

Leveraging Large Language Model(LLM) and V-LLM for Zero-Shot Multilingual Story Illustration

Kishan Kachhadiya^{1,*†}, Parth Patel^{1,†}

¹Dhirubhai Ambani University, Gandhinagar, Gujarat, India

Abstract

Story illustration is a growing field that blends textual narratives with visual media to produce engaging and culturally enriched story imagery. However, achieving coherent and visually consistent illustrations remains a significant challenge due to language diversity, variations in story structure, and the requirement for maintaining character continuity. In this work, we utilize Gemini-2.5-Pro to generate adaptive and detailed prompts for each segment of a story, which are subsequently employed with Imagen-4.0-Ultra to create high-quality illustrations. Our approach is designed for multilingual settings, preserves character consistency across multiple images, and dynamically adapts prompt generation to the required number of illustrations per story. Human evaluations conducted on English and Hindi story datasets—assessing metrics such as consistency, relevance, and visual quality—demonstrate that our method markedly enhances the alignment between narrative content and generated visuals. These findings underscore the effectiveness of integrating LLM-driven prompt engineering with state-of-the-art image synthesis for producing coherent, culturally resonant, and visually compelling story illustrations.

Keywords

Multilingual Story Illustration, Large Language Models, Prompt Engineering, Image Generation, Vision-Large Language Models

1. Introduction

Visual storytelling is a compelling medium that combines textual narratives and visual illustrations to convey engaging and culturally rich stories. The task of generating story illustrations from textual narratives has gained significant attention in recent years due to its potential applications in areas such as children’s education, entertainment, and digital content creation. The ability to automatically generate coherent and contextually relevant illustrations from multilingual story texts can revolutionize how stories are presented and consumed, making it an important research area in artificial intelligence and computer vision.

Despite its potential, the task of multilingual story illustration presents several challenges. One primary challenge is maintaining **character and scene consistency** across multiple images, which involves ensuring that the same characters retain visual continuity and recognizable traits throughout a story. Additionally, generating illustrations that accurately reflect the narrative while being visually appealing and culturally appropriate adds another layer of complexity. Existing models often struggle with these aspects due to limited task-specific datasets and the inherent difficulty of aligning textual and visual information effectively [2, 3].

The motivation for this research stems from the need to address these challenges by leveraging the capabilities of Large Language Models (LLMs) combined with advanced image generation models. In particular, we employ *Gemini-2.5-Pro* [4] to generate structured and detailed prompts for each segment of a story, which are then used with *Imagen-4.0-Ultra* [5] to synthesize high-quality illustrations. This

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

*Corresponding author.

†These authors contributed equally.

Official Page: [https://cse-iitbhu.github.io/MUSIA/\[1\]](https://cse-iitbhu.github.io/MUSIA/[1])

Dataset details:<https://cse-iitbhu.github.io/MUSIA/dataset.html>

✉ kishan_kachhadiya@dau.ac.in (K. Kachhadiya); 202411047@dau.ac.in (P. Patel)

>ID 0009-0009-5362-7302 (K. Kachhadiya); 0009-0002-1589-0221 (P. Patel)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

methodology allows us to adapt prompt generation to the number of images per story, maintain character continuity, and ensure that the visual output aligns closely with the story narrative, even across multiple languages, such as English and Hindi.

We utilized the dataset provided by the official MUSIA shared task.[1] where each story is annotated with the number of images to be generated. The LLM is tasked with producing prompts that are rich in visual detail and style guidance, ensuring that the generated images capture key narrative elements, character traits, actions, and emotions. To enhance visual coherence and thematic consistency, we incorporate instructions for style, color, shading, texture, framing, and character continuity. The illustrations are then generated sequentially using the Imagen model, producing high-quality story images while preserving character and scene consistency.

