MUSIA: Multilingual Story Illustration Corpus for Cross-Cultural Alignment and Generation

Krishna Tewari[1], **Supriya Chanda**[2], **Nirmit Patil**[1], **Sukomal Pal**[1]
[1] Indian Institute of Technology(BHU) Varanasi
[2] Bennett University, Greater Noida
{krishnatewari.rs.cse24, spal.cse}@iitbhu.ac.in
{suplife24, nirmit045}@gmail.com

Abstract

Recent advances in text-to-image generation have enabled automated visual storytelling, yet most existing datasets remain monolingual and culturally narrow. We introduce MUSIA, a Multilingual Story Illustration Corpus designed to advance research in cross-lingual and culturally grounded narrative illustration. MUSIA comprises bilingual (English-Hindi) story-image pairs drawn from open literary and folk sources, curated to reflect diverse cultural themes, artistic styles, and linguistic structures. Each story includes multiple illustrations aligned at the scene level, accompanied by quality-verified mappings for narrative-visual coherence. To establish a reproducible benchmark, we propose a two-stage baseline combining transformer-based semantic summarization with diffusion-based image generation, achieving strong performance in relevance, visual quality, and consistency. MUSIA represents the first step toward a scalable, culturally inclusive benchmark for multilingual visual storytelling, enabling fair and reproducible research across low-resource and underrepresented languages.

Keywords: Multimodal storytelling, Text-to-image generation, Cultural representation, Narrative illustration

1. Introduction

The rapid development of artificial intelligence (AI), especially with the rise of diffusion-based image generation models and multimodal foundation models, has changed how we create content. Earlier large language models mainly produced text, but today's LLMs and text-to-image (T2I) systems can generate both coherent narratives and visual illustrations that fit those narratives (Brown et al., 2020; Dosovitskiy et al., 2021; Radford et al., 2021).

Visual storytelling, which combines narrative text and images, has become an important research area merging natural language processing (NLP) and computer vision. Automatically creating illustrations that match story text is relevant in fields like education, entertainment, digital publishing, and support for content creation. Specifically, creating multilingual illustrations can improve access to visual content and enhance cultural representation in the materials produced.

However, the advancement of generative technologies also brings challenges related to cultural authenticity, language diversity, visual consistency, and aesthetic quality. These issues are especially pronounced in culturally rich multilingual contexts like India. Here, narrative traditions vary from classical texts like *Panchatantra* and *Hitopadesha* to contemporary children's literature and oral stories. These traditions carry meaning through words, symbols, idioms, and cultural aesthetics, which are often missing in Western-focused training data. As a result, most illustration pipelines tend to reflect Western attire, architecture, and symbolism.

Previous work such as Pororo-SV (Li et al., 2019) showed that we can link narrative sequences to image sequences. However, Pororo-SV relies on a fixed set of characters and entities, while real-world multilingual stories include diverse characters, settings, and aesthetics that are hard to define ahead of time. This issue is even more complicated in multilingual contexts, where models must work across different scripts like Devanagari and various Dravidian scripts, handle diverse language structures, and maintain cultural meaning visually.

From a technical viewpoint, several obstacles remain in multilingual visual storytelling. First, low-resource languages, including most Indian languages, lack extensive pretraining data and well-organized multimodal datasets. Second, models often struggle to understand and align narratives that use different styles of syntax and discourse. Third, keeping character and scene consistency across multiple images can be fragile, as even minor changes in rendering can harm narrative flow (Maharana et al., 2022). Fourth, many generative models show cultural bias, leaning towards Western aesthetics rather than adapting to local visual styles.

While recent advances in vision transformers (Dosovitskiy et al., 2021), contrastive vision-language pretraining (Radford et al., 2021), and diffusion-based illustration techniques have improved the quality of generated content, most systems still operate in a single language and do not focus on cultural relevance.

In this work, we introduce a new dataset (MUSIA) that addresses this gap. We created a dataset for

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1914–1924
11–16 May 2026.

multilingual story illustration because no suitable resource is currently available in this area. Unlike earlier efforts, our dataset features paired story and illustration content in both Hindi and English, recognizing that even English lacks culturally relevant illustrated resources. To our knowledge, this is the first dataset specifically curated for multilingual and culturally sensitive story illustration. We also provide baseline models to help establish benchmarks and outline the main challenges faced during the development of these models to support future research.

Importantly, this resource is an evolving effort rather than a fixed release. The current version contains only narrative stories in English and Hindi, but subsequent releases will broaden linguistic coverage to additional low-resource Indian languages, and broaden textual scope to include poems and other literary forms. Each future version will be frozen and documented to preserve reproducibility while enabling continuous expansion.

