Multilingual Story Illustration for MUSIA 2025 using One-Prompt-One-Story Image Generation

Sharma Nandini Surendra[1,*], Divya[1,*]

[1]Indian Institute of Technology (BHU), Varanasi, INDIA

Abstract

This working note presents our submission for the MUSIA 2025 Shared Task on multilingual story illustration. The task required generating a series of culturachlly relevant and narratively coherent illustrations for stories written in English and Hindi. To address this, we adopted the recently proposed *One-Prompt-One-Story* framework, implemented using Playground AI's generative image model. Our approach involved converting each story into a comma-separated sequence of prompts, where each segment corresponded to one illustration frame. For the English dataset, we used LLM to summarize and segment the stories into the required number of prompts. For the Hindi dataset, we first translated the stories into English using LLM and then applied the same segmentation process. The prompts were then fed into the image generation model, producing one illustration per narrative segment. The method is simple, training-free, and language-agnostic, yet effective in maintaining narrative flow across frames. Our results demonstrate that this approach produces visually consistent and contextually relevant illustrations, providing a practical baseline for multilingual story visualization.

Keywords

Story Illustration, Generative AI, Multilingual Processing, One-Prompt-One-Story, Image Generation

1. Introduction

This paper presents our participation in the MUSIA 2025 Shared Task [1, 2], which focuses on generating visual illustrations for multilingual stories, particularly in English and Hindi. The central aim of this task is to advance narrative-driven text-to-image (T2I) generation by producing coherent, culturally relevant, and visually engaging illustrations that complement the flow of a story. Such a capability has wide-ranging applications in areas such as education, where illustrated stories can enhance language learning and comprehension, and in the creative industries, where automated visual storytelling can accelerate content production for comics, animations, and other narrative media.

A key challenge in this domain lies in balancing richness of visual expression with consistency across characters and scenes. Unlike single-image generation, story illustration requires that recurring entities—such as protagonists or symbolic characters—retain their identity throughout the narrative. This problem is amplified in multilingual and culturally diverse stories, such as those drawn from folklore or the Panchatantra tradition, where characters often serve as anchors for moral or cultural meaning. Current T2I models often struggle with this requirement, producing inconsistencies in character appearance across different story segments. Previous approaches have attempted to mitigate this through fine-tuning, large-scale identity encoders, or complex module designs, but these solutions tend to be computationally expensive and are not easily adaptable to multilingual settings.

To address this challenge, we adopt the recently proposed One-Prompt-One-Story (1Prompt1Story) framework as the backbone of our system for MUSIA. This approach leverages the inherent ability of language models to maintain contextual consistency across longer prompts, thereby reducing the need for fine-tuning or additional training. By consolidating identity descriptions and narrative events into a single extended prompt, and applying techniques such as Singular-Value Reweighting and Identity-Preserving Cross-Attention, 1Prompt1Story achieves both strong subject consistency and reliable alignment with individual narrative segments. This makes it particularly well suited for the

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

*Corresponding author.

sharma.nandini.cse22@itbhu.ac.in (S. N. Surendra); divya.student.cse22@itbhu.ac.in (Divya)

MUSIA task, where multilingual stories must be visually illustrated without compromising narrative fidelity.

In this paper, we describe our application of the 1Prompt1Story model to the MUSIA 2025 Shared Task. We discuss how we structured story prompts, handled multilingual input, and generated illustrations that balance narrative relevance, visual quality, and cultural appropriateness. Through this work, we aim to highlight the effectiveness of prompt consolidation strategies for consistent and coherent story illustration, and demonstrate their potential for broader applications in multilingual narrative generation.

2. Related Work

Research on turning multi-sentence narratives into coherent sequences of images spans several directions, and the literature can be grouped roughly into (a) adversarial, task-specific story-visualization models, (b) methods that adapt large pretrained text-to-image models to sequential generation, and (c) recent efforts that explicitly target character/identity consistency across frames. Below we summarize representative works in each category and contrast them with the prompt-driven, training-free strategy we adopt.