Human evaluation was conducted using a combination of task-specific metrics, including image consistency, relevance, and visual quality. Quantitative results on both English and Hindi stories indicate that our method achieves high scores in consistency and relevance, demonstrating substantial improvements over baseline approaches in multilingual story illustration.

In this paper, we focus on the application and evaluation of existing LLMs and advanced image generation techniques for multilingual story illustration. Specifically, we:

- We utilize *Gemini-2.5-Pro* to generate detailed and structured visual prompts for both English and Hindi story texts, leveraging the model’s strong multilingual comprehension ability.
- Use *Imagen-4.0-Ultra* to generate high-quality illustrations based on the prompts, while maintaining character and scene consistency across multiple story images.
- Analyze and evaluate the results on English and Hindi story datasets using metrics such as image Consistency, Relevance, and Visual Quality, providing insights into the effectiveness of applying these models to the MUSIA task.

2. Literature Review

Visual storytelling has attracted significant attention in recent years, particularly with the rise of multimodal learning frameworks that integrate visual and textual information. Oliveira et al. [2] proposed a framework for story generation from visual inputs, leveraging paired image-caption datasets to train models capable of producing coherent narrative sequences. Using their dataset, the proposed models achieved substantial improvements in BLEU and METEOR scores over baseline captioning methods, highlighting the importance of structured narrative supervision.

In terms of task-specific approaches, Hong et al. [3] introduced visual coherence loss functions to explicitly maintain consistency in character appearances and scene layouts across multiple frames. Their framework was evaluated on annotated visual storytelling datasets, showing an increase in narrative alignment metrics and human-rated story quality. These methods highlight the necessity of addressing both semantic and visual consistency when generating extended narratives.

Recent advances in large vision-language models (L-VLMs) have further enhanced visual story generation. Zhang et al. [6] explored multimodal LLMs fine-tuned with instruction tuning for generating contextually rich stories from image streams. Their experiments utilized diverse datasets, including illustrated storybooks and comics, achieving significant improvements in narrative coherence and emotional alignment as evaluated by GPT-4 and human assessments. This work demonstrates the potential of instruction-guided LLMs to capture complex relationships across sequential visual inputs, a principle that underpins our methodology.

Our methodology builds upon these foundations by applying LLMs and LVLMs in a task-oriented framework. We utilize curated image sequences and structured instruction tasks to guide the model’s narrative generation, drawing inspiration from the instruction-tuning strategy reported in prior works. Unlike previous studies, we do not propose novel architectures but instead demonstrate the effectiveness of applying these advanced techniques to a structured multilingual visual storytelling task, assessing performance using established evaluation metrics such as image coherence, relevance, and visual consistency.

3. Methodology

The primary objective of this work is to develop an automated and scalable approach for generating high-quality visual illustrations from multilingual story texts, specifically targeting both English and Hindi narratives. By integrating LLMs for prompt generation and leveraging advanced image synthesis models, our goal is to ensure that the generated images are contextually relevant, visually appealing, and culturally appropriate. This methodology aims to bridge the gap between textual storytelling and visual imagination, enhancing the accessibility and engagement of story content for diverse linguistic audiences, especially children.

3.1. Multilingual Story Dataset

We utilized the dataset provided by the official MUSIA track, which is specifically curated for story illustration research. The dataset includes both English and Hindi stories, with separate partitions for training and testing. For our experiments, we focused exclusively on the testing set, which comprises 39 English stories and 30 Hindi stories. Additionally, the dataset specifies the exact number of illustrations to be generated for each story, thereby enabling precise alignment between the narrative structure and the corresponding prompt generation process.

