Findings of the Shared Task on Multilingual Story Illustration: Bridging Cultures through AI Artistry (MUSIA)

Krishna Tewari[1,*], Anshita Malviya[1], Supriya Chanda[2], Arjun Mukherjee[1] and Sukomal Pal[1]

[1]*Indian Institute of Technology (BHU), Varanasi, INDIA*
[2]*Bennett University, Greater Noida, INDIA*

Abstract

The Multilingual Story Illustration Shared Task (MUSIA), held under FIRE-2025, looks into the challenge of creating culturally grounded visual narratives for short stories in English and Hindi. As multimodal generative AI becomes more important in education, digital storytelling, and creative content creation, MUSIA offers the first benchmark focused on culturally accurate, multilingual story visualization. Eight teams signed up for the task; five submitted valid system runs, and four provided camera-ready papers. The systems used different strategies like narrative segmentation, translation, summarization, prompt design, retrieval-augmented methods, and diffusion-based generation. Human evaluation assessed three areas: visual quality, relevance, and consistency. The results showed that pipelines combining LLM-based story understanding with diffusion models achieved the best outcomes, especially when it came to creating visually coherent images. However, most systems had trouble maintaining narrative fidelity and consistency across panels. None used the culturally rich training illustrations offered in MUSIA, which led to a noticeable Westernization of the generated images. These results reveal ongoing limitations in current text-to-image models, particularly their struggle to accurately reflect Indian cultural elements such as regional clothing, landscapes, and folk designs. This paper discusses the MUSIA dataset, task formulation, team methods, and comparative results, laying the groundwork for creating multilingual, culturally aware story-illustration systems and guiding future research in inclusive multimodal generation.

Keywords

Mulitmodal storytelling, Text-to-image generation, Cultural representation, Narrative illustration

CEUR-WS.org/Vol-4173/T2-1.pdf

1. Introduction

Recent advances in multimodal Artificial Intelligence (AI), driven by large language models (LLMs) and diffusion-based text-to-image (T2I) generators, have greatly expanded the range of applications needing joint reasoning over language and vision. Among these applications, story illustration has become an important but underexplored task. It involves turning narrative passages into a series of images that accurately reflect the story's events, characters, and setting. While leading models perform reasonably well on English story datasets from Western media [1, 2], they often struggle with narratives that differ linguistically or culturally from their training data.

This limitation is especially clear in Indian storytelling, which is multilingual and culturally rich. Stories written in Hindi, English, Bengali, and other Indian languages often contain detailed cultural cues, including traditional clothing, local architecture, regional landscapes, folk motifs, and idiomatic expressions. However, diffusion models mainly trained on Western visuals frequently misrepresent these elements. As a result, illustrations for Indian stories often show incorrect clothing, non-Indian characters, or Western-style backgrounds, breaking cultural and narrative trust.

Forum for Information Retrieval Evaluation, December 17-20, 2025, India
*Corresponding author.
krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); anshitamalviya.rs.cse23@iitbhu.ac.in (A. Malviya); suplife24@gmail.com (S. Chanda); arjunmukherjee.rs.cse23@iitbhu.ac.in (A. Mukherjee); spal.cse@iitbhu.ac.in (S. Pal)
0009-0005-6599-9956 (K. Tewari); 0009-0004-8647-0245 (A. Malviya); 0000-0002-6344-8772 (S. Chanda); 0009-0007-4322-3537 (A. Mukherjee); 0000-0001-8743-9830 (S. Pal)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-24

Existing resources like VIST [3], PororoSV [4], FlintstonesSV [5], and OpenStory++ [6] have advanced story visualization, but they are limited to English narratives and Western imagery. Consequently, models based on these datasets are not well-suited for the multilingual and culturally diverse nature of Indian stories.

To fill this gap, we introduce the FIRE-2025 Shared Task on Multilingual Story Illustration (MUSIA)[1], a new benchmark designed to evaluate systems that generate culturally appropriate illustrations for short stories in Indian languages. This edition focuses on Hindi and English. Participants receive training stories with example illustrations and must produce a specific number of images for each test story indicated in a mapping file.

The task presents several challenges. Systems must capture key moments in the narrative while ensuring that characters, backgrounds, and visual styles remain consistent across multiple panels. At the same time, they must incorporate culturally relevant features that reflect the linguistic and regional context of the story. A fully fixed character template can make these requirements too simple, while unconstrained generation risks style drift and inconsistency. The MUSIA dataset offers diverse, attribution-compliant illustrations to help guide models toward culturally faithful generation.

Evaluation is entirely based on human assessment across three areas: relevance, consistency, and visual quality. Reviewers rate each area using a three-level scale (Good, Moderate, Fair), allowing for a detailed comparison of different modeling strategies in multilingual and multicultural settings.

