: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952v1 [cs.CV] 4 Jul 2023.

[8] Ch.Sita Kameswari , Kavitha J , T. Srinivas Reddy , Balaswamy Chinthaguntla , Senthil Jagatheesaperumal , Silvia Gaftandzhieva , Rositsa
Doneva: An Overview of Vision Transformers for Image Processing: A Survey. ResearchGate(2023).

[9] Chunye Li, Liya Kong, Zhiping Zhou: Improved-StoryGAN for sequential images visualization. ELSEVIER(2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018

[11] Punam V. Maitri, Dattatray S. Waghole, Vivek S.Deshpande “Low latency for file encryption and decryption using Byte Rotation Algorithm”, Proceedings of IEEE International conference on International Conference on Pervasive Computing, 2015.

[12] Vivek S Deshpande, Dattatray S Waghole, “Performance analysis of FMAC in wireless sensor networks”, pp. 1-5 ,IEEE , Eleventh International Conference on Wireless and Optical Communications Networks (WOCN), 2014

[13] Dattatray Waghole, Vivek Deshpande, Divya Midhunchakkaravarthy, Makarand Jadhav, “Position aware congestion control (PACC) algorithm for disaster management system using WSN to improve QoS”, Design Engineering, pp. 11470-11478, 2021

[14] Dattatray S Waghole, Vivek S Deshpande, Punam V Maitri,” pp. 1=5, IEEE, International Conference on Pervasive Computing (ICPC) 2015.

[15] Dattatray S Waghole, Vivek S Deshpande, ” Analyzing the QoS using CSMA and TDMA protocols for wireless sensor networks”, pp. 1-5, IEEE, International Conference for Convergence for Technology-2014.

[16] Omkar Udawant, Nikhil Thombare, Devanand Chauhan, Akash Hadke, Dattatray Waghole, “Smart ambulance system using IoT”,pp 171-176, IEEE, International conference on big data, IoT and data science (BID),2017

[17] Sohail Shaikh, Dattatray Waghole, Prajakta Kumbhar, Vrushali Kotkar, Praffulkumar Awaghade,” Patient monitoring system using IoT”, pp. 177-181, IEEE International conference on big data, IoT and data science (BID) 2017

[18] Dattatray S Waghole, Vivek S Deshpande,” Techniques of data collection with mobile & static sinks in WSN’s: A survey”, Vol. 5, issue.10, 501-505, International Journal of Scientific & Engineering Research, 2010.

[19] Dattatray S Waghole, Vivek S Deshpande,” Reducing delay data dissemination using mobile sink in wireless sensor networks”, Vol.3, issue.1, pp. 305-308, International Journal of Soft Computing and Engineering,2013

[20] Prajakta Patil, Dattatray Waghole, Vivek Deshpande, Mandar Karykarte, “Sectoring method for improving various QoS parameters of wireless sensor networks to improve lifespan of the network”, pp.37-43, vol.10, issue.6, 2022.