Bridging Cultures through AI: The Art of Multilingual Storytelling

Shazia Mannan, Asha Hegde and Sharal Coelho

Department of Computer Science, Mangalore University, India

Abstract

The use of Artificial Intelligence (AI) in visual storytelling has created a new way for cultural exchange, but there are still issues with producing consistent and culturally relevant graphics for multilingual stories. This paper introduces an AI-powered pipeline for multilingual story illustration, focusing on English and Hindi texts influenced by culturally rich sources such as the Panchatantra. Addressing limitations like fixed character sets that limit diversity and infinite characters that compromise consistency, we use publicly accessible datasets with appropriate attribution to produce coherent visuals. Our proposed methodology combines T5-large for abstractive summarization of story segments and Stable Diffusion XL (SDXL) for generating cartoon-style illustrations with a fixed stylistic prompt to ensure narrative relevance and visual uniformity. Results demonstrate high narrative fidelity and cultural appropriateness, evaluated through expert criteria on relevance, quality, and consistency. However, issues like summarization, generality, and model biases highlight areas for improvement. By bridging linguistic gaps, our study promotes inclusive storytelling and paves the way for AI systems that can adapt to diverse cultures within global storylines.

Keywords

Stable Diffusion XL, Text-to-Image, Visual Storytelling, Artificial Intelligence

1. Introduction

The Artificial Intelligence (AI) evolved toward image recognition and generation, expanding capabilities to visual data. The Large Language Models (LLMs) [1] and Text-to-Image (T2I) [2] [3] systems have demonstrated remarkable abilities in creating contextually rich textual and visual narratives. T2I generation is basically the process of generating a visually realistic image that matches a given text description [4]. AI systems that can produce outstanding, contextually aware, and stylistically varied images straight from verbal descriptions and automate visual storytelling, which previously depended almost entirely on the talent and creativity of human illustrators [5]. Within this broader adaptation, multilingual story illustration has become an important and socially significant domain. This paradigm change presents important issues regarding cultural authenticity, linguistic diversity, visual consistency, and many more, but it also has great potential to democratize creativity and speed up the creation of illustrated information [6].

Stories in culturally varied places like India cover a remarkable variety of languages, customs, and visual elements. Oral folktales, modern children's books, and classical compilations like the *Panchatantra* and *Hitopadesha* sometimes use culturally rooted symbols, idioms, and artistic norms in addition to words to express meaning. However, existing AI illustration pipelines are largely trained on datasets with a Western cultural bias, which results in illustrations that poorly reflect local aesthetics, environmental settings, and symbolic representations. T2I generation has achieved exceptional quality and has been extensively used for a variety of tasks due to the rapid development of diffusion models [7]. From a technical perspective, current T2I systems still struggle with fundamental bottlenecks in multilingual storytelling, such as narrative comprehension across languages, character and visual consistency, and cultural adaptability and inclusivity.

To address these gaps, we worked on task MUSIA: Multilingual Story Illustration: Bridging Cultures through AI Artistry. An AI-powered pipeline designed for end-to-end multilingual storytelling and illustration is proposed in the work. Our system introduces three key innovations:

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

smannan81@gmail.com (S. Mannan); hegdekasha@gmail.com (A. Hegde); sharalmucs@gmail.com (S. Coelho)

1. **Narrative Understanding Module** – A multilingual story-processing engine capable of segmenting narratives written in English and Hindi into semantically coherent plot points and generating concise, structured summaries optimized for illustration.
2. **Character-Consistency Mechanism** – A module that maintains visual identity across multiple illustrations, ensuring coherence in character attributes, environment, and stylistic representation throughout the narrative.
3. **Cultural Evaluation Framework** – An integrated evaluation methodology that combines automated quality metrics with human expert decisions to consider cultural relevance, narrative accuracy, and visual coherence. This helps in bridging the gap between technical performance and real-world applicability.

The goal of the MUSIA shared task is to enable culturally aligned, linguistically inclusive, and visually coherent storytelling at scale. As a result, this work has significant applications in the fields of education such as illustrated textbooks and children's books, entertainment like digital comics, graphic novels, and animated shorts. Beyond its technical contributions, this research highlights how AI can serve as a medium for cultural preservation and exchange, bridging divides across languages, geographies, and artistic traditions, and making literature more accessible to global audiences.

2. Related Work

Visual storytelling involves generating a series of images based on narrative text prompts to create a continuous and coherent story. Sudharsan et al. introduced the *KAHANI* pipeline, which focuses on non-Western cultural narratives [8]. Their framework uses GPT-4 Turbo and Stable Diffusion XL. Maharana and Bansal [9] proposed an approach to enhance story-to-image generation. By employing parse trees, dense captioning, and dual learning frameworks, their method improves narrative segmentation, spatial coherence, and visual consistency. This work provides valuable technical insights into the challenges of maintaining coherence across sequential illustrations.

