StoryVizAI: Generating Visual Narratives from Textual Descriptions

Text-to-Image Story Generation: Bridging Words and Visuals

Dyuti Dasmahapatra[1] , Palak Gupta[2], Ankur Kaushal[3], Himanshu Sharma[4], Priyanshu Yadav[5]
Dyuti Dasmahapatra (BML Munjal University, (SOET), Kapriwas, (Haryana), India
Palak Gupta (BML Munjal University, (SOET), Kapriwas, (Haryana), India
Ankur Kaushal (BML Munjal University, (SOET), Kapriwas, (Haryana), India
Himanshu Sharma (BML Munjal University, (SOET), Kapriwas, (Haryana), India
Priyanshu Yadav (BML Munjal University, (SOET), Kapriwas, (Haryana), India

***Abstract*— Background :** In recent years, significant progress has been made in the field of text-to-image synthesis, facilitated by the development of large pretrained transformers. However, such models often struggle with specialized tasks like story visualization, where a coherent sequence of images is required to accompany a narrative. Furthermore, these models face challenges in generalizing unseen plots and characters.To address these limitations, our research introduces the task of story continuation, wherein a sequence of images is generated based on a given source image, allowing for better adaptation to new narratives. Recognizing the scarcity of large-scale datasets for training such models from scratch, especially due to the requirement for narrative continuity among images, we propose leveraging the pretrained knowledge of text-to-image synthesis models. Specifically, we enhance pretrained text-to-image synthesis models with task-specific modules for sequential image generation and copying relevant elements from an initial frame. We explore both full-model finetuning and prompt-based tuning to adapt the pretrained model to our task efficiently.

Our research explores the intersection of generative AI and narrative-driven projects, focusing on the task of story visualization. We detail a comprehensive methodology aimed at generating visual narratives from textual prompts, emphasizing the use of state-of-the-art techniques and datasets. Leveraging the HugChat API for prompt generation, we guide the creation of textual descriptions that serve as blueprints for subsequent image generation. Utilizing the stable-diffusion-xl-base-1.0 text-to-image model, we translate these prompts into vivid visual scenes, forming the foundation of our narrative generation process. To enhance the model's adaptability and narrative coherence, we employ fine-tuning techniques using the Flicker30 dataset, enriching its ability to generalize across diverse narratives and scenes. Through empirical evaluation, we demonstrate the effectiveness of our approach in generating cohesive and visually compelling narratives. Our research not only advances the capabilities of generative AI in narrative visualization but also underscores the importance of fine-tuning and dataset selection in achieving robust performance in low-resource scenarios.

1. Introduction

Pretrained text-to-image synthesis models, exemplified by stable-diffusion-xl-base-1.0, have exhibited remarkable proficiency in translating input captions into coherent visualizations. These models, alongside powerful multimodal approaches, signify a democratization of art by rendering artistic renditions of textual descriptions. However, their utility has been primarily confined to processing single, concise captions. Contrastingly, real-world applications often demand the processing of longer narratives and metaphorical expressions, necessitating models capable of conditioning on existing visuals and generating multiple images to encapsulate textual meaning comprehensively. While prior works have explored specialized Generative Adversarial Networks (GAN) for tasks like image-to-image translation and style transfer, the emergence of transformer-based pretrained models presents an opportunity to harness latent knowledge from extensive pretrained datasets, akin to fine-tuning for downstream tasks in language understanding. Motivated by these advancements, our paper delves into the adaptation of pretrained text-to-image synthesis models for intricate downstream tasks, particularly focusing on the realm of story visualization.

Story visualization, a convergence of image generation and narrative comprehension, presents distinct challenges. Here, agents are tasked with generating a series of images aligned with a sequence of narrative captions, effectively depicting the story's progression. While previous research has explored the potential of story visualization, practical implementation faces obstacles. Models trained on fixed sets of characters, settings, and events lack the ability to represent newly introduced characters during testing. To overcome this constraint and enable adaptation to novel story elements, we introduce the concept of "story continuation." This task involves providing an initial scene, allowing for the replication and adjustment of elements as subsequent images unfold. Our approach shifts the focus beyond mere text-to-image conversion, prioritizing narrative structure and the dynamic evolution of images to capture nuanced storytelling details.