MUSIA thus establishes the first standardized framework for multilingual, culturally grounded story illustration in the Indian context. By emphasizing both linguistic diversity and cultural authenticity, this shared task aims to encourage the development of generative models that can better support applications in education, children's media, digital storytelling, and creative content production.

The rest of the paper is structured as follows: Section 2 discusses related work; Section 3 describes the dataset; Section 4 presents the proposed methodology; Section 5 reports results and analysis; and Section 6 concludes with key findings.

2. Related Work

Recent advances in vision language models have led to highly capable systems that can generate images and interpret narratives from text. Earlier work mainly targeted single-image generation from short descriptions, but newer approaches move toward visualizing entire stories from longer passages. This section highlights key methods and datasets that underpin the MUSIA multilingual story illustration track.

2.1. Text-to-Image Generation

Early work on text-conditioned image generation largely relied on GAN-based architectures, with an emphasis on improving visual realism and semantic alignment for single images described by short captions. DM-GAN, for example, augments the generator with a dynamic memory module that repeatedly updates visual features using textual cues, leading to sharper images and better coverage of fine-grained details in complex descriptions [7]. Follow-up approaches move toward more structured forms of conditioning: Make-A-Scene lets users provide a high-level scene layout or semantic map, which is then fused with the input text so that the generated images adhere more closely to spatial structure and human-specified priors [8].

In parallel, another thread of research scales text-to-image models using large datasets and autoregressive formulations. Zero-shot systems such as DALL·E cast image synthesis as predicting sequences of discrete visual tokens conditioned on a textual prompt, and show strong compositional generalization without any task-specific fine-tuning [9]. Autoregressive transformers trained on web-scale image–text corpora further increase diversity and richness of generated content, illustrating that sheer scale can bridge much of the performance gap relative to supervised models [10]. These advances are

[1] https://cse-iitbhu.github.io/MUSIA/

underpinned by stronger visual and multimodal encoders, including Vision Transformers [11] and contrastive vision–language pretraining frameworks like CLIP [12], which supply robust priors for aligning images and text.

More recently, diffusion-based methods have become the predominant choice for high-quality image synthesis. Foundational work on unconditional diffusion models shows that, with appropriate parameterization, they can outperform GANs both in sample quality and in coverage of the data distribution [13]. Building on this, GLIDE demonstrates that text-guided diffusion can produce photorealistic images and supports flexible, prompt-driven editing via classifier-free guidance [14]. Latent diffusion further improves computational efficiency by running the diffusion process in a learned latent space, enabling high-resolution generation at manageable cost [15]. Large-scale systems such as Imagen and related models extend these ideas with powerful language encoders and massive training corpora, achieving strong zero-shot performance across standard text-to-image benchmarks [16]. Complementary work on vector-quantized diffusion [17] explores discrete latent representations, offering alternative trade-offs in training stability and sampling speed.

2.2. Story Visualization and Visual Storytelling

While the models discussed above are mainly designed for generating a single image from text, story visualization has the additional requirement of preserving coherence across a series of images driven by a longer narrative. StoryGAN casts this task as sequential conditional generation, using a story encoder together with a recurrent generator so that each panel reflects not only the current sentence but also prior context across the story [18, 1]. PororoGAN extends this idea by strengthening character-specific representations and improving temporal consistency on the Pororo-SV benchmark [4]. Moving beyond these early GAN-based solutions, StoryDALL-E leverages large pretrained text-to-image transformers for story continuation: each frame is conditioned on the current sentence as well as a compact summary of earlier frames, which substantially enhances character and style consistency over longer sequences [19, 20]. More recent character-preserving methods explicitly model visual plans and token-level alignments to keep track of entities across panels [21].

Advances in story visualization are tightly coupled with the availability of dedicated datasets and benchmarks. Pororo-SV and FlintstonesSV provide short, cartoon-like narratives with aligned frame sequences that are well suited for studying story-level generation [18, 4, 22]. FlintstonesSV++ further enriches this setting with additional annotations and visual scene graphs, enabling more detailed reasoning about objects and their interactions [23]. Complementary resources such as the Visual Storytelling Dataset (VIST) pair real-world photo streams with human-written stories, supporting research on aligning visual and textual narratives [3]. OpenStory++ extends this direction to large-scale, open-domain, instance-aware visual storytelling with more diverse story types and visual content [6]. Taken together, these datasets underscore the need to evaluate both image quality and narrative coherence, but they largely focus on English and on relatively narrow visual domains (for example, specific cartoons or TV shows).