English Dataset Story Example Instance:

"Everyone enjoys a holiday. So, one fine Sunday, the Sun also took a holiday. It suddenly disappeared from the sky. Darkness covered the earth. A squirrel was chasing its friend. It stopped and wondered at the darkness. It guessed something was wrong and ran home. A little plant searched for the Sun. It couldn't grow without the Sun's rays. Flowers and leaves of the plant bent low to the ground. Other trees also missed the Sun. Mother bird peeped out of her nest and whispered to its little ones about the darkness. She was sad because she couldn't go out to find food for them. The bee couldn't find any honey because the flowers didn't bloom, so it went back to its hive. The Sun wasn't there, so the wind got stuck up the hill. Men, women and children stopped working. They opened their windows and said, "What happened to the Sun?" It was very quiet everywhere and the warmth of the Sun was missing too. All of them prayed for the Sun to rise. The Sun looked down from its abode. The stillness on the earth shocked the Sun. The Sun felt sorry. The earth seemed lifeless. This made the Sun very sad. So he decided to stop his holiday and start shining again. The plants grew and flowers bloomed. Bees and birds started singing. The river flowed joyfully and the wind blew. Mother bird was happy and everyone on the earth began to work! Indeed, the Sun can never go on a holiday!"

Hindi Dataset Story Example Instance:

"हर दिन रीना सुबह जल्दी उठती है। उठकर बिस्तर को ठीक से लगाती है। नीम की दातुन से अपने दाँत साफ़ करती है। साबुन से नहाकर रीना स्वच्छ कपड़े पहनती है। वह अपने बाल में तेल लगाकर कंधी करती है। रीना माँ के बनाए पराठे और सब्ज़ी आनंद के साथ खाती है। रीना माँ के गले लगती है और फिर स्कूल जाती है। स्कूल के रास्ते में रीना अपनी सहेली दीपा से मिलती है। दोनों एक-दूसरे से सुप्रभात कहती हैं और हँसती-खेलती स्कूल जाती है। स्कूल में प्रार्थना के बाद रीना अपनी कक्षा में जाती है। जैसे ही उनकी अध्यापिका कक्षा में आती हैं, सभी बच्चे खड़े हो जाते हैं और नमस्ते करते हैं। अध्यापिका भी मुस्कुरी हुई नमस्ते करती हैं। रीना स्कूल में मन लगाकर पढ़ाई करती है। वह अपनी सहेलियों के साथ खेलती है और थोड़ी शरारत भी करती है। घर आकर वह हाथ-मुह धोती है। फिर वह अपनी स्कूल की सभी बातें अपने परिवार को बताती है। रीना अपने प्यारे से छोटे भाई के साथ भी खेलती है। रीना को रात को जल्दी ही नींद आ जाती है। दाढ़ी प्यार से रीना को शभुरात्रि कहकर सुला देती हैं।"

3.2. Image Prompt Generation from story

We employed the Gemini-2.5-Pro[4] model's API to generate the prompt for specified number of images for each story. It has input token limit of 1,048,576 and output token limit of 65,536 so it can also generate prompt from very large story. To ensure visual consistency across all illustrations associated with a given story, we prompted the language model to first generate a system prompt that defines the overall thematic style for the images. The exact prompt used for this process is provided below.

Prompt:

Your task is defining {n_image} prompts in english language for image generation for key aspects of below english/hindi story. Give me prompts in very detailed manner so the children can imagine whole story without reading the story.

story:

{story}

Define a system prompt for the theme which are most liked by children for above story:
fixed system prompt from: do not mixup many styles keep it simple

#Style

#Cartoon style, Anime style, Digital painting, Watercolor illustration, Oil painting, Comic book style, Fantasy art, #Pixel art, Sketch art, Realistic style, Flat illustration, Low-poly style, Chibi style, Papercut style, Cel-shaded style, Line art style, Ink wash painting

#Color

#Studio Ghibli colors, Pastel colors, Vibrant colors, Muted tones, Warm color palette, Cool tones, Neon lights, Earth tones, Duotone scheme, Retro color scheme, #Desaturated tones, High contrast colors, Cinematic color grading, Sepia tone, Monochrome palette, Rainbow gradient