In our implementation, we leverage the HugChat API for prompt generation, facilitating user interaction with the chatbot to initiate narrative prompts. Additionally, we incorporate a library for scene description and character portrayal, enabling detailed contextualization of story elements. Furthermore, we utilize a state-of-the-art text-to-image pipeline, specifically the StableDiffusionXLPipeline, to translate textual prompts into visual representations. To enhance model capabilities and ensure adaptability to diverse narratives, we plan to fine-tune the pretrained model using the Pokemon-Blip-Captions dataset. This dataset, containing text-image pairs, offers a rich source of diverse story elements for model training. Moreover, we have integrated Google Translator to enable prompt generation and processing in multiple languages, including Hindi, thereby enhancing the inclusivity and accessibility of our approach. Through these technical implementations, we aim to demonstrate the efficacy of adapting pretrained models for complex narrative-driven tasks, paving the way for advancements in story visualization and generative AI applications.

2. Literature Review

The literature surrounding automated story generation and multimodal content creation has witnessed remarkable progress in recent years, driven by advancements in generative models and natural language processing techniques. Researchers have explored various avenues to address the challenges inherent in creating coherent narratives and enhancing storytelling experiences, particularly in the realm of multimodal storytelling where text and images converge. This review examines several key contributions in this evolving landscape, each offering unique insights and methodologies to push the boundaries of story generation, visualization, and comprehension. Through an exploration of recent research endeavors, we aim to delineate the state-of-the-art approaches, identify emerging trends, and underscore the significance of multimodal content in fostering engaging and immersive storytelling experiences.

In "**FAIRYTAILOR: A MULTIMODAL GENERATIVE FRAMEWORK FOR STORYTELLING**" the authors delve into the complexities of previous story generation models, highlighting issues such as repetition, incoherence, and limited control over themes. Additionally, they identify shortcomings in existing story visualization systems, which often struggle with specific text inputs or lengthy image generation processes. To overcome these obstacles, FairyTailor introduces a novel approach that combines human input with multimodal generation, leveraging both text and image data to enhance storytelling experiences, particularly for young readers.

Building upon these insights, "**StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation**" tackles the limitations of existing story visualization tasks and models, which often falter in maintaining consistency across narratives and fail to generalize to unseen characters and settings. By introducing the concept of story continuation, the paper reframes story visualization by providing an initial scene as input, thereby enabling better generalization and continuity throughout the narrative. Leveraging pretrained text-to-image transformers and task-specific modifications, such as a global story encoder and retro-fitted cross-attention layers, StoryDALL-E enhances its performance in generating coherent and visually appealing stories, thus advancing the state-of-the-art in multimodal storytelling.

In contrast, "**Zero-shot Generation of Coherent Storybook from Plain Text Story using Diffusion Models**" focuses on text-guided image generation and multimodal storytelling, acknowledging previous research efforts in this domain. The proposed method distinguishes itself by offering a zero-shot generation capability, enabling the creation of coherent storybooks directly from plain text stories without the need for pre-trained models or large annotated datasets. Through comparative evaluations and user studies, the authors demonstrate the superiority of their approach in terms of coherency preservation, background preservation, and overall user satisfaction, contributing significantly to the field of multimodal storytelling.

Similarly, "**StorVi (Story Visualization)**: A Text-to-Image Conversion" positions itself within the broader context of enhancing children's storytelling experiences through visual aids. By addressing the specific needs of children aged 4-7 and emphasizing the importance of visualization in their development, the researchers provide a valuable tool for educators and parents to facilitate storytelling and enhance children's engagement with narrative content. Through user surveys with teachers and parents, StorVi demonstrates usability, user-friendliness, and positive feedback on the content of the generated frames, highlighting its potential as an effective educational resource.

Finally, in "**Make-a-Story: Visual Memory Conditioned Consistent Story Generation**," recent advancements in generative models are leveraged to produce high-quality images based on textual descriptions, with a focus on maintaining narrative coherence and consistency across characters and backgrounds. By introducing an autoregressive diffusion-based framework with a visual memory module, the proposed method resolves ambiguous references and establishes appropriate correspondences, resulting in visually coherent and semantically consistent story frames. Through comprehensive evaluations, the authors demonstrate superior performance compared to previous state-of-the-art techniques, marking a significant advancement in the field of story generation and multimodal content creation.