Early work in story visualization built on conditional GANs and introduced architectures tailored to the sequential nature of the task. StoryGAN and its variants established a baseline by conditioning per-sentence image synthesis on learned story representations and by using recurrent or memory modules to inject temporal context into the generator [3]. Subsequent GAN-based improvements focused on architectural refinements to improve both image quality and narrative adherence: for example, Improved-StoryGAN integrates dilated convolutions, gated convolutions, and a Weighted Activation Degree (WAD) mechanism to expand receptive fields and emphasize sentence–image alignment [4], while PororoGAN augments the pipeline with an aligned sentence encoder (ASE), an attentional word encoder (AWE), and a patch-level discriminator to strengthen global and local relevance on cartoon datasets [5]. These approaches demonstrate that careful encoder–decoder design and attention to multi-scale realism can boost both fidelity and coherence, but they typically require substantial task-specific training data and bespoke losses.

A second strand of work explores how to leverage large pretrained text-to-image models for story tasks rather than training from scratch. StoryDALL-E [6] is an illustrative example: it "retro-fits" a pretrained T2I transformer with modules for sequential generation and copying elements from a source image, and explores both full fine-tuning and parameter-efficient prompt tuning. This line of research shows that pretrained generators bring strong visual priors and can be adapted to low-resource story tasks, but adapting them effectively often requires additional modules or fine-tuning to preserve continuity and character identity across frames.

More recent methods explicitly target character- and identity-preservation as core objectives. Character-Preserving Coherent Story Visualization (CP-CSV) [7] combines story/context encoders, figure–ground segmentation as an auxiliary task, and figure–ground–aware generation to better maintain character details and scene consistency; it also introduces evaluation metrics (e.g., Frechet Story Distance) tailored to the sequential setting. These works underscore two important lessons: (1) preserving character-specific attributes often benefits from explicit representation or auxiliary supervision, and (2) evaluation for story visualization needs to measure not only single-image quality but also cross-frame coherence.

The One-Prompt-One-Story (1Prompt1Story) approach departs from the above trends by exploiting an orthogonal insight: language models naturally encode contextual identity information when multiple frame descriptions are consolidated into a single prompt [8]. Rather than relying on additional training, dedicated encoders, or heavy fine-tuning, 1Prompt1Story uses prompt consolidation together with prompt-embedding reweighting and attention-layer interventions (Singular-Value Reweighting and Identity-Preserving Cross-Attention) to produce consistent multi-frame illustrations in a training-free manner. In contrast to GAN-based pipelines that require dataset-specific training and to retrofit approaches that fine-tune large models, 1Prompt1Story aims for a practical middle path—preserving subject

identity across frames through prompt engineering and light-weight, model-agnostic modifications—making it particularly attractive for multilingual, low-resource, or culturally diverse story illustration tasks such as MUSIA.

In summary, prior work has advanced story visualization through architectural innovation, encoder/attention designs, and adaptation of large pretrained models, while more recent work emphasizes explicit character preservation and specialized evaluation. Our choice to build on the 1Prompt1Story paradigm is motivated by its training-free character-consistency mechanism and its suitability for the MUSIA setting, where multilingual and culturally specific narratives demand a practical, adaptable solution that can maintain identity and narrative alignment without extensive retraining.

3. Dataset

For this shared task, we were provided with a multilingual dataset consisting of narrative stories and corresponding illustrations in two languages: English and Hindi. The dataset is organized into training and testing sets, each following a systematic naming convention for both stories and images.

3.1. Training Data

The training data includes 360 English stories and 185 Hindi stories. Each story is given as a plain text file inside the `Stories` folder, with the following naming convention:

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

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

For each story, one or more manually created illustrations are provided in the `Images` folder, named as:

- `eng_story_XXXX_01`, `eng_story_XXXX_02`, ...
- `hin_story_XXXX_01`, `hin_story_XXXX_02`, ...

The images are shared in standard formats such as `.jpg`, `.jpeg`, and `.png`. This setup aligns each narrative with a sequence of illustrations, enabling models to learn how to map text to visuals in a consistent, story-driven manner.