The rest of this paper is organized as follows. Section 2 reviews the existing work in this domain and highlights the key innovations of previous approaches. The dataset creation is described in section 3. A baseline system and it's methodology is described in section 4. Section 5 presents the experimental results along with key evaluation metrics and discussion on the findings. Finally, the paper is concluded in section 6 along with future directions.

2. Related Work

In recent years, text-to-image generation has emerged as a core component in story synthesis, enabling the visual realization of narratives from textual descriptions. These generative models play a crucial role in maintaining semantic coherence between text and imagery and are frequently adopted as the generative backbone for story illustration tasks.

Contemporary advancements in text-to-image synthesis largely revolve around three major modeling paradigms: Generative Adversarial Networks (GANs), auto-regressive models, and diffusion models. Text-to-Image GANs (Zhu et al., 2019) rely on adversarial training between a generator and a discriminator to produce visually realistic images. Large-scale auto-regressive models such as DALL·E (Ramesh et al., 2021), Make-A-Scene (Gafni et al., 2022), and Parti (Yu et al., 2022) demonstrate impressive scalability and strong performance in coherent image generation from textual prompts. Diffusion-based approaches (Dhariwal and Nichol, 2021) including VQ-Diffusion (Gu et al., 2022), GLIDE (Nichol et al., 2022), DALL·E 2 (Ramesh et al., 2022), Latent Diffusion Models (LDM) (Rombach et al., 2022), and Imagen (Saharia et al., 2022) have gained substantial attention in recent years. As likelihood-based models, diffusion methods mitigate common issues found in GANs, such as mode collapse and training instability, while supporting the synthesis of diverse and semantically faithful visual outputs. Their robustness and capacity for high-fidelity image generation have made them the preferred foundation for recent research in narrative visualization and story illustration tasks.

Research on story visualization has developed through three main areas: adversarial architectures, adapting pretrained text-to-image models, and integrating multimodal large language models. Early methods like StoryGAN (Li et al., 2019) and its variants created conditional generative frameworks that produced sequential images from narrative text. They used recurrent and memory mechanisms to capture the timing relationships between sentences. Building on these ideas, models such as Improved StoryGAN (Li et al., 2020) and PororoGAN (Zeng et al., 2019) improved the generator with better convolutional structures like dilated and gated convolutions, as well as attention modules. These included aligned sentence encoders and patch-level discriminators to enhance both local realism and overall narrative coherence. However, these GAN-based approaches typically needed a lot of task-specific training data and tailored objective functions to balance fidelity and timing consistency.

Later research shifted to using pretrained text-to-image transformers for story visualization. This change helped lower training costs and improve generalization. StoryDALL-E (Maharana et al., 2022), for example, adapted a pretrained DALL-E model with sequential generation modules and copying mechanisms to keep visual continuity across frames. While these methods benefit from strong visual foundations, they often require fine-tuning or additional layers to maintain character identity and narrative flow. Other approaches, such as Character Preserving Coherent Story Visualization (CP-CSV) (Chen et al., 2022), added figure-ground segmentation and specific metrics like Frechet Story Distance to assess story-level coherence, emphasizing the need for structured supervision to keep character and scene consistency. Other efforts have looked at the reverse task of generating text from visuals. Frameworks that learn narrative structure from image-caption pairs demonstrated significant improvements in BLEU and METEOR scores compared to captioning baselines (Oliveira et al., 2025). Similarly, incorporating visual coherence losses helps maintain character and scene consistency across generated narratives (Hong et al., 2023). More recently, instruction-tuned multimodal LLMs have been used to achieve bet-

ter narrative coherence, emotional alignment, and contextual fidelity across image sequences (Yuan et al., 2025).

Several datasets have been proposed to facilitate research in story visualization and multimodal narrative understanding. The VIST (Visual Storytelling) dataset (Hu et al., 2024) provides sequences of images paired with corresponding narrative text, emphasizing the alignment between story events and visual content. PororoSV (Zeng et al., 2019) focuses on character-driven story sequences from the Pororo animated series, supporting tasks that require temporal consistency and identity preservation. Extensions such as FlintstoneSV (Gupta et al., 2018) and FlintstoneSV++ (Kapuriya and Buitelaar, 2025) capture narrative sequences from the Flintstones cartoons, with FlintstoneSV++ offering additional annotations for character segmentation and scene layout to better support structured story generation. OpenStory++ (Ye et al., 2024) expands the domain to open-ended stories, providing diverse textual narratives and associated illustrations to evaluate generalization across genres. Collectively, these datasets span character-centric, genre-specific, and open-domain narratives, providing essential benchmarks for advancing story visualization research.