2.3. Multimodal Models for Narrative Understanding and Generation

Beyond pure image synthesis, a line of work studies multimodal models that tell stories from visual inputs and jointly reason over images and text. Narrative generation frameworks for image sequences aim to produce flowing, coherent stories that unfold over time, instead of treating each frame as an isolated captioning problem [24]. To better enforce temporal consistency, some approaches introduce visual coherence losses for image-based story generation, discouraging sudden shifts in meaning or style between adjacent sentences [25]. At a larger scale, multimodal language models such as Flamingo [26] bring few-shot generalization to image–text tasks by conditioning a strong language model on interleaved visual and textual tokens. On the purely textual side, GPT-3 shows that large language models can handle a wide range of narrative and generative tasks with minimal examples [27], and later

work adapts such models to storytelling by instruction-tuning and extending them to multimodal settings [28].

These advances rely heavily on general-purpose visual and multimodal backbones. Architectures like Vision Transformers [11] and contrastively trained vision–language models such as CLIP [12] offer robust shared representations for images and text, which many narrative generation systems reuse as frozen encoders or as building blocks within larger pipelines. Still, most existing multimodal narrative systems are designed for a single language or a limited set of visual domains, and they rarely tackle the challenge of maintaining consistent characters and style across long story arcs.

Taken together, prior work provides strong building blocks for text-to-image generation, story visualization, and multimodal narrative modeling. However, there is still a lack of benchmarks that explicitly target *multilingual*, culturally diverse story illustration, where models must generate a *sequence* of panels that are both narratively appropriate and visually consistent. Datasets such as Pororo-SV, FlintstonesSV, VIST, and OpenStory++ [18, 4, 22, 3, 6, 23] tend to focus on one language, a specific visual style, or do not place strong emphasis on fine-grained character consistency in illustrated stories. The MUSIA track addresses this gap by casting story illustration as a multilingual, story-level text-to-image task rooted in children's storybooks, with evaluation criteria that deliberately stress relevance, coherence, and visual quality across the full narrative.

3. Dataset

The MUSIA dataset is constructed from openly licensed storybooks and digital archives that provide narrative text and illustrations under permissive terms (CC-BY or Public Domain). Our sources span educational repositories, community-driven children's literature platforms, and curated folk-story collections. This diversity enables the dataset to capture a wide spectrum of Indian storytelling traditions, including classic folktales, moral narratives, fables, and contemporary short fiction. Each candidate story undergoes manual verification to confirm that its licensing permits redistribution and that both its textual and visual content are appropriate for child-oriented contexts.

For every story that satisfies these criteria, we compile the complete narrative alongside all corresponding illustrations. The training split is offered in two languages, English and Hindi. It is organized into language-specific directories with separate `Stories` and `Images` folders. Story files follow a consistent naming pattern:

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

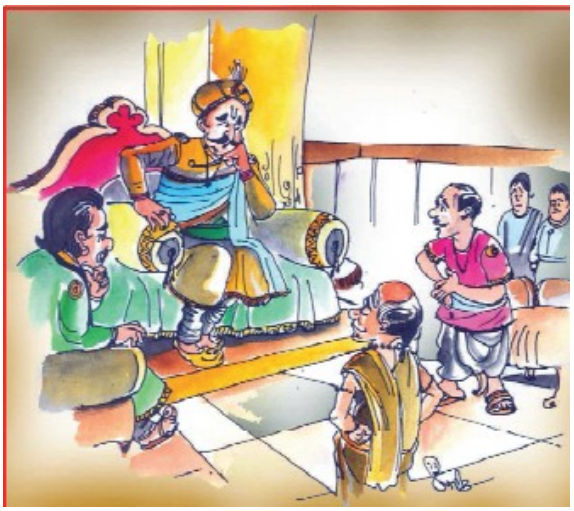
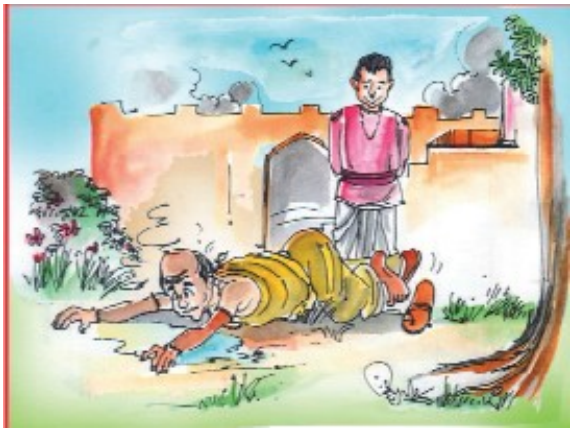
where `XXXX` denotes a four-digit, zero-padded identifier (e.g., `eng_story_0001`). Images retain the same numeric base and include an additional two-digit index indicating their order within the story (e.g., `eng_story_0001_01`). All images are stored in standard formats such as `.jpg`, `.jpeg`, or `.png`.