3. Dataset

Our dataset comprises a subset of the pokemon blip captions sourced from Hugging Face, containing a total of around 700 images. Each image in this dataset is paired with a corresponding caption, providing valuable textual context for the visual content. The captions vary in length and complexity, encompassing a wide range of descriptive styles and narrative elements. To streamline the fine-tuning process and address computational constraints, we meticulously select around 700 images from the dataset for model fine-tuning. This subset is carefully curated to ensure diversity in visual content, encompassing various scenes, objects, and settings depicted in the images. By working with a subset of images, we optimize

computational resources while still benefiting from the rich textual and visual diversity present in the dataset. This approach ensures that our model is trained on a representative sample of visual data, enhancing its ability to generate coherent and contextually relevant images in response to textual prompts.

Fig 1. Images inside Pokemon Blip Caption Dataset

4. System Architecture

4.1 System Framework

The system architecture is designed to facilitate the generation of scene images based on user input, employing the Hugging Face Chat API and a pre-trained model, specifically the stable-diffusion-xl-base-1.0 model. At the forefront of the architecture lies the User Interface, serving as the entry point for users to input their prompts describing the desired scene. This interface could take various forms, such as a text box or any other interactive element enabling users to articulate their scene descriptions effectively.

Once the user input is provided, it is transmitted to the Hugging Face Chat API, which serves as an intermediary processing layer. Here, the input prompt undergoes analysis and transformation, leveraging the capabilities of the Hugging Face Chat API to generate a coherent scene description. This description encapsulates the essence of the user's input, translating it into a textual representation of the envisioned scene.

In tandem with the Hugging Face Chat API, an additional option for text translation is available through Google Translate, enhancing accessibility and accommodating diverse linguistic preferences among users.

The generated scene description then proceeds to the Pre-trained Model component, where the stable-diffusion-xl-base-1.0 model comes into play. This pre-trained model is specialized in converting textual descriptions into corresponding image representations. Employing encoder to decoder architecture, the model deciphers the textual scene description and generates a visual depiction of the described scenario. Leveraging the encoder to decoder mode ensures a seamless transition from textual input to visual output, capturing the intricacies and nuances of the scene description effectively.

Following the image generation process, the resulting image is produced as the final output of the system. This image, representing the visual manifestation of the user's input prompt, is then displayed to the user via the Output component. The user can interact with the displayed image, assessing its fidelity to the input prompt and potentially providing feedback or further instructions for refinement.

Overall, this system architecture enables the seamless transformation of user-provided prompts into corresponding scene images, leveraging the combined capabilities of natural language processing and image generation models. By orchestrating the flow of information from user input to visual output, the architecture facilitates intuitive and interactive scene generation, catering to diverse user preferences and creative requirements.

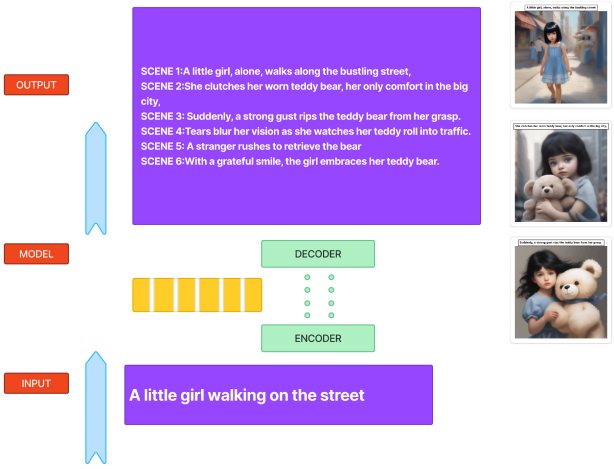

Fig 2 System Framework

4.2 Architecture Overview: Scene Image Generation

This system generates scene images based on user input, utilizing the Hugging Face Chat API and the pre-trained model stable-diffusion-xl-base-1.0.

The user begins by entering a prompt describing their desired scene. This prompt could be a simple description of the scene, or it could include more specific details about the desired elements, such as the time of day, weather conditions, or desired mood. The user interface then sends the prompt to the Hugging Face Chat API.

The Hugging Face Chat API receives the user's prompt and uses it to generate a scene description. This scene description is a more detailed and nuanced interpretation of the user's prompt, taking into account various factors such as the context of the scene, the desired mood, and the potential for symbolic or metaphorical interpretations.