3.2. Testing Data

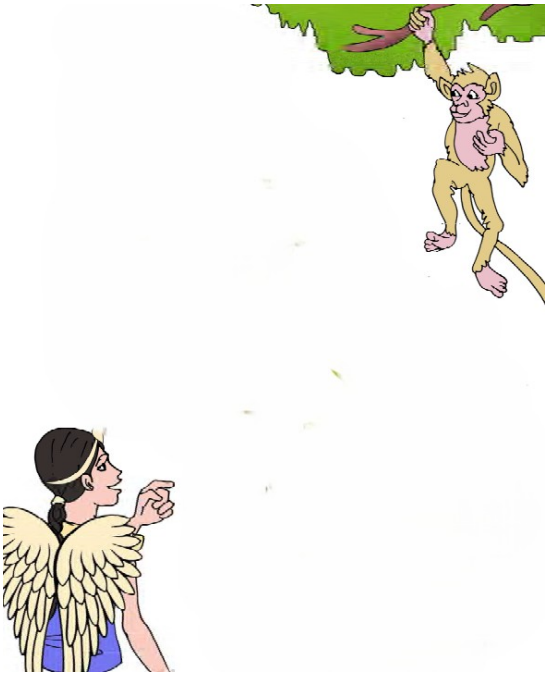
The testing set consists of 40 English stories and 30 Hindi stories. Similar to the training set, the stories are provided as text files, but without ground-truth illustrations. Instead, a mapping file specifies the required number of images to generate for each story. Participants were expected to generate the specified number of illustrations per story, following the same file-naming conventions as the training data (e.g., `hin_story_0021_01`).

The outputs were to be organized into separate folders for English and Hindi and submitted as a single archive.

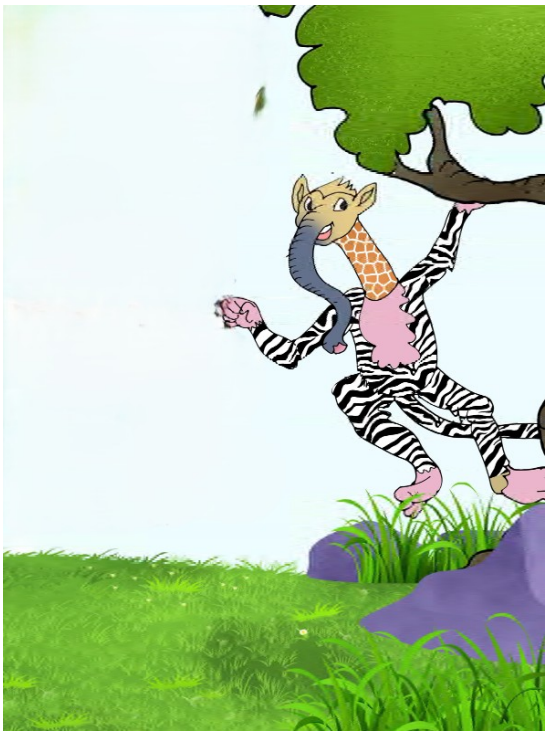
3.3. Example Illustration

To help participants understand the task, the organizers also provided an example illustration. One such story is about a character called *Little Monkey*, who wishes to be big and strong. A wise woman gives him a magic wand, and his wishes begin to come true, transforming him with the features of other animals he admires. In the second illustration, he goes to the river, sees his reflection, and is horrified by his appearance. His mother reminds him that he wished for those features, after which he longs to return to his original self.

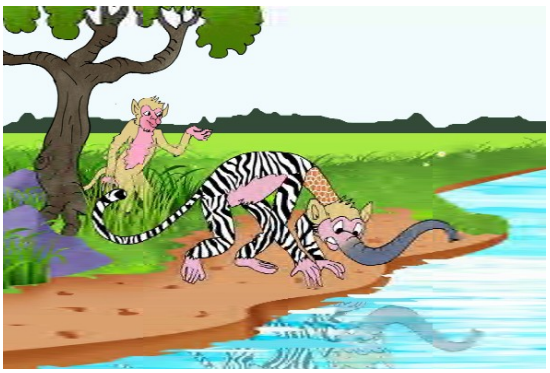
This example demonstrates the type of narrative flow and visual consistency expected in the generated illustrations, where characters evolve across multiple scenes while remaining identifiable.



(a) “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.”



(b) 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.



(c) 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.

Figure 1: Example sequence of generated illustrations from an English story (eng_story_1252).

3.4. Summary

Table 1 summarizes the dataset statistics.

Language	Training Stories	Test Stories
English	360	40
Hindi	185	30

Table 1
Dataset statistics for MUSIA 2025 Shared Task.