#Shading

#Soft shading, Hard shading, Cel shading, Volumetric lighting, Ambient occlusion, Global illumination, Flat lighting, Soft lighting, Harsh shadows, Backlighting, #Rim lighting, Subsurface scattering, Bounce lighting, Ray-traced lighting

#Texture

#Hand-drawn texture, Painted texture, Smooth texture, Sketch-like strokes, Grainy texture, Rough brush strokes, Inked outlines, Crayon texture, Chalk texture #Marker rendering, Pencil sketch texture, Watercolor wash, Canvas texture, Digital airbrush, Etching lines

#Character Consistency

#same characters, Consistent clothing and face, Repeating character model, Fixed hairstyle and outfit, Identical facial features across frames, Character continuity #Preserve facial structure, Consistent outfit design, No change in appearance, Character template unchanged, Use same character across all frames, Maintain character identity, Uniform costume across scenes

#Framing / Composition

#Wide frame, Close-up, Medium shot, Portrait frame, Landscape frame, Bird's eye view, Worm's eye view, Over-the-shoulder view, Centered frame, Rule of thirds composition, #Dynamic camera angle, Symmetrical framing, Diagonal composition, Isometric view, Cinematic framing, Panoramic shot

Ensure that output for system_prompt and image_prompt_i each comes in python dictionary format only, all in one code, discard any extra details except prompt,
e.g: {"system_prompt": [Prompt for theme], "image_prompt_1": ["Prompt for image 1"], ...}

3.3. Image Generation

To ensure thematic consistency across all generated illustrations for each story, we concatenated the system prompt—responsible for defining the overall visual and stylistic guidelines—with each individual image prompt prior to image synthesis. This approach allowed each illustration to adhere closely

to a unified artistic direction, maintaining coherence in terms of color palette, composition, character representation, and other stylistic attributes throughout the entire narrative.

For image generation, we employed Google’s Imagen-4.0-Ultra-Generate-Preview-06-06[5] model released on 2025-06-06, supporting generation with 1:1, 9:16, 16:9, 3:4, 4:3 aspect ratios and watermarking (SynthID). It excels at converting detailed textual prompts into high-quality, visually coherent images. By supplying the combined prompt (system plus image-specific prompt) to the model, we ensured each illustration captured the narrative content while maintaining visual harmony across all images for a given story. This approach enabled the reliable production of illustrations closely aligned with both the story and the desired artistic style. Our system is zero-shot: neither Gemini-2.5-Pro nor Imagen-4.0-Ultra was finetuned. The MUSIA training split was used only to calibrate prompt templates (manual inspection), not to update model parameters; all results reported are on the official test stories.

4. Results

The performance of our approach was evaluated by MUSIA track organizers. A total of 39 stories in English and 30 stories in Hindi were evaluated. The evaluation followed the official criteria of the track, which focused on three dimensions of the generated images: **Consistency**, **Relevance**, and **Visual Quality**. Consistency measures whether the generated images for a story maintain continuity of characters and scenes across multiple frames. Relevance assesses how well the generated images align with the story content. Visual quality evaluates the overall aesthetic quality and clarity of the illustrations. Among these, consistency was assigned the highest weight, followed by relevance, and finally visual quality. Each story was rated on a three-point scale: *Good*, *Moderate*, or *Fair*. For each criterion, the ratings follow the order: *Good* > *Moderate* > *Fair*.

Our system, ranked **1st** in both the English and Hindi tracks of the shared task, outperforming all baseline and competing systems across all three evaluation metrics. The following subsections present detailed results and comparisons.

4.1. English Results

Table 1 reports the result for English stories. Consistency showed slightly lower performance compared to Visual quality and Relevance. This indicates that while the majority of images captured consistent characters and scenes across frames, a subset of stories displayed variations in character attributes or scene continuity.