The generated scene description is then sent to the pre-trained model stable-diffusion-xl-base-1.0. This model is a powerful text-to-image generator that has been trained on a massive dataset of images and their corresponding text descriptions. The model uses this knowledge to generate an image that accurately reflects the scene described in the text.

Finally, the generated image is displayed to the user. The user can then review the image and make any necessary adjustments to the prompt or scene description. This iterative process allows the user to fine-tune the image generation process until they are satisfied with the results.

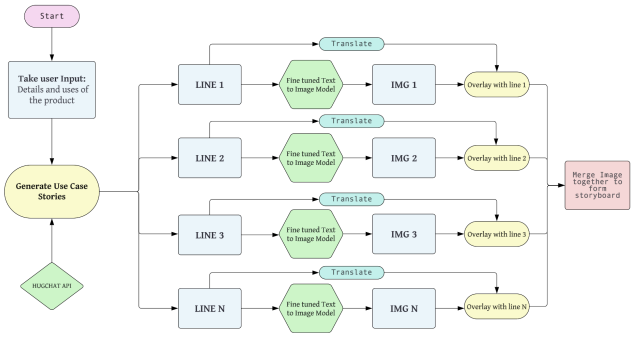

Fig 3. Architecture Overview

This system provides a powerful and flexible tool for generating scene images based on user input. By leveraging the capabilities of the Hugging Face Chat API and the pre-trained model stable-diffusion-xl-base-1.0, the system can generate realistic and detailed images that match the user's desired scene. This technology has the potential to be used in a wide variety of applications, such as creating storyboards for movies or video games, generating concept art for products or environments, and even creating personalized artwork for individuals.

4.3 Model Architecture

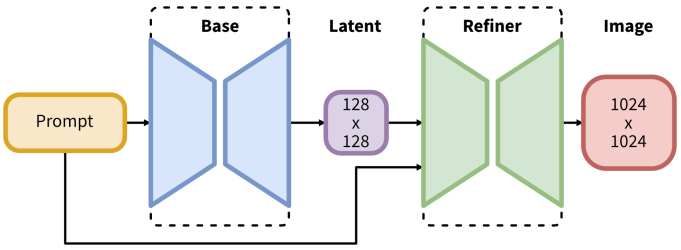

Fig 4 ArchitectureSDXL model

A strong framework for latent diffusion, the stable-diffusion-xl-base-1.0 (SDXL) model [8] has strong capabilities for producing high-quality images from textual prompts. Fundamentally, SDXL utilizes a pipeline of experts, utilizing a two-phase method for latent diffusion. Based on the given prompt, the basic model creates initial (noisy) latents in the first stage. After that, a specific refinement model designed for the last stages of denoising is applied to these latents. It's crucial to remember that the base model offers flexibility in its use by functioning as a standalone module as well. As an alternative, SDXL provides a two-stage pipeline in which the desired output size is first generated as latents by the basic model. Using the same cue, a specialized high-resolution model is used in the second stage to refine the latents created in the first step using a technique called SDEdit. Because it requires more function evaluations, this technique might be a little slower, but it produces images with better resolution and fidelity, which makes it a useful choice in some cases.With the help of this model, images can be created and altered in response to text commands. Two fixed, pretrained text encoders (OpenCLIP-ViT/G and CLIP-ViT/L) are used in this Latent Diffusion Model.

5. Experiments

In our experiment, we aim to generate a storyboard based on product descriptions and usage through a two-step process: generating story descriptions using natural language processing (NLP) and generating corresponding images using a text-to-image model.

Generating Story Descriptions

For the first step, we employ the Hugging Face Chat API to generate story descriptions based on product descriptions and usage. The API utilizes a pretrained model to generate coherent and contextually relevant text based on user prompts. We input the product description and usage details into the API, which then generates a descriptive narrative reflecting the product's features and scenarios of use.

Generating Scene Images

In the second step, we utilize a text-to-image model to generate images corresponding to the story descriptions generated in Step 1. Specifically, we employ the StableDiffusionXLPipeline model, pretrained on the "stable-diffusion-xl-base-1.0" architecture. This model is capable of converting text inputs into visually coherent images. We provide the generated story descriptions as prompts to the model, which then produces corresponding scene images.