3.5. Workflow

Our workflow for generating story illustrations consists of four main stages: dataset processing, prompt preparation, illustration generation with the One-Prompt-One-Story (1Prompt1Story) framework[8], and final rendering using a diffusion-based model[9]. In what follows, we describe each stage in detail and explain the rationale for adopting the 1Prompt1Story paradigm.

3.5.1. Dataset Processing

The MUSIA 2025 organizers provided a multilingual dataset comprising stories in English and Hindi. Since our illustration pipeline required a consistent prompt format, we processed these languages separately. For English stories, we directly utilized a Large Language Model (LLM) to summarize and segment the narrative into a predefined number of frames. For Hindi stories, we first translated the text into English using an LLM, ensuring semantic fidelity, and then applied the same segmentation process. This approach allowed us to standardize the inputs, making the subsequent stages language-agnostic.

3.5.2. Prompt Preparation

A key requirement of the 1Prompt1Story framework [8] is that the input must be a single concatenated prompt, where each frame of the story is represented as a comma-separated segment. Unlike conventional methods that issue separate prompts for each illustration, 1Prompt1Story leverages the global context encoded in a single prompt to enforce consistency across all frames.

To meet this requirement, we designed an LLM-based pipeline. Given the number of illustrations required for a story, we instructed the LLM to divide the narrative into that exact number of concise, descriptive segments. These segments were then concatenated into a single comma-separated prompt, with each segment corresponding to one illustration frame. This ensured that the global story context was preserved while still providing sufficient detail for each individual frame.

3.5.3. Illustration Generation with One-Prompt-One-Story

Traditional story illustration approaches often suffer from inconsistencies when generating multiple images from separate prompts. For example, the appearance of a main character may drift from one illustration to the next, or background details may fail to remain coherent across frames. Fine-tuning or specialized architectures can mitigate these issues, but they are computationally expensive and often impractical in a competition setting.

The 1Prompt1Story framework provides an elegant, training-free solution to this problem. By encoding all frame-level descriptions into a single prompt, the framework exploits the diffusion model's inherent capacity to maintain semantic coherence across scenes. It further refines the generation process through two mechanisms: (i) Singular-Value Reweighting (SVR), which amplifies dominant identity-preserving features and prevents character drift, and (ii) Identity-Preserving Cross-Attention (IPCA), which strengthens alignment between the global narrative and local scene descriptions. Together, these

techniques ensure that characters remain visually consistent and story-relevant across all illustrations, while still allowing natural variation in backgrounds, actions, and perspectives.

3.5.4. Final Rendering

The consolidated prompt, structured according to the 1Prompt1Story requirements, was then passed to Playground AI's[9] diffusion-based model. Each frame was generated sequentially, with the shared global prompt acting as a unifying context. This setup yielded illustrations that were not only stylistically coherent but also semantically faithful to the original story narrative. The resulting images demonstrated strong character fidelity, thematic consistency, and narrative alignment—qualities that are often difficult to achieve with multi-prompt approaches.

3.5.5. Why One-Prompt-One-Story?

The decision to adopt 1Prompt1Story was motivated by both theoretical and practical considerations. First, it eliminates the need for additional training or fine-tuning, making it highly efficient and accessible. Second, it provides a principled way to maintain cross-frame consistency, a challenge that traditional prompt-per-frame pipelines often struggle with. Third, the framework aligns naturally with the storytelling task, since narratives are inherently global in scope, with local events contextualized by the broader storyline. By embedding the entire narrative into a single structured prompt, 1Prompt1Story enables diffusion models to capture this global coherence, resulting in illustrations that are faithful not just to individual sentences, but to the story as a whole.

4. Results

Our system successfully generated illustrations for both English and Hindi datasets. The outputs were coherent with the story flow, visually aligned with narrative events, and culturally appropriate in representation. We observed that:

- Story segmentation via LLM produced concise and effective prompts.
- Translation from Hindi to English preserved most narrative elements, though minor nuances were occasionally lost.
- The One-Prompt-One-Story framework produced consistent visual identities across frames without requiring explicit fine-tuning.