The final training set consists of 360 English stories and 185 Hindi stories, each linked with its full set of illustrations. The test set contains 40 English and 30 Hindi stories. However, only the story text and a mapping file specifying the required number of illustrations per story are provided, ensuring that evaluation proceeds in a completely generative manner without reference images.

To maintain textual quality and consistency, all stories were normalized to UTF-8 encoding, cleaned to remove stray symbols or formatting artifacts, and filtered to eliminate non-narrative content such as author notes, advertisements, and extraneous metadata. Paragraph structure was preserved to maintain the narrative flow and to facilitate downstream tasks such as scene segmentation and panel-level generation.

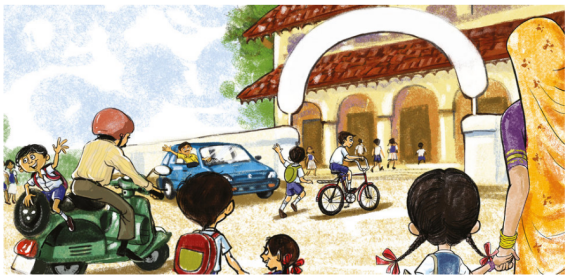
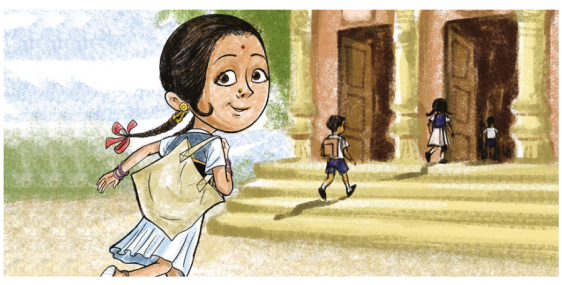
For the visual component, illustrations were kept at their original resolution and aspect ratio to preserve artistic fidelity. Images with severe noise, compression artifacts, low resolution, or intrusive watermarks were removed through a mix of automatic checks and manual review. Each remaining image was then cross-verified for relevance and narrative alignment to ensure that it accurately reflected the corresponding portion of the story.

Table 1
Example from the English dataset showing story text excerpts (left) and corresponding illustrations (right).

Story Excerpt (Text)	Illustration (Image)
Raja Krishna Chandra ruled over a part of Bengal about two hundred years ago. His court jester was Gopal Bhand. Though Gopal Bhand had not studied books, he was a very wise man. Once, a very learned man, Mahagyani Pandit came to the court. He spoke all the Indian languages fluently and perfectly. He answered all the questions very wisely. People were amazed to talk to him but no one could identify his mother tongue. Whenever they asked him, he smiled arrogantly. He said, "A truly wise man will easily know my mother tongue." Raja Krishna Chandra was very upset. So he announced a reward for anyone who could tell the Pandit's mother tongue. All the scholars listened to Mahagyani attentively. But no one could identify his mother tongue. "Shame on you", said the king angrily. All the scholars were silent. Gopal Bhand stood up hesitantly. He said, "Your Highness, give me a chance." "How could you tell?", asked the king. "Your Highness! I won't talk. He will tell you himself", answered Gopal Bhand.	
The next morning the king was walking in his garden. Gopal Bhand ran upto him quickly and said, "I have told Mahagyani Pandit that you are going to honour him with a garland of roses." "What!", said the king surprisingly. The next moment the king saw Mahagyani Pandit walking in expectantly. He was in silk clothes.	
Gopal Bhand hid himself behind the hedge. As soon as the Pandit came near the hedge, he put his leg out and tripped the Pandit. The Mahagyani pandit fell down on the freshly watered ground. He sat up and shouted at Gopal Bhand in his mother tongue. Gopal Bhand said, "Your Highness, now you know, what the Pandit's mother tongue is!" Mahagyani Pandit got up and said to Gopal Bhand, "You wise man, you have trapped me intelligently," and he went away.	

Quality control involved three bilingual annotators who independently assessed every text–image pair for linguistic accuracy, cultural appropriateness, and narrative correspondence. Only pairs that received unanimous approval were included in the final dataset, ensuring high standards of multimodal

Table 2
Example from the Hindi dataset showing story text excerpts (left) and corresponding illustrations (right).