Table 1

English story visuals result (39 stories)

Metric	Consistency			Relevance			Visual Quality		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
Story Count	24	14	1	34	5	0	39	0	0

Qualitative example (English): The following JSON contains the system and image prompts; the four illustrations in Figure 1 were generated from these prompts.

Generated Json Instance from Gemini-2.5-pro for English story:

```
{
  "system_prompt": "Cartoon style, Studio Ghibli colors, Soft shading, Smooth texture, same characters, Cinematic framing",
  "image_prompt_1": "A wide frame, cinematic still of a whimsical, cartoon world on a beautiful Sunday. On the left, a cute, anthropomorphic sun with a smiling face is waving goodbye as it steps out of the sky, which is turning from bright blue to a dark, starry night."
}
```



Figure 1: English story's qualitative generated images example.

```

Below, a startled cartoon squirrel with big, expressive eyes has stopped mid-run, looking up in confusion at the sudden darkness enveloping the lush green landscape. The transition from light to dark should be clear and dramatic, capturing the moment the holiday begins.", "image_prompt_2": "Medium shot, a heart-wrenching scene under a dark, moonless sky. In the foreground, a small, sad-faced plant droops its head, its flowers and leaves bent low to the greyish ground. Nearby, a mother bird with a worried expression peeks out from her cozy nest, whispering to her little chicks. In the background, through the illuminated windows of cartoon houses, silhouettes of men , women, and children are seen looking out, their hands together in prayer, with questioning looks on their faces. The entire scene is cast in cool, muted tones and soft lighting to convey sadness and stillness.", "image_prompt_3": "Over-the-shoulder view, from behind the cute, round Sun character. The Sun is sitting on a fluffy, celestial cloud, looking down with a very sad and sorry expression. Below is the Earth , depicted as a quiet and lifeless sphere, painted in desaturated, grey tones to show its lack of energy and joy . A single tear rolls down the Sun's cheek. The Sun itself is beginning to glow with a warm, apologetic light, signaling its decision to end its holiday and return to its duty.", "image_prompt_4": "A vibrant panoramic shot of the world bursting back to life. The smiling, radiant Sun is high in the center of the sky, casting warm, golden rays everywhere. Below, the landscape is full of vibrant colors. The little plant is now standing tall with brightly bloomed flowers. Bees are happily buzzing around them. A river flows joyfully with sparkling water . The mother bird soars in the sky. People on the ground are active and smiling, working and playing. The entire image radiates happiness, warmth, and energy, a stark contrast to the previous scenes."
}

```

4.2. Hindi Results

The result for Hindi story visuals is shown in Table 2. For consistency, 27 stories were rated *Good* and 3 *Moderate*, with no stories falling into the *Fair* category. Hindi performance is more stable compared to English.

Table 2
Hindi story visuals result (30 stories)

Metric	Consistency			Relevance			Visual Quality		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
Story Count	27	3	0	30	0	0	30	0	0

Qualitative example (Hindi): The following JSON contains the system and image prompts for hindi story instance; the four illustrations in Figure 2 were generated from these prompts.



Figure 2: Hindi story's qualitative generated images example.

Generated Json Instance from Gemini-2.5-pro for Hindi story:

```
{
  "system_prompt": "Cartoon style, Vibrant colors, Soft shading, Smooth texture, same characters, Consistent clothing and face, Medium shot",
  "image_prompt_1": "A cheerful 7-year-old Indian girl named Reena, with her hair neatly braided and oiled, wearing a clean school uniform, is giving her mother a warm hug before leaving for school. They are in a simple, sunlit kitchen. A plate with a paratha and some vegetables sits on the table nearby. The atmosphere is full of love and warmth. The shot is a medium frame focusing on their happy expressions.",
  "image_prompt_2": "Reena and her friend Deepa, both in identical school uniforms, are walking hand-in-hand down a cheerful, green village path. They are both laughing joyfully, captured mid-giggle, on their way to school, which is visible in the background. The morning sun creates a bright and happy scene. This is a landscape frame showing the two friends and their playful journey.",
  "image_prompt_3": "Inside a bright and colorful classroom, Reena and her classmates, all in uniform, are standing up from their wooden desks. They are all smiling and greeting their kind-faced female teacher with folded hands in a 'Namaste' gesture. The teacher is standing at the front of the class, smiling back warmly. The scene captures a moment of respect and happiness in a learning environment.",
  "image_prompt_4": "In a cozy, softly lit living room at night, Reena is kneeling on a rug, playing happily with her cute toddler brother and some colorful wooden toys. In the background, their loving grandmother is tucking a blanket around Reena, who is looking up at her with a sleepy, content smile, as if she has just been wished 'Shubhratri' (Good night). The scene is peaceful, warm, and full of family love."
}
```