In addition to datasets developed for story visualization, several benchmarks explore multimodal understanding in illustrated media such as cartoons and narrative imagery. One example is a benchmark derived from the New Yorker Caption Contest, which evaluates multimodal humor understanding by pairing cartoon images with multiple candidate captions and requiring systems to select the caption that best aligns with the visual context. The benchmark highlights the difficulty of modeling subtle cross-modal relationships where humor arises from implicit interactions between visual cues and linguistic interpretation rather than direct description (Hessel et al., 2023). Related work has also examined emotional understanding in cartoon imagery through large annotated datasets designed to capture affective signals present in illustrated narratives. Such approaches demonstrate that multimodal analysis of cartoons can provide insights into emotion, narrative intent, and contextual interpretation in visual storytelling scenarios (Jain et al., 2021).

Complementing these datasets, a community evaluation effort was introduced through the Multilingual Story Illustration Shared Task, organized by the authors as part of the Forum for Information Retrieval Evaluation (FIRE). The task focused on generating culturally grounded visual narratives from multilingual story texts in Hindi and English. Participating systems were required to produce coherent sequences of illustrations aligned with narrative events, enabling comparative evaluation across different modeling strategies. Human evaluation assessed system outputs using dimensions such as relevance, visual quality, and narrative consistency. Results from the shared task indicated that pipelines combining large language models for narrative-aware prompt construction with diffusion-based image generation produced the most coherent visual outputs, while maintaining character identity and stylistic consistency remained challenging problems (Tewari et al., 2026b). These findings further motivate the creation of the MUSIA dataset presented in this work, which extends the initial shared-task corpus into a larger benchmark designed to support multilingual and culturally grounded visual storytelling research.

3. Dataset Creation

This section describes the creation of our multilingual story illustration dataset, detailing the motivation and objectives, annotation strategies, dataset statistics, its organization and structure, preprocessing steps, quality assurance and other such details. The dataset and baseline notebook can be found at `https://github.com/Kriss-Tewari/MUSIA-at-LREC-2026`.

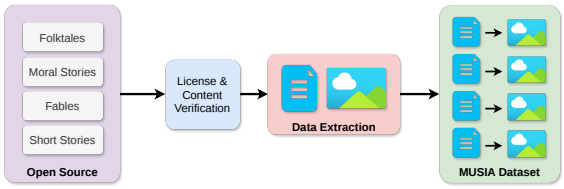


Figure 1: Pipeline for MUSIA dataset creation from open story sources

3.1. Motivation and Objectives

The primary objective of the MUSIA dataset is to facilitate research in *multilingual visual storytelling*, enabling AI systems to generate culturally grounded illustrations from narrative text. While previous multimodal datasets have focused on English text paired with captions or short descriptions, there is a scarcity of datasets that align *long-form narratives* with *scene-level visual representations*, particularly in Indian languages.

MUSIA dataset addresses this gap by introducing a bilingual corpus (English and Hindi) of narrative texts paired with human-created illustrations, reflecting diverse cultural, thematic, and visual styles.

3.2. Data Sources and Collection

The dataset was curated from *publicly available storybooks* and *open-licensed digital archives* that pro-

vide narrative content and accompanying illustrations under permissive licenses (e.g., CC-BY, Public Domain). Sources include educational repositories, open children's literature collections, and folk story archives.

To ensure cultural diversity, stories were selected from multiple genres such as folktales, moral stories, fables, and modern short stories, representing both contemporary and traditional Indian storytelling. Each story was manually verified for license compatibility and content suitability. For every selected story, the textual narrative and its corresponding images were extracted and organized into parallel files for the two target languages.

3.3. Data Organization and Structure

The dataset is provided in two languages: English and Hindi. The corpus is organized into two main folders, `Stories` and `Images`, under separate language-specific directories. It also contains a mapping file, which defines the numbers of images counts per story. The `Stories` folder contains narrative text files, with each story named following a consistent convention:

- `eng_story_XXXX` for English stories and
- `hin_story_XXXX` for Hindi stories,

where `XXXX` is a zero-padded numeric identifier (e.g., `eng_story_0001`).

Corresponding illustrations for each story are stored in the `Images` folder, with image file names following the pattern:

- `eng_story_XXXX_01`, `eng_story_XXXX_02`, and so on for English stories, and
- `hin_story_XXXX_01`, `hin_story_XXXX_02`, etc., for Hindi stories.

All image files are provided in standard formats such as `.jpg`, `.jpeg`, or `.png`.

This hierarchical organization is first divided by language, then by story and corresponding images that ensures clear narrative-visual alignment and facilitates seamless access for model training and evaluation.

