AI Story Craft: A Platform for collaborative storytelling, blending human creativity with AI driven text and visuals to craft engaging narratives

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-The rapid advancements in both image generation and open-form text generation have opened new avenues for creating interleaved image-text content. This paper focuses on multimodal story generation, an innovative task that integrates narrative text with rich visual elements in a cohesive manner. While promising, this end story presents substantial challenges, particularly in understanding the intricate relationship between text and images, as well as in generating extended sequences of coherent, contextually relevant narratives and visuals. We introduce Story Teller, a groundbreaking approach that harnesses a Multimodal Large Language Model (MLLM) to create comprehensive multimodal stories. Our model excels in predicting both text and visual tokens, employing an adapted visual de tokenizer to generate images that maintain character consistency and stylistic coherence. We also introduce a novel multimodal attention sink mechanism that facilitates the efficient generation of stories with up to 25 interleaved sequences, surpassing the training limit of 10. To support our model and evaluate the multimodal story generation task, we present StoryStream, a large-scale, high-resolution dataset designed for comprehensive training and quantitative analysis. This work aims to advance the state of multimodal storytelling, offering insights and tools for future research in this dynamic field.

Keywords: MLLM, Vit, SD-XL, Detokenizer, Tokenizer, Diffusion Model, Story-Teller

I. INTRODUCTION

Interleaved image-text data is pervasive across the internet, where multiple images are seamlessly integrated with text. Recently, there has been a growing interest in generating such interleaved content, fueled by significant advances in both image and text generation technologies. This evolution has led to the emergence of Multimodal Story Generation[1][2], a compelling and valuable task that involves creating narratives with interspersed text and vivid imagery. It transcends traditional storytelling by blending visuals and text, crafting an immersive experience where both elements dynamically enhance one another. However, multimodal story generation presents considerable challenges due to the complexity of the data and the high standards required for both text and image output[3][4]. The task demands a deep understanding of interleaved data, where the text is not only descriptive and narrative but also closely linked with intricate visual elements. The model must skillfully capture the relationship between images and text to maintain narrative coherence. Additionally, it requires generating compelling visuals that align with the story, ensuring character and style consistency across both text and images. Recent developments in Multimodal Large Language Models (MLLMs)[5][6] have demonstrated impressive capabilities in comprehending and processing multimodal data, making them well suited for generating interleaved image-text narratives. To leverage these strengths, we introduce Story Teller, a novel approach built on the MLLM[7][8] framework. It enhances the model's ability to generate coherent images in tandem with narrative texts. Utilizing pre-trained image tokenizers and de-tokenizers, Story Teller decodes realistic images with SD-XL[9] by employing the features of a pre-trained ViT[10] (Vision Transformer). During training, the model uses interleaved visual and textual data, incorporating next-word prediction and image feature regression to regularize multimodal generation. This setup allows the model to reconstruct ViT[11] features, which are fine-tuned through de-tokenizer adaptation for consistent image quality and narrative cohesion. To further improve the model's ability to generate long-form stories, we propose a multimodal attention sink mechanism[12]. This mechanism, inspired by window attention, manages a sliding window on key-value states to

handle longer sequences efficiently, surpassing the limitations imposed by training sequence lengths. Our model, equipped with this mechanism, can generate extended multimodal stories featuring rich text plots and varied visual settings. Additionally, we introduce StoryStream, a large scale dataset designed specifically for training and evaluating multimodal story generation. StoryStream[13], derived from animated videos, offers a significant increase in data volume, image resolution, sequence length, and narrative depth compared to existing datasets. We also propose new evaluation metrics that assess image consistency, narrative engagement, and image-text coherence. Results show that Story Teller achieves superior performance across these dimensions. In summary, our contributions are threefold: (1) We present Story Teller, a novel method leveraging MLLM[1,2,6] for generating rich multimodal stories; (2) We introduce a multimodal attention sink mechanism to efficiently handle long-form story generation; (3) We release StoryStream[14][15], a comprehensive dataset designed to benchmark and enhance multimodal story generation tasks.