Fine-Tuning with Pokemon-Blip-Captions Dataset

After the initial experimentation phase, our next step involves fine-tuning the model using the Pokemon-Blip-Captions dataset. This dataset, sourced from Hugging Face, comprises a diverse collection of images paired with corresponding captions. By fine-tuning our model on this dataset, we aim to refine its ability to generate high-fidelity images aligned with the provided textual prompts. This process enhances the model's proficiency in understanding and translating narrative descriptions into visually compelling scenes, contributing to overall performance improvement. Fine-tuning on this dataset aims to improve the model's ability to generate high-quality images that accurately reflect the given story descriptions.The fine tuning process incorporates Exponential Moving Average (EMA) for stability and model generalization. Images are standardized to 512x512 pixels with center cropping for consistency. Random horizontal flipping augments the dataset for improved model robustness. Training efficiently processes individual samples with a batch size of 1 and accumulates gradients over 4 steps to effectively increase batch size without exceeding memory constraints. Gradient checkpointing reduces memory consumption, enabling training with larger models. Mixed precision training accelerates computation with 16-bit floating-point format. Training is limited to 400 steps to control duration and prevent overfitting. Learning

rate is gradually adjusted for model convergence, while gradients are clipped to prevent instability. Checkpoints are shared to the Hugging Face Model Hub for collaboration and reproducibility, and saved at intervals for tracking progress. Output directory is defined for storing trained model checkpoints and logs.

Multilingual Support with Google Translator

Additionally, we have implemented support for multilingual prompts using Google Translator. This feature allows users to input prompts in languages other than English, such as Hindi,Spanish, French, etc. By leveraging Google Translator, we aim to enhance the accessibility and usability of our system for users from diverse linguistic backgrounds.

Hardware Setup

All experiments are conducted on hardware equipped with CUDA-enabled GPUs to accelerate model training and inference processes, ensuring efficient utilization of computational resources.

6. Results

Baseline Model:

While the base model demonstrated promising performance in generating image representations based on textual prompts from scenes generated using Hugging Face Chat, it lacked fine-tuning on dataset-specific features. As a result, the generated images exhibited some inconsistencies in capturing nuanced details and visual fidelity, particularly in achieving a cartoonish effect conducive to creating visually appealing storyboards.

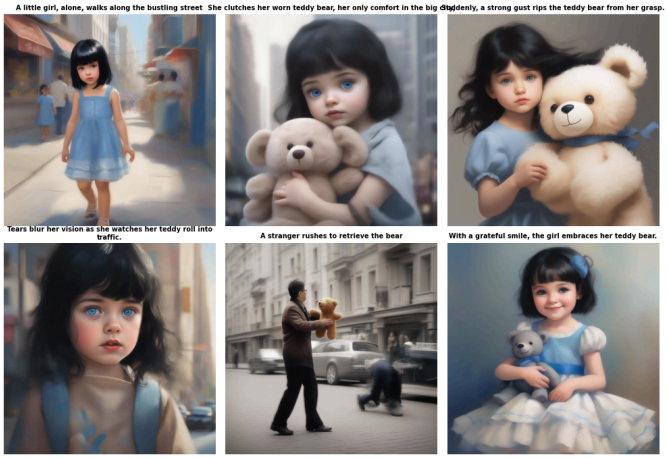

Fig 5. Storyboard generated using baseline model

Fine-Tuned Model:

After training with the dataset, we observed significant improvements in the model's ability to generate images with the desired cartoonish effect, crucial for virtual storyboards. The fine-tuned model demonstrated enhanced proficiency in understanding and translating textual descriptions into visually compelling scenes aligned with Pokémon themes. The generated images exhibited greater consistency with dataset-specific features and showed improved fidelity to the provided captions. Overall, fine-tuning on the Pokémon-Blip dataset resulted in a notable enhancement in the model's performance, emphasizing the effectiveness of dataset-specific fine-tuning for generating contextually relevant and visually appealing images for virtual storyboards.

Fig 6. Storyboard generated using fine-tuned model

After integrating Google Translate into our system, we successfully expanded our capabilities to translate text into various languages, including Hindi, Spanish, and more.