3.4. Data Statistics

Our MUSIA dataset comprises two languages: English and Hindi. The English corpus contains approximately 508 stories, while the Hindi corpus includes around 401 stories. Each story is accompanied by one or more illustrative images, ranging from 1 to 20 per story, capturing key narrative events and character actions. A line chart (Figure 2) visually depicts the distribution of stories based on the number of images they contain for both Hindi and English datasets, highlighting variability in illustration density across stories.

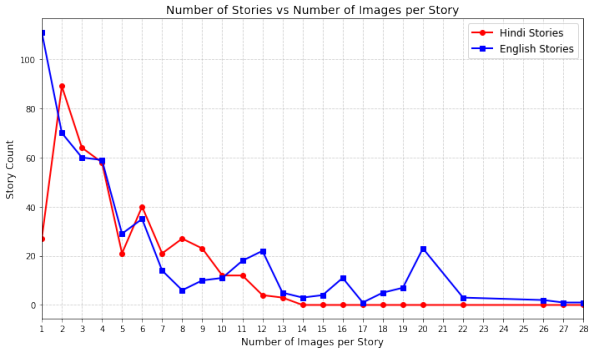


Figure 2: Line graph depicting the distribution of stories based on the number of images they contain for both Hindi and English datasets.

Table 1 summarizes key textual and narrative statistics for both languages, including the total word count, average words per story, number of unique characters, and average story length. These statistics provide an overview of dataset diversity and complexity, which can guide model design, training, and benchmarking.

Statistics	English	Hindi
Number of Stories	504	401
Average Characters per Story	5177.99	1239.45
Average Words per Story	963.48	257.69
Average Images per Story	5.92	5.85
Min Images per Story	1	1
Max Images per Story	22	13

Table 1: Dataset Statistics for English and Hindi Story Illustrations

To facilitate reproducible experiments and consistent model evaluation, we provide a mapping file that specifies the number of images associated with each story. This mapping can be used when creating training and testing splits to ensure that generated images match the number of ground-truth illustrations per story, preserving narrative-visual alignment and enabling fair benchmarking across models.

3.5. Preprocessing and Quality Assurance

To ensure textual consistency across the dataset, all stories were first normalized to UTF-8 encoding and carefully cleaned to eliminate typographical artifacts or formatting inconsistencies. Non-narrative elements such as author notes, advertisements, or unrelated metadata were systematically excluded to retain only the core narrative content. Paragraph

boundaries were intentionally preserved, as they play an essential role in maintaining the logical flow of the story and serve as useful markers for subsequent scene segmentation and image generation tasks.

For the image component, all illustrations were preserved in their original resolution and aspect ratio without any resizing or modification to ensure that the visual fidelity and artistic intent of the source material remained intact. Images exhibiting redundancy or low quality, such as those that were blurred, pixelated, or contained intrusive watermarks, were removed through a combination of automated filtering and manual inspection. Furthermore, each retained image underwent a careful visual verification process to confirm its relevance and alignment with the corresponding story segment, ensuring narrative coherence between text and visuals.

To further guarantee data quality and cultural authenticity, three bilingual reviewers independently validated the curated pairs. They assessed each story-image pair for linguistic accuracy, cultural appropriateness, and narrative correspondence. Only those pairs that achieved consensus among the reviewers were included in the final dataset, thereby ensuring both linguistic and visual integrity suitable for downstream multimodal analysis.

Table 2 shows a sample example from the curated MUSIA dataset for interpretation.

4. Methodology & Experiments

The proposed baseline establishes a fully automated and reproducible framework for generating story-aligned illustrations. The system is designed as a two-stage text-to-image synthesis pipeline that transforms narrative text into coherent visual scenes. In the first stage, each story is semantically condensed into short summaries that capture the essential narrative elements, characters, emotions, and actions, while filtering out linguistic redundancy. In the second stage, these condensed representations are converted into visually descriptive prompts and passed through a diffusion-based image generation model. The approach ensures that the generated illustrations are not only semantically faithful to the narrative but also consistent in their visual appearance across multiple story segments. All modules in this pipeline are implemented using the Hugging Face `transformers` and `diffusers` libraries, enabling seamless integration, reproducibility, and scalability.

4.1. Story Segmentation and Summarization

The pipeline begins by segmenting each story into smaller, semantically coherent units that represent distinct narrative events. These sub-stories correspond to different moments or scenes that together form the complete storyline. The segmentation process is guided by a predefined mapping file, `EN_story_image_counts.json`, which specifies how many illustrations should be produced for each story. Each sub-story is treated as an independent input to the summarization model, ensuring localized semantic representation while maintaining the overall continuity of the narrative.