Story Excerpt (Text)	Illustration (Image)
स्कूल में आज मेरा पहला दिन है। माँ मेरा हाथ पकड़े हुए हैं और मेरे साथ चल रही हैं। “मैं अब बड़ी हो गई हूँ." मैं कहती हूँ। “चलो... चलो !" माँ ने मेरा हाथ कसकर पकड़ा हुआ है।	
स्कूल के पास बहुत सारे बच्चे हैं। कुछ बस से आते हैं। कुछ कार से आते हैं। कुछ रिक्शे से आते हैं। कुछ साइकिल से और कुछ पैदल आते हैं, मेरी तरह। हम फाटक तक पहुँचे। माँ मेरा हाथ छोड़ती हैं।	
वह गेट पर रुक गई। मुझे अंदर अकेले जाना है। मेरे चारों तरफ़ बहुत से अनजाने चेहरे हैं। मैं एक क़दम चलती हूँ। मैं दूसरा क़दम बढ़ाती हूँ। मैं पीछे मुड़कर देखती हूँ। जैसे मैं आगे बढ़ती जाती हूँ, माँ छोटी दिखती जाती हैं। क्या वह गायब हो जाएँगी?	
मैं दौड़कर उनके पास जाती हूँ। मुझे नहीं लगता कि मैं बड़ी हो चुकी हूँ। मैं उनका हाथ पकड़ती हूँ और कहती हूँ, “मत जाओ!" सभी अंदर जा चुके हैं। सिर्फ मैं बाहर हैं।	
टीचर दीदी बाहर आती हैं। वे मुझे देख मुस्कराती हैं। मैं भी मुस्कराती हूँ।माँ कहती हैं, “रानी जब तुम बाहर आओगी मैं यहीं मिलूँगी।" मैं उनका हाथ छोड़ देती हूँ। वह हाथ हिलाती हैं।	
मैं दौड़कर अंदर जाती हूँ। माँ छुट्टी होने पर वहीं मिलेंगी!	

coherence and cultural integrity.

Table 1 and Table 2 presents representative examples from MUSIA-2025 english and hindi dataset respectively, illustrating the alignment between narrative passages and their associated illustrations.

4. Methodology

Eight teams registered to participate in the MUSIA-2025 shared task, indicating strong engagement with the challenge of multilingual story illustration. As the evaluation stage drew near, five of these teams were able to submit valid system runs for review, and four continued all the way through to prepare and submit camera-ready papers that carefully presented their approaches and results.

The team **NandiniDivya** [29] builds on the One-Prompt-One-Story (1Prompt1Story) framework, using it as a training-free backbone to generate consistent illustrations for each story. Their pipeline begins by putting all the text into a common form: English stories are used as they are, while Hindi stories are first translated into English with a large language model so that the rest of the process runs in a single language. The MUSIA mapping file is then consulted to find out how many images are needed per story, and the LLM is asked to rewrite the narrative into exactly that many short, scene-wise descriptions. Instead of treating these scenes separately, the team stitches all of them together into one long, comma-separated prompt, so that character details and overall plot context are visible to the model at once rather than in isolation. This combined prompt is fed into the 1Prompt1Story framework, which augments a diffusion-based generator (Playground AI) with components such as Singular-Value Reweighting and Identity-Preserving Cross-Attention to stabilise character identity and important visual cues across the full sequence, while still allowing natural changes in viewpoint and background. From this shared prompt, the system produces one image for each scene, resulting in a run of illustrations that stay aligned with the story and look coherent as a set, all without any extra fine-tuning on the MUSIA dataset.

The **NLPFusion** team [30] design a hybrid pipeline that combines story understanding with diffusion-based image generation to address the MUSIA shared task. Their system handles each English or Hindi story in three broad steps: preparing the text, summarising the narrative, and finally creating the illustrations. To begin with, the raw story is cleaned to remove unwanted symbols and formatting, then broken into sentences and grouped into consecutive chunks so that the number of "mini-stories" lines up with the number of images required for that story. Each chunk is meant to capture one scene that deserves its own illustration. These chunks are then fed into a T5-large abstractive summarisation model, run with fixed length constraints and deterministic decoding, to produce short scene descriptions that focus on the main actions and characters while avoiding unnecessary repetition. These summaries serve as the core of the image prompts. In the last stage, the team uses the Stable Diffusion XL (SDXL 1.0) model, running in 16-bit precision on a GPU, to generate the images. A fixed cartoon-style prefix is added to every summary to keep the visual look consistent, and the combined prompt is passed to the SDXL pipeline to produce one PNG image per scene, stored in language-specific folders with basic logging and error handling for smooth batch processing. Overall, this step-by-step setup converts raw multilingual stories into a set of visually consistent, story-faithful illustrations tailored to the MUSIA evaluation.