Fig 7. Storyboard generated after translation to French

Impact of Dataset Size on Storyboard Quality:

It's worth noting that due to the smaller size of the Pokémon-Blip dataset, consisting of approximately 700 images, there was a slight difference observed in the quality of the generated storyboards compared to larger datasets. While the fine-tuned model showed remarkable improvement in image generation, the limited diversity in the dataset may have impacted the variety and richness of the generated storyboards. Despite this limitation, the fine-tuned model still managed to produce coherent and visually appealing narratives, underscoring the importance of dataset size in model training and its subsequent impact on storyboard quality.

7.Evaluation:

In evaluating the performance of our model, we employed two key metrics: the CLIP score and the step loss. The CLIP score serves as a measure of the model's alignment with the textual prompts provided as input. A higher CLIP score signifies a stronger correlation between the generated outputs and the given prompts. For our experiments, the model attained a

commendable CLIP score of 35.7038, indicating a substantial coherence between the generated content and the input text. Conversely, the step loss metric reflects the average loss computed during each step of the training process. A lower step loss value denotes a more effective convergence and improved performance of the model. In our analyses, the model demonstrated a low step loss of 0.072, suggesting robust training and convergence capabilities. These metrics collectively attest to the efficacy and reliability of our model in generating outputs that closely adhere to the provided textual prompts, while also showcasing its proficiency in minimizing training loss and achieving convergence.

8. Future Scope

Incorporating datasets with larger and more diverse images can further enrich the model's understanding of visual concepts and improve its ability to generate high-quality images. Such datasets contain vast collections of images spanning various categories and contexts, offering ample opportunities for training more robust models.

9. Conclusion

In this study, we embarked on a journey to explore the potential of generative AI models in generating visual narratives from textual descriptions. Leveraging the stable-diffusion-xl-base-1.0 model obtained from Hugging Face as our base framework, we ventured into the realm of story visualization using the Pokémon-Blip dataset.We established a robust architecture capable of transforming user-provided prompts into visually compelling scene images. Through meticulous fine-tuning on dataset-specific features, such as the Pokémon-Blip dataset, we witnessed substantial improvements in image quality and coherence, achieving desired cartoonish effects conducive to virtual storyboarding. Moreover, the incorporation of Google Translate broadened our system's reach, enabling seamless translation of text prompts into multiple languages, further enhancing accessibility and user engagement. Looking forward, the future holds promising opportunities for advancing our system's capabilities, from leveraging larger datasets to exploring domain-specific training and enhancing computational efficiency. By continually innovating and embracing emerging technologies, we are poised to unlock new frontiers in scene image generation, catering to diverse creative needs and preferences while fostering immersive storytelling experiences.

References

[1] Nanda, Shreya D., et al. "Story Visualization: Generation of Scenes sequentially for a given fable using NLP and Image Processing." 2021 Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2021.

[2] Alcantara, Kim D., et al. "Storvi (story visualization): A text-to-image conversion." International Journal of Future Computer and Communication 3.5 (2014): 363.

[3] Maharana, Adyasha, Darryl Hannan, and Mohit Bansal. "Storydall-e: Adapting pretrained text-to-image transformers for story continuation." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[4] Jeong, Hyeonho, Gihyun Kwon, and Jong Chul Ye. "Zero-shot generation of coherent storybook from plain text story using diffusion models." arXiv preprint arXiv:2302.03900 (2023)

[5] Lovenia, Holy, et al. "Every picture tells a story: Image-grounded controllable stylistic story generation." arXiv preprint arXiv:2209.01638 (2022)

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Rahman, Tanzila, et al. "Make-a-story: Visual memory conditioned consistent story generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[8] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0/

Gen AI FINAL REPORT.docx.pdf

ORIGINALITY REPORT

7%	5%	4%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.arxiv-vanity.com Internet Source	2%
2	huggingface.co Internet Source	1%
3	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 Publication	1%
4	ora.ox.ac.uk Internet Source	1%
5	Submitted to Colorado State University Fort Collins Student Paper	<1%
6	arxiv.org Internet Source	<1%
7	Juntae Kim, Yoonseok Heo, Hogeon Yu, Jongho Nang. "A Multi-Modal Story Generation Framework with AI-Driven Storyline Guidance", Electronics, 2023 Publication	<1%

www.researchgate.net

8	Internet Source	<1 %
9	papers.nips.cc Internet Source	<1 %
10	www.ecva.net Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 6 words
Exclude bibliography	On		