To produce these summaries, we employ the transformer-based model, a fine-tuned version of summarizer trained specifically on the MUSIA training corpus. This model performs high-level semantic abstraction by condensing long-form stories into compact representations without losing narrative richness. The summarization pipeline, implemented through Hugging Face's `pipeline("summarization")` interface, automatically detects available CUDA hardware and leverages GPU acceleration for faster inference. Each sub-story is passed through this summarizer, which outputs concise yet contextually complete summaries that highlight essential narrative features such as protagonist identity, emotional tone, and primary actions. These summaries serve as semantically dense inputs for visual prompt construction in the subsequent stage.

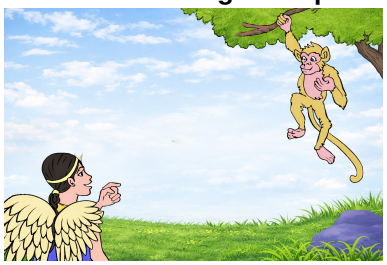
4.2. Visual Prompt Construction

Once the narrative summaries are generated, they are converted into visual prompts by appending a fixed visual style descriptor. This descriptor plays a critical role in maintaining stylistic consistency and visual coherence across all generated illustrations, ensuring that recurring characters and settings are represented uniformly throughout the story. The descriptor encodes several stylistic parameters, including the overall art style, color palette, shading intensity, texture, character consistency, and framing orientation. These visual specifications act as soft constraints that guide the diffusion model toward a uniform artistic domain.

The general structure of the descriptor follows a simple template:

```
[Style], [Color],
[Shading], [Texture],
[Character Consistency], [Framing].
```

For the baseline implementation, the descriptor is instantiated as: *Cartoon style, warm color palette, soft shading, hand-drawn texture, same characters, consistent clothing and face, wide frame.* This

Ground-truth image for Span 1	Ground-truth image for Span 2	Ground-truth image for Span 3
		

Story Span 1	"I want to be big," says Little Monkey. "I want to be strong." A wise woman hears him. "Take this magic wand," she says, "and all your wishes can come true."
Story Span 2	A giraffe comes by. He stretches his long neck. He eats the sweet leaves at the top of the trees. "I want a long neck," says Little Monkey. "POP!" His neck grows long, just like the giraffe's. Little Monkey is happy. An elephant comes down to the river. He fills his trunk with water. He blows it all over himself. "I want to do that too!" says Little Monkey. "BANG!" Just like that, he grows a trunk. He is very happy. "This is fun!" he says. Next, Little Monkey sees a zebra. "I want stripes like those," he says. "WHIZZ!" Little Monkey has stripes all over his body, just like the zebra. He is very, very happy.
Story Span 3	He goes to the river to try out his new trunk. He looks down. He sees himself in the water. "Mother!" he cries. "Help! A monster!" "That's not a monster," says his mother. "That's you." "You want a giraffe's neck, an elephant's trunk and stripes like a zebra. Don't you remember?" Little Monkey cries and cries. "I look awful!" he says. "I want to be myself again." There is a POP, a BANG and a WHIZZ. Little Monkey is himself again. He jumps for joy. He throws the magic wand into the river. He never wants to be anyone else again.

Table 2: Examples from the MUSIA-2025 dataset showing story text excerpts (below) and corresponding illustrations (top).

choice was empirically determined to yield visually pleasing and stylistically stable results while preserving narrative integrity. The final visual prompt for each segment is constructed by concatenating the fixed descriptor with its corresponding summary text. Formally, the overall prompt can be represented as:

$$\text{Prompt} = \text{Fixed Descriptor} + \text{Summary Text}$$

This ensures that every sub-story produces a distinct yet stylistically consistent visual output, resembling an illustrated narrative sequence.

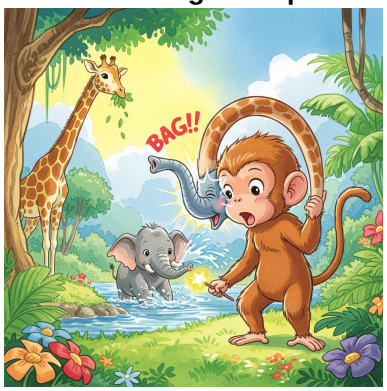
4.3. Image Generation

The generation of illustrations from the constructed prompts is performed using the Stable Diffusion XL (SDXL) model, developed by Stability AI. The model is accessed via the Hugging Face `diffusers` library and instantiated using the pre-trained checkpoint `stabilityai/stable-diffusion-xl-base-1.0`. The SDXL model is configured to operate in mixed precision (`torch.float16`) for optimized memory efficiency and computational throughput. To further enhance memory management, attention slicing is enabled, allowing the model to process large images within limited GPU memory constraints.