4.3. Analysis and Discussion

The results demonstrate that our pipeline achieved **state-of-the-art visual quality and relevance** across both English and Hindi datasets. This confirms the effectiveness of using Gemini-2.5-Pro for detailed prompt generation and Imagen for producing visually appealing illustrations. However, consistency remains the most challenging criterion. While Hindi stories achieved a high consistency score (0.90 *Good*), English stories showed lower performance (0.62 *Good*, 0.36 *Moderate*). This discrepancy may be explained by the higher narrative complexity and greater character diversity in the English dataset.

Our findings are consistent with prior research on multimodal storytelling, such as the work of Lin et al. [6], which highlights consistency as a persistent bottleneck in visual story generation. In our case, inconsistencies primarily arose in multi-character stories where the LLM occasionally produced prompts with subtle variations in character descriptions across frames. These errors manifested as changes in clothing, hairstyles, or scene backgrounds between images.

Despite these challenges, high relevance and visual quality performance indicate that our zero-shot pipeline successfully preserved narrative fidelity and produced visually engaging outputs. These results underline the strengths of prompt-driven story illustration while pointing toward the need for additional mechanisms—such as character embedding alignment or post-generation refinement—to further improve consistency across complex storylines.

5. Conclusion

This work presented a zero-shot pipeline for multilingual story illustration in the MUSIA track, combining Gemini-2.5-Pro for structured prompt generation with Imagen for text-to-image synthesis. The approach achieved near-perfect visual quality and relevance in English and Hindi, with stronger consistency in Hindi (90%) compared to English (62%). These results confirm the effectiveness of prompt-driven generation for producing coherent, visually engaging illustrations without fine-tuning, while also highlighting consistency as the primary limitation, especially in longer and more complex narratives. The strong Hindi performance demonstrates the potential of such methods for low-resource languages, and future research should focus on enhancing consistency through character embeddings or memory-augmented models, to advance automated multilingual story illustration.

References

- [1] K. Tewari, S. Chanda, A. Malviya, A. Mukherjee, S. Pal, Overview of the shared task on multilingual story illustration: Bridging cultures through ai artistry, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, Association for Computing Machinery, New York, NY, USA, 2026.
- [2] D. Oliveira, E. Ribeiro, D. de Matos, Story generation from visual inputs: Techniques, related tasks, and challenges, arXiv preprint arXiv:2406.02748 (2024). URL: <https://arxiv.org/abs/2406.02748>.
- [3] X. Hong, V. Demberg, A. Sayeed, Q. Zheng, B. Schiele, Visual coherence loss for coherent and visually grounded story generation, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 9456–9470. URL: <https://aclanthology.org/2023.findings-acl.745/>.
- [4] Gemini 2.5 pro, 2024. URL: <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-2.5-pro>.
- [5] Imagen 4.0 ultra, 2024. URL: <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/imagen-4.0-ultra-generate-preview-06-06>.
- [6] X. Zhang, Y. Lin, H. Chen, Improving visual storytelling with multimodal large language models, arXiv preprint arXiv:2407.02586 (2024). URL: <https://arxiv.org/abs/2407.02586>.