4.1. Qualitative Results

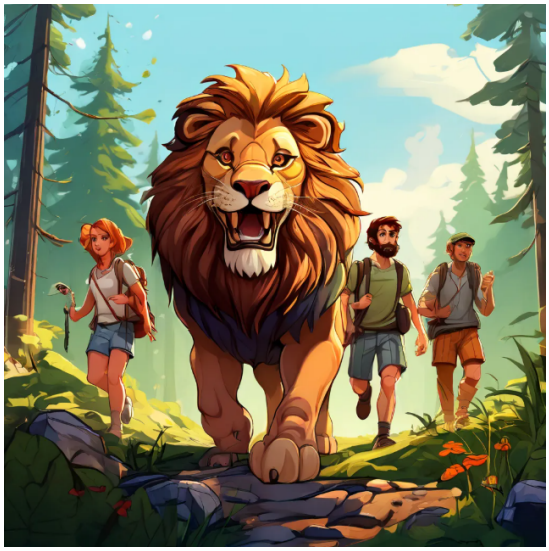
Figure 2 shows an example from the English dataset where the generated sequence of illustrations captures key narrative events. The model maintained a consistent depiction of characters (e.g., the lion) across frames while adapting to changes in environment and action.

4.2. Quantitative Evaluation

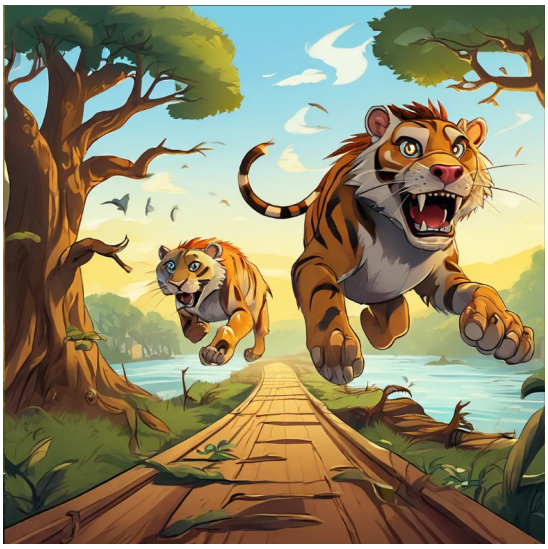
The generated outputs were evaluated using three criteria provided by the MUSIA 2025 organizers:

- **Consistency:** Maintaining identity and style across multiple frames (highest weight).
- **Relevance:** Faithfulness of illustrations to the narrative descriptions.
- **Visual Quality:** Overall perceptual quality and clarity of the generated images.

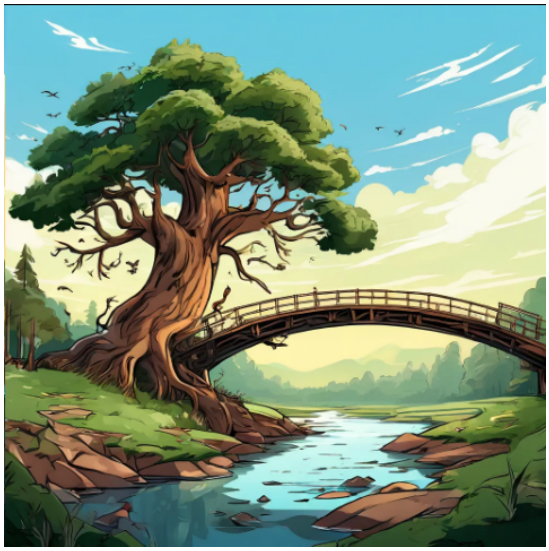
For each criterion, ratings were assigned as *Good*, *Moderate*, or *Fair*. The tables below present the distribution of stories across these categories for both English and Hindi test sets.



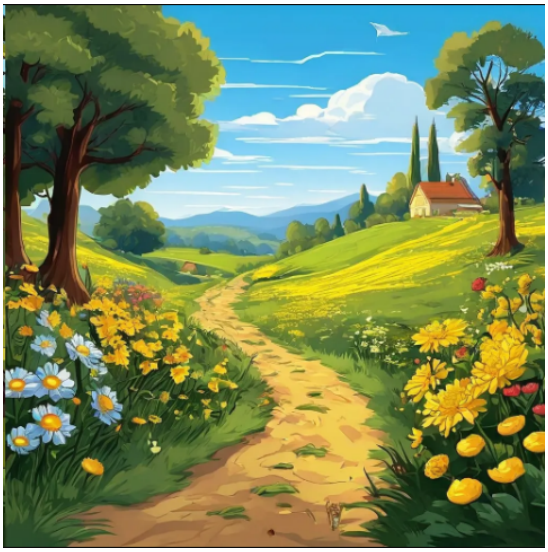
(a) Prompt 1: "A group of travelers and a lion walking through a dense forest."