For each prompt, the model generates a high-resolution illustration using 30 inference steps and a guidance scale of 9 under the Classifier-Free Guidance (CFG) mechanism. This setup strikes a balance between visual fidelity and textual adherence, ensuring that the generated images remain faithful to the underlying summaries while exhibiting high perceptual quality. To maintain reproducibility, all random seeds are fixed prior to generation, ensuring that identical inputs yield consistent outputs across multiple runs. The resulting images are saved in 1024×1024 resolution, providing sufficient visual detail for downstream evaluation and qualitative analysis. Each story thus yields a sequence of visually coherent images that together form a structured multimodal narrative aligned with the source text.

4.4. Experimental Setup

All experiments were conducted on a high-performance computing setup equipped with NVIDIA GPUs supporting CUDA acceleration. The summarization model and image generator were both executed within the same Python environment

Generated image for Span 1	Generated image for Span 2	Generated image for Span 3
		

Story Span 1	"I want to be big," says Little Monkey. "I want to be strong." A wise woman hears him. "Take this magic wand," she says, "and all your wishes can come true." A giraffe comes by. He stretches his long neck. He eats the sweet leaves at the top of the trees. "I want a long neck," says Little Monkey. "POP!" His neck grows long, just like the giraffe's. Little Monkey is happy. An elephant comes down to the river. He fills his trunk with water. He blows it all over himself. "I want to do that too!" says Little Monkey. "BANG!" Just like that, he grows a trunk. He is very happy. "This is fun!" he says. Next, Little Monkey sees a zebra. "I want stripes like those," he says. "WHIZZ!" Little Monkey has stripes all over his body, just like the zebra. He is very, very happy.
Story Span 2	He goes to the river to try out his new trunk. He looks down. He sees himself in the water. "Mother!" he cries. "Help! A monster!" "That's not a monster," says his mother. "That's you." "You want a giraffe's neck, an elephant's trunk and stripes like a zebra. Don't you remember?" Little Monkey cries and cries. "I look awful!" he says. "I want to be myself again."
Story Span 3	There is a POP, a BANG and a WHIZZ. Little Monkey is himself again. He jumps for joy. He throws the magic wand into the river. He never wants to be anyone else again.

Table 3: Generated illustrations from the baseline experimentation

to ensure end-to-end compatibility. The required libraries included `transformers`, `diffusers`, `torch`, and `PIL`, which collectively support text processing, diffusion inference, and image handling. The summarization and image generation modules were executed sequentially to manage GPU memory effectively, allowing the system to scale across multiple stories without manual resource reallocation.

The complete experimental workflow follows a deterministic and modular structure. First, the mapping file is loaded to determine the number of substories and corresponding images per story. Next, each story is divided into smaller narrative units, and these units are summarized using the fine-tuned MUSIA summarizer. The summaries are then concatenated with the fixed visual style descriptor to form the final image prompts. These prompts are passed through the SDXL pipeline to generate corresponding illustrations, which are stored systematically in submission-ready directories following the MUSIA 2025 format. The final output comprises aligned text-image pairs that can be utilized for downstream multimodal tasks such as visual question answering, story comprehension, or narrative retrieval.

Overall, this baseline provides a clear, modular, and reproducible methodology for automated story-to-illustration generation. Its design emphasizes semantic accuracy, stylistic coherence, and reproducibility, qualities essential for multimodal storytelling research. By combining transformer-based summarization with diffusion-based synthesis, the pipeline bridges the gap between narrative abstraction and visual realization, producing coherent and artistically unified illustrations that faithfully represent the narrative semantics of the source stories.

5. Results and Discussion

Evaluating the quality of story-to-illustration generation remains a challenging task due to the inherently subjective nature of creativity, visual semantics, and narrative coherence. Following the evaluation strategy adopted in prior literature on text-to-image

Approach	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
LLM-guided prompt engineering + Imagen (Kachhadiya and Patel, 2025)	39	0	0	34	5	0	24	14	1
1Prompt1Story diffusion (Surendra and Divya, 2025)	36	2	0	5	21	12	8	18	12
Translation + summarization + diffusion (Sadhukhan et al., 2025)	15	22	2	3	10	26	3	8	28
Hybrid summarization + SDXL diffusion (Mannan et al., 2025)	0	0	39	0	6	33	0	2	37

Table 4: Human evaluation results from the shared task (Tewari et al., 2026a) across English stories for different generation approaches.