1.1 Introduction of domain :

Artificial Intelligence (AI) is used to describe machines' abilities to do things that require human intelligence, like learning, problem-solving, decision-making, and natural language comprehension. Machine Learning (ML) is an area of AI that specifically trains machines to identify patterns and make decisions based on data so that systems can learn and improve autonomously over time rather than being programmed for specific tasks. Techniques in ML include supervised, unsupervised, and reinforcement learning and is employed in applications such as recommendation systems, image recognition, and predictive analytics.

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

1. Transformer Attention Mechanism (used in Qwen MLLM and ViT)

The self-attention mechanism helps the model focus on relevant parts of the input sequence. It is calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- QQ, KK, VV are the query, key, and value matrices.
- dkd_k is the dimension of the key.
- The output is a weighted combination of values based on attention scores.

2. Text-to-Speech Mel Spectrogram Generation

For audio narration, the input text is converted into mel spectrogram, which is later turned into audio:

$$S(f,t) = \sum_{n=0}^{N-1} x(n) \cdot w(n-t) \cdot e^{-j2\pi fn/N}$$

Where:

- x(n)x(n) is the time-domain signal
- w(n−t)w(n - t) is a window function
- S(f,t)S(f, t) is the spectrogram at frequency ff and time tt

IV. UNITS

Time: Seconds (s)
Frequency (Audio): Hertz (Hz), Kilohertz (kHz)
Image Dimensions: Pixels (px), Megapixels (MP)
Data Size: Megabytes (MB), Gigabytes (GB)
Computation Time: Seconds (s), Milliseconds (ms)
Model Parameters: Millions (M), Billions (B)

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. Multimodal Deep Learning :

In speech recognition, people learn to combine visual and auditory information for understanding. Speech provides an early illustration of the McGurk effect [1], wherein the image for /ga/ is paired with the voiceover producing /ba/. Most participants perceived /da/ as indicative of a certain typology of image. The joints' positions [2] and muscle activities most often help distinguish them. The words with identical phonological features are analyzed in this paper. Multimodal learning refers to using deep architectures to acquire multiple representations

2. Multimodal Large Language Models Survey

Large Language Models (LLMs) have become remarkably proficient in understanding performing tasks like generating text, following instructions, and imitating examples. However they can be Limited as these AI systems are that can only process text– they can't "see" or understand images. On the other hand, Large Vision Models LVMs are extremely proficient in identifying objects within images. they are not very good at it but yet can do reasoning or understand a language. To amalgamate the strengths of A new class of models emerged from this merger, too: Multimodal Large Language Models (MLLMs).

These models incorporate both language and vision capabilities, allowing them to operate on textual as well as visual inputs. simultaneous. These make excellent candidates for tasks requiring both kinds of knowledge, Generating code for a website from a design image,Solving mathematical problems by reading them directly from a picture.

3. Vision Transformer (ViT)-based Applications in Image Classification

Internet technology has grown rapidly and the computer systems have advanced quickly in the past decade As noted in references [3]-[5], artificial intelligence has been applied in many sectors of societal working, which not only improves the intelligence and modernization of many industries but also carries great convenience to people's life [6]–[8]. Image classification is one of the most primitive tasks.

Computer vision, as part of artificial intelligence, has turned into a significant research area. Image. classificatory is a type model, the principal mission of which is to extract features from the original rendering, categorize the image into different categories and perform minimum classification system.

4. Qwen Technical Report

We introduce the Qwen-VL series, a new advanced line of vision-language models.

(VL) models mark yet another significant step forward in improved multimodal capabilities, combining

image processing and linguistic understanding. The Qwen-7B is foundational to all these models. Architecture combines sight and language into one new, though complex, visual construct. a receptor and position-aware adapter, thus overcoming the limitations of earlier LLMs which were The Qwen-VL models can process and interpret both text and images, positioning that step further in the multilingual text-based attitude system.