(b) Prompt 2: "They found a thick dark and gloomy forest which had monsters which looked like tigers."



(c) Prompt 3: "They came across a wide and deep gulf and thought to cut down a tree and the Lion pushed it on the gulf to make a bridge."



(d) Prompt 4: "They finally exited the dense forest and reached an open valley with bright flowers and blue sky."

Figure 2: Example sequence of generated illustrations from an English story (eng_story_1252).

Table 2
Evaluation results for English stories (39 total).

Criterion	Good	Moderate	Fair
Visual Quality	36	2	0
Relevance	5	21	12
Consistency	8	18	12

4.3. Observations

- Visual quality was consistently high across both languages, demonstrating the effectiveness of the diffusion-based generation.

Table 3
Evaluation results for Hindi stories (30 total).

Criterion	Good	Moderate	Fair
Visual Quality	30	0	0
Relevance	10	15	5
Consistency	15	10	5

- Relevance varied more widely, particularly in English, where detailed narrative elements were sometimes oversimplified in prompts.
- Consistency improved with the One-Prompt-One-Story framework but remains a challenge in complex multi-character scenes.

5. Conclusion

In this work, we presented our approach for the MUSIA 2025 Shared Task on multilingual story illustration. Our system combined large language models for summarization, prompt segmentation, and translation with a diffusion-based text-to-image generator to produce coherent illustrations from narrative text. The results on both English and Hindi datasets demonstrate that our pipeline can generate visually appealing and narratively aligned sequences of images without additional model training. Quantitative evaluation showed strong performance in terms of visual quality, with moderate effectiveness in relevance and consistency, highlighting the inherent challenges of capturing fine-grained narrative details and preserving character identity across multiple frames.

While our method benefits from its simplicity and training-free design, there remains room for improvement in handling nuanced cultural context, maintaining long-term consistency in multi-character stories, and improving relevance to complex narrative descriptions. Future work may involve fine-tuning the summarization and prompt segmentation modules on the provided training data, as well as exploring advanced frameworks such as the One-Prompt-One-Story paradigm for more consistent identity preservation. We believe that combining these directions with multilingual adaptation will further enhance the quality and reliability of automatic story illustration systems.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Perplexity AI 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] K. Tewari, A. Malviya, S. Chanda, A. Mukherjee, S. Pal, Findings 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, CEUR Working Notes, 2026.

[2] K. Tewari, A. Malviya, S. Chanda, A. Mukherjee, S. Pal, 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, Association for Computing Machinery, New York, NY, USA, 2026.

[3] Y. Li, et al., Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[4] C. Li, L. Kong, Z. Zhou, Improved-storygan for sequential images visualization, Journal of Visual Communication and Image Representation 73 (2020) 102956. URL: https://www.sciencedirect.

com/science/article/pii/S1047320320301826. doi:https://doi.org/10.1016/j.jvcir.2020.102956.

[5] G. Zeng, Z. Li, Y. Zhang, Pororogan: An improved story visualization model on pororo-sv dataset, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2020.

[6] A. Maharana, D. Hannan, M. Bansal, Storydall-e: Adapting pretrained text-to-image transformers for story continuation, arXiv preprint arXiv:2209.06192 (2022).

[7] Y.-Z. Song, Z.-R. Tam, H.-J. Chen, H.-H. Lu, H.-H. Shuai, Character-preserving coherent story visualization, in: Proceedings of the ACM International Conference on Multimedia (ACM MM), 2020.

[8] T. Liu, K. Wang, S. Li, et al., One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt, in: Proceedings of the International Conference on Learning Representations (ICLR), 2025.

[9] B. Liu, E. Akhgari, A. Visheratin, A. Kamko, L. Xu, S. Shrirao, C. Lambert, J. Souza, S. Doshi, D. Li, Playground v3: Improving text-to-image alignment with deep-fusion large language models, arXiv preprint arXiv:2503.12345 (2025).