Approach	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
LLM-guided prompt engineering + Imagen (Kachhadiya and Patel, 2025)	30	0	0	30	0	0	27	3	0
1Prompt1Story diffusion (Surendra and Divya, 2025)	30	0	0	10	15	5	15	10	5
Translation + summarization + diffusion (Sadhukhan et al., 2025)	19	11	0	0	0	30	0	3	27
Hybrid summarization + SDXL diffusion (Mannan et al., 2025)	0	0	30	2	4	24	0	2	28

Table 5: Human evaluation results from the shared task (Tewari et al., 2026a) across Hindi stories for different generation approaches.

generation and multimodal storytelling, we relied primarily on human evaluation, as automatic metrics such as FID or CLIPScore do not fully capture the narrative or contextual alignment required for this task. To ensure consistency and reliability, four independent annotators were recruited to assess the generated outputs using a three-point Likert scale where 1 corresponds to *Fair*, 2 to *Moderate*, and 3 to *Good*. Each generated story was rated across three distinct dimensions: Relevance, Visual Quality, and Consistency.

Relevance measures how accurately the generated illustrations capture the core events, characters, and emotions described in the textual story (Text → Image alignment). Visual Quality assesses the overall artistic appeal, realism, and aesthetic coherence of the generated outputs, focusing on color balance, texture, and composition. Consistency evaluates whether the entire sequence of images collectively maintains a coherent narrative flow (Image → Image continuity), including aspects such as recurring character depiction, background stability, and visual transitions between scenes.

Each annotator evaluated the full sequence of illustrations corresponding to a story, and the average scores across annotators were computed to derive a consensus judgment. This evaluation mechanism allows for a more nuanced understanding of multimodal coherence than purely quantitative metrics, aligning with methodologies used in contemporary text-to-image benchmarks.

To further provide transparency regarding the human evaluation process, we report results obtained during the Multilingual Story Illustration shared task pilot benchmark (Tewari et al., 2026a), where multiple generation approaches were evaluated on a subset of the dataset using the same evaluation criteria. Table 4 summarizes the performance according to the underlying generation approaches used in the shared task systems for English language and Table 5 depicts similar results for Hindi language. These results highlight the relative strengths of different generation pipelines across the three evaluation dimensions.

Table 3 presents representative illustrations generated by the proposed baseline model. The visual outputs demonstrate strong performance across all three evaluation dimensions. In particular, the generated images exhibit high *Relevance*, effectively capturing key narrative moments while visually representing both central and secondary characters. The *Visual Quality* of the images is notably high, featuring balanced color tones, well-defined shapes, and stylistic uniformity that aligns with the "cartoon style" descriptor used during prompt construction. Furthermore, the *Consistency* between consecutive frames remains robust, characters retain their identity and attire across scenes, and the visual environment evolves smoothly to reflect the unfolding storyline.

An additional qualitative observation is that the baseline model successfully generates multiple characters in a single frame while preserving spatial and emotional coherence. Moreover, it captures subtle elements of cultural diversity, reflected through background motifs, and color themes that are faithful to the regional and narrative context of the source stories. These qualitative strengths collectively indicate that even a simple summarization-diffusion pipeline can yield visually coherent and semantically rich storytelling results when trained on culturally rich and diverse dataset.

6. Conclusion and Future Work

In this paper, we introduced MUSIA, a multilingual and culturally grounded story illustration dataset designed to advance research in multimodal narrative understanding. By aligning bilingual story texts with corresponding illustrations, MUSIA addresses a major gap in current multimodal resources that primarily emphasize monolingual and Western-centric narratives. The dataset includes diverse English and Hindi stories across genres such as folktales,

moral stories, and children's literature, highlighting linguistic diversity, cultural authenticity, and narrative-visual coherence.

We also presented a reproducible baseline for automated story-to-illustration generation that combines transformer-based semantic summarization with diffusion-based image synthesis. In the baseline experiments we observed that this two-stage design may produce faithful illustrations, demonstrating reliable performance in terms of relevance, visual quality and narrative consistency.

In future work, we plan to expand MUSIA to include additional Indian and low-resource languages such as Bengali, Marathi, and Dravidian languages, along with new literary forms like poems, comics, and folk narratives. We also aim to explore instruction-tuned multimodal large language models (MLLMs) with reinforcement learning from human feedback (RLHF) for enhanced contextual and cultural alignment. Ultimately, we envision MUSIA as a continuously evolving benchmark that supports fair, reproducible, and culturally inclusive research in multilingual visual storytelling.

Ethics Statement

All content included in the MUSIA-2025 dataset is sourced exclusively from open-access and public-domain materials, with clear attribution provided to original creators wherever applicable. The dataset contains no personally identifiable information, and its use is restricted strictly to educational and research purposes.

To ensure responsible research and prevent potential misuse, all stories and illustrations have been carefully screened to exclude any material containing violent, religiously sensitive, or otherwise inappropriate content. The dataset is released under the Creative Commons Attribution - NonCommercial 4.0 International (CC BY - NC 4.0) license, permitting use for academic and non-commercial research only.