5. StoryGAN : A Sequential Conditional GAN for Story Visualization

StoryGAN is an inventive tool that works on a principle of converting story into images. based on one paragraph containing several sentences. This is a tough task since it involves for literally understanding the story's words and portraying them visually. Two main challenges arise here. First, the images need to be consistent, that is, co- clearly tell the whole story. Conventional algorithms generate images from short captions, fail with longer stories because they tend to focus on only one image. in the same way, it is possible to miss the lessons and overlook practical points in what may on first impression look like a simple or not too complex story.

6. Lora: Low rank Adaptation of large language models

This article discusses Low-Rank Adaptation (LoRA), an innovative technique developed to improve the efficiency of adapting large language models, especially in comparison to full fine-tuning, which is becoming increasingly impractical. With the expansion of models like GPT-3, which has 175 billion parameters, retraining all model parameters becomes too costly and requires resources. Additionally: Figure 1. To address this problem, LoRA freezes the pre-trained model weights and introduces trainable low-rank decomposition matrices into each layer of the Transformer architecture. This novel technique reduces the number of trainable parameters by up to 10,000 times compared to full fine-tuning, and it also reduce GPU memory usage by three times. Unlike other adaptation techniques like adapters, LoRA

preserves or even improves model performance across various architectures such as RoBERTa.

7. Tokenization_as_the_initial_phase_in_NLP

Tokenization, the initial stage of Natural Language Processing (NLP), is discussed by the authors in their paper "TOKENIZATION AS THE INITIAL PHASE IN NLP", which highlights its complexity and significance. The understanding of "word" and its "token" is examined using two distinct methods, the one provided by a lexicographer and the other carried out in NLP. The lexicographer's perspective emphasizes the identification of formal items, or "lexical item(s)" that can be defined by their patterns of occurrence in relation to other linguistic items. The understanding of how words are used in language and how they can be categorized into meaningful units is dependent on this viewpoint.

8. SDXL: Improving Latent Diffusion Modfor High-Resolution Image Synthesis

SDXL signifies a considerable advancement in the field of text-to-image synthesis (1). It represents a new iteration of latent diffusion models, which are distinguished by an upgraded architecture: this includes a UNet backbone that is three times larger than those present in earlier versions of Stable Diffusion. The escalation in model parameters is primarily attributable to the addition of further attention blocks and an enlarged cross-attention context; moreover, SDXL incorporates a second text encoder. This structural enhancement facilitates an improved capacity to manage intricate prompts, thereby generating images of superior quality. In addition to its robust design, SDXL unveils several innovative conditioning schemes and is trained across diverse aspect ratios, which increases its versatility in image generation.

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. CONCLUSION

The literature on multimodal story generation underscores the rapid advancement of techniques, models, and methodologies designed to seamlessly integrate text and image synthesis into cohesive narratives. Models like Story Teller, which utilize technologies such as Vision Transformers (ViT), Tokenizers, SDXL, and LoRA, have made significant progress in aligning generated visuals with textual content while maintaining narrative coherence across multiple frames. These models effectively address challenges such as visual consistency, plot progression, and the production of high-quality images paired with rich storytelling. Attention mechanisms, including self- and cross-attention, enhance text-image alignment, while fine-tuning approaches like LoRA contribute to computational efficiency. Emerging techniques, such as diffusion-based models and autoregressive language models, are increasingly adept at managing complex storytelling elements, such as character consistency and scene transitions, ensuring a more seamless and engaging user experience. Furthermore, the importance of user customization and personalization has gained traction, opening the door to the development of interactive, dynamic, and personalized story generation systems. These ongoing improvements in multimodal models mark a significant step forward in automated.

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 and character 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] 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

[12] 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

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

[14] 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.

[15] 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