The team **Retriever** [31] approaches the MUSIA shared task with a purely zero-shot strategy, relying on large foundation models and prompt engineering instead of training any custom models. They work directly on the official MUSIA test set, which contains 39 English and 30 Hindi stories, each annotated with the exact number of images to be generated. The same pipeline is applied to both languages by making use of Gemini-2.5-Pro's multilingual capabilities. For each story, Gemini is first asked to produce a single high-level "system prompt" that defines a child-friendly visual style for the entire story, and then to write one detailed image prompt for every required illustration, each capturing a specific scene or turning point in the narrative. A carefully crafted meta-prompt guides Gemini using structured tags for aspects such as style, colour, shading, texture, character persistence, and framing, and instructs it to return the outputs in a Python dictionary format for easy downstream processing.

In the generation phase, the global system prompt is prepended to each scene-level prompt, so that all images share a common artistic look and consistent character depiction while still reflecting the unique content of each scene. These combined prompts are then passed to Google's Imagen-4.0-Ultra model to create the final images, yielding story-wise sequences that aim to stay faithful to the text and visually coherent across all frames.

The **team_meoooo** [32] designs a multilingual story-to-image pipeline that treats Hindi and English stories within a single, unified framework by combining translation, text segmentation, summarisation, and diffusion-based image generation. Hindi stories are first translated into English using the pretrained `facebook/nllb-200-distilled-600M` model, while English stories are used directly. The resulting texts are cleaned to remove extra spaces, line breaks, and other formatting artefacts. The cleaned story is then split into sentences and organised into smaller "sub-stories". To do this, each sentence is represented with a TF–IDF vector, cosine similarity scores are computed between sentences, and sentences with high similarity are grouped together, with low similarity marking natural breaks in the narrative. Because summarisation models have input length limits, these initial groups are further adjusted based on token counts: each chunk is tokenised using NLTK's `word_tokenize`, tokens are redistributed so that segment lengths remain reasonably balanced, and the final segments are checked against a 512-token threshold using the `facebook/bart-large-cnn` tokenizer, with only occasional truncation of very long inputs. Each of these balanced segments is then summarised using the `facebook/bart-large-cnn` abstractive model to produce short descriptions that highlight the central events and characters in that part of the story. These summaries serve as prompts for the `stable-diffusion-X1-base-10` model, which generates one image per segment, aiming to produce a sequence of illustrations that both follow the narrative flow and maintain a coherent visual style across the entire story.

5. Results

The evaluation process followed a structured human-annotation protocol designed to capture both visual and narrative fidelity. Each system's outputs were examined by trained annotators who assessed every image independently along three dimensions. Visual Quality measured clarity, realism, absence of distortions, and overall aesthetic coherence. Relevance captured how well the image reflected the textual description of the specific story segment, focusing on attributes such as depicted actions, objects, characters, and contextual cues. Consistency evaluated whether images belonging to the same story maintained stable character appearances, settings, and thematic progression. All judgments were made using a uniform three-point scale i.e., Good, Moderate, and Fair providing interpretable and comparable cross-team performance indicators. This evaluation framework ensured that systems were not rewarded solely for generating attractive images but were also assessed on their ability to respect narrative structure, linguistic cues, and multi-panel continuity.

Table 3
English Results

Team Name	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
Retriever-run1	39	0	0	34	5	0	24	14	1
NandiniDivya-run1	36	2	0	5	21	12	8	18	12
Team_meoooo-run1	15	22	2	3	10	26	3	8	28
NLPFusion-run3	0	0	39	0	6	33	0	2	37
NLPFusion-run1	0	1	38	0	1	38	0	1	38
JU Team MCSE-l_run1	0	0	39	0	2	37	0	0	39
NLPFusion-run4	0	1	38	0	1	38	0	0	39
NLPFusion-run2	0	1	21	0	0	22	0	0	22

Human evaluation scores for all systems submitted on English and Hindi stories are shown in Table 3

Table 4
Hindi Results

Team Name	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
Retriever-run1	30	0	0	30	0	0	27	3	0
Nandini_Divya-run1	30	0	0	10	15	5	15	10	5
Team_meoooo-run1	19	11	0	0	0	30	0	3	27
NLPFusion-run3	0	0	30	2	4	24	0	2	28
NLPFusion-run1	0	0	30	0	2	28	0	0	30
NLPFusion-run4	0	0	30	0	1	29	0	0	30
NLPFusion-run2	0	0	21	0	0	21	0	0	21

and Table 4. In English, *Retriever-run1* clearly excels. All 39 outputs receive a Good rating for Visual Quality. Its Relevance scores are also strong, with 34 Good and 5 Moderate ratings. Consistency is a bit more challenging; 24 outputs are rated Good, 14 Moderate, and just 1 Fair. This indicates that this system generally preserves story flow but occasionally loses track between panels. Team *NandiniDivya-run1* also achieves high Visual Quality with 36 Good and 2 Moderate ratings. However, its Relevance and Consistency scores are more evenly spread between Moderate and Fair. This suggests that while the images are visually appealing, they do not always connect tightly with the narrative. *Team_meoooo-run1* is in the middle: most images are rated Moderate rather than Good for Visual Quality (22 v/s 15). Relevance and Consistency are mostly rated Fair, indicating challenges in capturing nuances of the story. The remaining systems, including *NLPFusion* runs and *JU Team MCSE-1_run1*, primarily receive Fair ratings across all three criteria, suggesting they struggle with both basic visual quality and narrative alignment in English stories.