We recognize the importance of ethical stewardship in multimodal AI research and have taken deliberate measures to align this work with principles of transparency, inclusivity, and cultural sensitivity. The broader impact of this resource lies in supporting fair, reproducible, and socially responsible advances in story-to-image generation and multimodal understanding.

7. References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. 2022. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8259–8272, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 604–620. Springer.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10696–10706.

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, page 610–626, Berlin, Heidelberg. Springer-Verlag.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. *Transactions of the Association for Computational Linguistics*, 11:688–705.

Xudong Hong, Vera Demberg, Asad Sayeed, Qiankun Zheng, and Bernt Schiele. 2023. Visual coherence loss for coherent and visually grounded story generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9456–9470, Toronto, Canada. Association for Computational Linguistics.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2024. Visual storytelling dataset (vist). https://service.tib.eu/ldmservice/dataset/visual-storytelling-dataset--vist-. Accessed: 2025-10-25.

Ankit Jain, Anand Mishra, and C. V. Jawahar. 2021. Understanding cartoon emotion using integrated deep neural network on large dataset. *Pattern Recognition Letters*, 145:148–154.

Kishan Kachhadiya and Parth Patel. 2025. Leveraging large language model (llm) and v-llm for zero-shot multilingual story illustration. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2025)*, volume 4173 of *CEUR Workshop Proceedings*, pages 113–120. CEUR-WS.org.

Janak Kapuriya and Paul Buitelaar. 2025. Flintstonessv++: Improving story narration using visual scene graph. In *Proceedings of the 8th Workshop on Narrative Extraction From Texts (Text2Story 2025)*, volume 3964 of *CEUR Workshop Proceedings*. Accessed: 2025-10-25.

Chunye Li, Liya Kong, and Zhiping Zhou. 2020. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, 73:102956.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6322–6331, Los Alamitos, CA, USA. IEEE Computer Society.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 70–87, Berlin, Heidelberg. Springer-Verlag.

Shazia Mannan, Asha Hegde, and Sharal Coelho. 2025. Bridging cultures through ai: The art of multilingual storytelling. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2025)*, volume 4173 of *CEUR Workshop Proceedings*, pages 139–143. CEUR-WS.org.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. Glide: Photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR.

Daniel A. P. Oliveira, Eugénio Ribeiro, and David Martins de Matos. 2025. Story generation from visual inputs: Techniques, related tasks, and challenges. *Information*, 16(9).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society.

Mrinmoy Sadhukhan, Indrajit Bhattacharya, and Paramartha Dutta. 2025. Narrart: Multilingual story illustration with ai for english and hindi narratives. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2025)*, volume

4173 of *CEUR Workshop Proceedings*, pages 130–138. CEUR-WS.org.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sharma Nandini Surendra and Divya. 2025. Multilingual story illustration for musia 2025 using one-prompt-one-story image generation. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2025)*, volume 4173 of *CEUR Workshop Proceedings*, pages 121–129. CEUR-WS.org.

Krishna Tewari, Anshita Malviya, Supriya Chanda, Arjun Mukherjee, and Sukomal Pal. 2026a. Findings of the shared task on multilingual story illustration: Bridging cultures through ai artistry (musia). In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2025) Working Notes*, volume 4173 of *CEUR Workshop Proceedings*, pages 101–112. CEUR-WS.org.

Krishna Tewari, Anshita Malviya, Supriya Chanda, Arjun Mukherjee, and Sukomal Pal. 2026b. Overview of the shared task on multilingual story illustration: Bridging cultures through ai artistry (musia). In *Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '25, pages 1–4. Association for Computing Machinery.

Zilyu Ye, Jinxiu Liu, Ruotian Peng, Jinjin Cao, Zhiyang Chen, Yiyang Zhang, Ziwei Xuan, Mingyuan Zhou, Xiaoqian Shen, Mohamed Elhoseiny, Qi Liu, and Guo-Jun Qi. 2024. Openstory++: A large-scale dataset and benchmark for instance-aware open-domain visual storytelling. *arXiv preprint arXiv:2408.03695*.

Jiahui Yu, Ting Wang, Zhe Zhang, Zhiwei Zhang, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yangshu Yuan, Heng Chen, and Christian Ng. 2025. Instruction tuning for story understanding and generation with weak supervision. *arXiv preprint arXiv:2501.15574*.

Gangyan Zeng, Zhaohui Li, and Yuan Zhang. 2019. Pororogan: An improved story visualization model on pororo-sv dataset. *Proceedings of the 3rd International Conference on Computer Science and Artificial Intelligence*, pages 1–5.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810.