From an educational and user-centered perspective, Han and Cai [10] examined the role of generative AI in children's visual storytelling. Their research incorporated feedback from parents, teachers, and researchers and ultimately proposed a prototype system called AIStory for literacy development. Their findings highlight the importance of usability and child-friendly design in visual storytelling applications. Antony et al. [11] proposed *ID.8*, a co-creative system for collaborative visual storytelling, which emphasizes creativity, author control, and user experience in generating illustrative narratives. Similarly, Kim et al. [12] explored multimodal story-to-image frameworks with a particular focus on character preservation and narrative consistency, addressing one of the key technical bottlenecks in long-form illustrated storytelling.

In summary, previous works have advanced the field through innovations in cultural grounding, linguistic integration, and user-centered design. However, most existing systems remain monolingual or culturally limited, with insufficient attention to multilingual narratives, script diversity, and sustained character coherence across culturally rich storylines. Building upon these foundations, our work contributes a multilingual, culturally adaptive pipeline that integrates narrative segmentation, character-consistency mechanisms, and human-centered evaluation to bridge cultural divides and enable inclusive AI-driven storytelling.

3. Task

This MUSIA shared task highlights the generation of visual illustrations for multilingual stories, particularly in English and Hindi languages [13, 14]. A key challenge in this domain lies in character representation: approaches that rely on a limited, predefined set of characters often fail to capture the richness and diversity inherent in culturally rooted narratives such as the Panchatantra, whereas allowing for an unrestricted number of characters can compromise visual consistency, weakening the reader's connection with the story.

4. Methodology

Our methodology adopts a hybrid AI pipeline that integrates Natural Language Processing (NLP) for narrative understanding with diffusion-based generative models for illustration. The pipeline is designed to process stories in both English and Hindi, thereby ensuring multilingual compatibility. The complete pipeline is organized into three main stages: dataset preparation, narrative summarization, and illustration generation.

4.1. Dataset Preparation

The initial stage focuses on preparing the textual corpus for downstream processing. Each story is provided as a full text file and preprocessed to remove extraneous whitespace. The cleaned text is then segmented into sentences using period-based splitting to maintain logical narrative units. To ensure balanced image generation across a given story, the text is partitioned into sub-stories. Specifically, the sentences are divided into equal-sized chunks, based on the number of images specified for the story. Each chunk, therefore, represents a key narrative segment corresponding to one illustration frame.

4.2. Narrative Summarization

Once the text is segmented, each sub-story undergoes abstractive summarization to filter the narrative into concise prompts suitable for image generation. We employed the `T5-large`[1] model from the Hugging Face Transformers library, a state-of-the-art model for abstractive summarization tasks.

For each sub-story, the model is configured with the following parameters: `max_length=50`, `min_length=35`, and `do_sample=False`. These constraints ensure summaries that are both concise and semantically accurate, while remaining deterministic across runs. The summarization step captures the essence of the plot, filtering out redundant information and highlighting the critical narrative events. This refined representation serves as the semantic backbone for the subsequent illustration prompts, enabling the diffusion model to focus on the most relevant story elements.

4.3. Illustration Generation

The last stage of the pipeline involves generating visual illustrations using the Stable Diffusion XL (SDXL) base model[2] from Stability AI. To optimize performance, the model is loaded with 16-bit floating-point precision and `SafeTensors`, thereby reducing memory usage and improving inference efficiency. The pipeline execution environment is GPU-enabled (Kaggle platform), with CUDA acceleration prioritized when available, to support the computational demands of large-scale diffusion models.

To maintain stylistic uniformity across all illustrations, a fixed prompt prefix is appended to each summary. The prefix is concatenated with the corresponding narrative summary to form the final input prompt for SDXL. Each prompt is then passed to the diffusion pipeline (`pipe(prompt).images[0]`) to generate a single image, which is subsequently stored in PNG format within language-specific directories (e.g., `generated_images_hindi` for Hindi stories).

Through this multi-stage design, our methodology ensures that story texts are transformed into visually coherent and culturally adaptive illustrations, while preserving narrative flow and character consistency across different languages and story genres.

5. Experiments and Results

The dataset used for the multilingual story illustration task was divided into training and testing sets, covering narratives in both English and Hindi. It was carefully curated from culturally rich sources, particularly Panchatantra-inspired stories, and each entry consisted of textual narratives paired with

[1] https://huggingface.co/google-t5/t5-large

[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

illustrations in the case of training data, or mapping files for the test data that indicated how many images were required to be generated per story. The training set comprised 360 English stories and 185 Hindi stories, organized into a "Stories" folder containing the texts and an "Images" folder containing the corresponding illustrations. The testing set included 40 English stories and 30 Hindi stories, with a similar folder structure for narratives and an accompanying mapping file to guide the image generation process.