The Hindi evaluations show a similar pattern. *Retriever-run1* again performs best, with all 30 outputs rated Good for both Visual Quality and Relevance. Consistency scores are slightly lower but still strong, with 27 Good and 3 Moderate ratings. Team *NandiniDivya-run1* matches *Retriever-run1* in Visual Quality with 30 Good ratings. However, its other two criteria have a more varied profile: Relevance rates as Good, Moderate, and Fair (10/15/5), and Consistency shows the same 15/10/5 distribution. This reflects the English evaluations, where images are visually appealing, but sometimes loosely connect to specific story sections. For *Team_meoooo-run1*, Visual Quality is mostly Good or Moderate (19 and 11 ratings, respectively). Yet, Relevance and Consistency are mostly rated Fair. This means the system often produces decent-looking images that do not closely match the Hindi text. The NLPFusion systems again rank lower, with most outputs receiving Fair ratings across all metrics.

Looking across teams and languages, three key trends emerge. First, visual quality appears to be the easiest dimension to satisfy. Even comparatively weaker systems are often capable of producing or retrieving images that are sharp, visually coherent, and free from obvious artifacts such as blurriness, distortions, or malformed objects. In contrast, Relevance and especially Consistency are much harder to meet since they require the model to focus on specific entities, actions, and story progression. Second, *Retriever-run1* stands out as the top-performing system overall and is also the most consistent across both English and Hindi. Third, systems like *NLPFusion* and *Team_meoooo-run1* reveal an important gap: they can create visually appealing outputs but struggle to maintain a strong connection to the text and ensure characters and scenes stay consistent across multiple panels. This gap is what MUSIA aims to highlight, distinguishing systems that produce merely ”nice” images from those that effectively follow and understand the story’s context.

6. Conclusion

The MUSIA-2025 shared task highlights the growing need for generative systems that go beyond creating visually appealing images. These systems should focus on narrative accuracy and cultural relevance. Eight teams registered at first, but only five submitted valid system runs, and just four completed

their final papers. Overall, the systems showcased impressive visual quality, mainly due to strong diffusion models and LLM-guided prompt engineering. However, they often fell short in narrative relevance and consistency across panels, revealing ongoing issues in today's multimodal generation setups. A significant finding is that none of the teams used the culturally rich training illustrations from the dataset. As a result, most generated images reflected Western stylistic biases. This points to a broader issue with current text-to-image models, which are mainly trained on Western-centric image collections and often struggle to depict Indian cultural elements, including traditional attire, regional settings, indigenous motifs, and local storytelling practices. These observations reinforce MUSIA's main goal: to promote systems that not only understand multilingual story narratives but also create visuals that truly represent cultural contexts. Future methods should incorporate the provided illustrations using techniques like retrieval-augmented prompting, cultural style transfer, or prototype-based learning to reduce bias. Expanding MUSIA to include more Indian languages and diverse narrative traditions could further enhance story structures and visual diversity. Additionally, developing automated metrics for cultural accuracy would support human evaluations of relevance, consistency, and quality. Promising paths include multimodal fine-tuning using Indian illustration datasets and memory-enhanced structures to maintain character identity across panels. Overall, MUSIA-2025 shows both the potential and current limitations of multimodal generation systems, stressing the need for models that are more culturally aware and context-sensitive in creative and educational settings.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao, Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 993–1002.

[2] A. Maharana, D. Hannan, M. Bansal, Improving Generation and Evaluation of Visual Stories via Semantic Consistency, in: NAACL, 2021, pp. 2427–2442.

[3] J. Hu, Y. Cheng, Z. Gan, J. Liu, J. Gao, G. Neubig, Visual storytelling dataset (vist), https://service.tib.eu/ldmservice/dataset/visual-storytelling-dataset--vist-, 2024. Accessed: 2025-10-25.

[4] G. Zeng, Z. Li, Y. Zhang, Pororogan: An improved story visualization model on pororo-sv dataset, Proceedings of the 3rd International Conference on Computer Science and Artificial Intelligence (2019) 1–5. URL: https://dl.acm.org/doi/10.1145/3374587.3374649. doi:10.1145/3374587.3374649.