Since the field of multilingual story illustration currently lacks standardized automatic evaluation metrics, the evaluation of results was performed entirely through human judgment. The evaluation was carried out across three key dimensions: relevance, which measured how effectively the generated illustrations captured the important moments of the story; visual quality, which reflected the artistic and innovative aspects of the illustrations; and consistency, which examined whether the illustrations maintained a coherent visual narrative when considered together.

To ensure reliability in the evaluation, Cohen's kappa was calculated to measure inter-annotator agreement. Only those stories for which the agreement value exceeded 0.75, indicating a strong level of consistency among the annotators, were considered in the analysis. As this threshold was achieved for almost all the stories, the dataset used for evaluation remained comprehensive. The scores were further categorized into three buckets: values between 4 and 5 were classified as good, a score of 3 was treated as moderate, and scores between 1 and 2 were considered fair. For each metric, the number of stories falling into these categories was counted, and performance was judged based on the distribution of stories across the buckets. In this way, a higher number of stories evaluated as good indicated stronger performance, followed by those rated as moderate and then fair.

6. Discussion

The proposed work demonstrated its effectiveness in generating illustrations for multilingual stories, achieving strong alignment between narrative content and visual output. In the case of Hindi narratives, the segmentation and summarization stages ensured that each frame captured critical plot moments, such as character interactions and moral resolutions, while the fixed stylistic prompt preserved consistency in character design and cultural aesthetics. The use of warm color palettes and hand-drawn textures contributed to evoking traditional Indian artistic sensibilities. Similarly, the English narratives yielded coherent results, underscoring the pipeline's language-agnostic adaptability.

Despite the promising outcomes, certain limitations remain. The T5 summarizer sometimes generated outputs that were overly generic for complex narratives. Similarly, while SDXL proved robust in producing visually compelling illustrations, issues of character consistency occasionally surfaced, especially when depicting culturally specific elements such as traditional attire or symbolic artifacts. From a practical standpoint, the high computational demands of diffusion-based models restrict accessibility in resource-constrained environments, limiting broader adoption.

7. Conclusion

This research proposed an AI-driven framework for multilingual story illustration, combining narrative summarization with generative visual models to enhance the accessibility and cultural inclusiveness of storytelling. By focusing on English and Hindi narratives, the study demonstrated how technology can bridge linguistic and cultural gaps, making stories more engaging and visually coherent across diverse audiences. Human evaluation based on relevance, quality, and consistency further validated the effectiveness of the approach.

Despite these promising outcomes, several challenges remain. Character consistency, narrative summarization precision, and mitigation of cultural or linguistic biases require further refinement.

Declaration on Generative AI

In preparing this work, the author(s) utilized Grok[3] for grammar and spelling checks. Paraphrasing was handled via QuillBot. With this tool, the author(s) reviewed and revised the content as required, while assuming full responsibility for the publication's integrity.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Advances in neural information processing systems 35 (2022) 36479–36494.

[4] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1505–1514.

[5] G. Kim, P. Lui, Story illustration using generative adversarial networks (gans) (2021).

[6] Z. Chen, L. Chen, Z. Zhao, Y. Wang, Ai illustrator: Art illustration generation based on generative adversarial network, in: 2020 IEEE 5th International conference on image, vision and computing (ICIVC), IEEE, 2020, pp. 155–159.

[7] S. Dong, I. Shaheen, M. Shen, R. Mallick, S. A. Bargal, Vista: Visual storytelling using multi-modal adapters for text-to-image diffusion models, arXiv preprint arXiv:2506.12198 (2025).

[8] D. Sudharsan, A. Seth, R. Budhiraja, D. Khullar, V. Jain, K. Bali, A. Vashistha, S. Segal, et al., Kahani: Culturally-nuanced visual storytelling pipeline for non-western cultures, arXiv e-prints (2024) arXiv–2410.

[9] A. Maharana, M. Bansal, Integrating visuospatial, linguistic and commonsense structure into story visualization, arXiv preprint arXiv:2110.10834 (2021).

[10] A. Han, Z. Cai, Design implications of generative ai systems for visual storytelling for young learners, in: Proceedings of the 22nd annual ACM interaction design and children conference, 2023, pp. 470–474.

[11] V. N. Antony, C.-M. Huang, Id. 8: Co-creating visual stories with generative ai, ACM Transactions on Interactive Intelligent Systems 14 (2025) 1–29.

[12] J. Kim, Y. Heo, H. Yu, J. Nang, A multi-modal story generation framework with ai-driven storyline guidance, Electronics 12 (2023) 1289.

[13] K. Tewari, A. Malviya, S. Chanda, A. Mukherjee, S. Pal, Overview of the Shared Task on Multilingual Story Illustration: Bridging Cultures through AI Artistry, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '25, Association for Computing Machinery, New York, NY, USA, 2025.

[14] K. Tewari, A. Malviya, S. Chanda, A. Mukherjee, S. Pal, Findings of the Shared Task on Multilingual Story Illustration: Bridging Cultures through AI Artistry, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '25, CEUR Working Notes, 2025.

[3] https://grok.com