[5] T. Gupta, D. Schwenk, A. Farhadi, D. Hoiem, A. Kembhavi, Imagine This! Scripts to Compositions to Videos, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2018, pp. 610–626. URL: https://link.springer.com/chapter/10.1007/978-3-030-01237-3_37. doi:10.1007/978-3-030-01237-3_37.

[6] Z. Ye, J. Liu, R. Peng, J. Cao, Z. Chen, Y. Zhang, Z. Xuan, M. Zhou, X. Shen, M. Elhoseiny, Q. Liu, G.-J. Qi, Openstory++: A large-scale dataset and benchmark for instance-aware open-domain visual storytelling, arXiv preprint arXiv:2408.03695 (2024). URL: https://arxiv.org/abs/2408.03695.

[7] M. Zhu, P. Pan, W. Chen, Y. Yang, Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis, 2019. URL: https://arxiv.org/abs/1904.01310. arXiv:1904.01310.

[8] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, Y. Taigman, Make-a-scene: Scene-based text-to-image generation with human priors, 2022. URL: https://arxiv.org/abs/2203.13131. arXiv:2203.13131.

[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, 2021. URL: https://arxiv.org/abs/2102.12092. arXiv:2102.12092.

[10] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, Y. Wu, Scaling autoregressive models for content-rich text-to-image generation, 2022. URL: https://arxiv.org/abs/2206.10789. arXiv:2206.10789.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, 2021.

[13] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, 2021. URL: https://arxiv.org/abs/2105.05233. arXiv:2105.05233.

[14] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. URL: https://arxiv.org/abs/2112.10741. arXiv:2112.10741.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. URL: https://arxiv.org/abs/2112.10752. arXiv:2112.10752.

[16] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL: https://arxiv.org/abs/2205.11487. arXiv:2205.11487.

[17] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, B. Guo, Vector quantized diffusion model for text-to-image synthesis, 2022. URL: https://arxiv.org/abs/2111.14822. arXiv:2111.14822.

[18] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, J. Gao, Storygan: A sequential conditional GAN for story visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, p. 9–17. Uses the Pororo-SV dataset for evaluation.

[19] A. Maharana, D. Hannan, M. Bansal, Storydall-e: Adapting pretrained text-to-image transformers for story continuation, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2022, pp. 70–87. URL: https://arxiv.org/abs/2209.06192. doi:10.1007/978-3-031-19826-4_5.

[20] A. Maharana, D. Hannan, M. Bansal, Storydall-e: Adapting pretrained text-to-image transformers for story continuation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13697–13706.

[21] Z. Song, Z. Zhang, Z. Li, Y. Zhang, Character preserving coherent story visualization via visual planning and token alignment, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 1234–1243.

[22] T. Gupta, A. Gupta, M. He, A. G. Bansal, Imagine this! scripts to compositions to videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 1–10. URL: https://arxiv.org/abs/1804.03608. doi:10.1109/CVPR.2018.00001.

[23] J. Kapuriya, P. Buitelaar, 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*, 2025. URL: https://ceur-ws.org/Vol-3964/paper3.pdf, accessed: 2025-10-25.

[24] R. Oliveira, L. Santos, F. Silva, R. Lima, E. Costa, Narrative generation from visual inputs: A framework for storytelling from images, in: Proceedings of the 2021 Annual Meeting of the Association for Computational Linguistics, 2021, pp. 567–577.

[25] J. Hong, M. Kim, S. Lee, Visual coherence losses for story generation from images, in: Proceedings

of the 2022 Conference on Neural Information Processing Systems, 2022, pp. 2345–2356.
[26] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, Advances in Neural Information Processing Systems 35 (2022).
[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, P. Nakkiran, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020).
[28] Y. Zhang, J. Li, Y. Chen, W. Xu, X. Wang, Instruction-tuned multimodal language models for storytelling, in: Proceedings of the 2023 International Conference on Learning Representations, 2023, pp. 3456–3467.
[29] N. S. Sharma, Divya, Multilingual Story Illustration for MUSIA 2025 using One-Prompt-One-Story Image Generation, in: FIRE 2025 Working Notes, CEUR Workshop Proceedings, 2025.
[30] S. Mannan, A. Hegde, S. Coelho, Bridging Cultures through AI: The Art of Multilingual Storytelling, in: FIRE 2025 Working Notes, CEUR Workshop Proceedings, 2025.
[31] K. Kachhadiya, P. Patel, Leveraging Large Language Model(LLM) and V-LLM for Zero-Shot Multilingual Story Illustration, in: FIRE 2025 Working Notes, CEUR Workshop Proceedings, 2025.
[32] M. Sadhukhan, I. Bhattacharya, P. Dutta, NarrArt: Multilingual Story Illustration with AI for English and Hindi Narratives, in: FIRE 2025 Working Notes, CEUR Workshop Proceedings, 2025.