NarrArt: Multilingual Story Illustration with AI for English and Hindi Narratives

Mrinmoy Sadhukhan[1,*,†], Indrajit Bhattacharya[2,‡] and Paramartha Dutta[3,‡]

[1]*Department of Computer & System Sciences, Visva-Bharati, Santiniketan, Birbhum, 731235, West Bengal, India.*
[2]*Kalyani Government Engineering College, Kalyani, Nodia, 741235, West Bengal, India*
[3]*Department of Computer & System Sciences, Visva-Bharati, Santiniketan, Birbhum, 731235, West Bengal, India*

Abstract

This work addresses the challenge of multilingual story illustration, with an emphasis on generating visual depictions for narratives in hindi and english. The task holds substantial value in domains such as education and the entertainment industry, where visual storytelling enriches children's literature by adding character visualizations that foster engagement and imagination, and support the creation of comics and animated illustrations that enrich the learning experience. A central problem in this domain is balancing character diversity with narrative consistency, where unconstrained character generation can result in inconsistency across illustrations. To overcome this, a framework is proposed that leverages a curated dataset of publicly available stories and illustrations combined with strategies for maintaining both cultural diversity and visual coherence. The proposed approach can produce illustrations for multilingual stories that are not only semantically aligned with narrative segments but also consistent across an entire story, paving the way for richer, culturally rooted applications in storytelling and creative media.

Keywords

Visual Story-telling, Diffusion Model, Tf-Idf, Summarization,

1. Introduction

Storytelling has long played a vital role in a child's cognitive and imaginative development. In earlier times, children often heard stories narrated by their grandparents, where imagination was fueled by verbal illustrations. Through these narratives, children envisioned characters and scenes in their minds, guided by the expressive storytelling of their elders. However, with the shift towards nuclear family structures in modern society, many children are deprived of this rich storytelling experience. Artificial Intelligence (AI) now offers new opportunities to recreate this experience by providing visual illustrations of stories. Such AI-generated illustrations can spark imagination in children, support teachers in classrooms, and assist illustrators—whose numbers are steadily decreasing—in producing visuals for stories. These illustrations can also enhance comprehension, helping students to better understand narrative content. Despite these advantages, current AI-based visual illustration models face significant limitations. Most state-of-the-art text-to-image generation models are predominantly trained on english language roman script, creating barriers for multilingual storytelling. For example, stories written in Hindi or other Indian languages often cannot be directly processed, restricting the accessibility of such tools. A straightforward solution might involve translating stories into English before feeding them into diffusion-based image generation models. However, this approach encounters challenges when handling long narratives, where translation quality and consistency degrade.

To address these challenges, we propose a pipeline that leverages multiple pretrained models. First, stories are preprocessed and, depending on whether they are in Hindi or English, passed through an

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

[*] Corresponding author.

[†] This author contributed by coding, writing, and editing the paper.

[‡] These authors contributed to reviewing and editing the manuscript.

mrinmoy.sadhukhan1996@gmail.com (M. Sadhukhan); indrajit.bhattacharya@kgec.edu (I. Bhattacharya); paramartha.dutta@visva-bharati.ac.in (P. Dutta)
0000-0002-3246-8295 (M. Sadhukhan); 0000-0002-3881-2755 (I. Bhattacharya); 0000-0003-3946-2440 (P. Dutta)

appropriate translation model. The translated or original text is then segmented into semantically coherent sub-stories. Each segment is summarized using abstractive summarization techniques to produce concise yet meaningful representations. Finally, these summarized sub-stories are used as input to diffusion models, enabling the generation of consistent and contextually accurate illustrations for multilingual storytelling.

The remainder of this paper is organized as follows. Section 2 reviews the existing work in this domain and highlights the key innovations of previous approaches. The description about the dataset in presented in section 3. The proposed methodology is described in section 4. Section 5 presents the experimental results along with key evaluation metrics. A discussion on the findings and potential directions for future improvement is provided in section 6. Finally, the paper is concluded in section 7.

2. Related Work

In the field of text-to-image generation, three primary categories of models have been explored, namely Generative Adversarial Network (GAN)-based models, diffusion-based models, and transformer-based models. In GAN-based models, the generator produced images conditioned on noise vectors and text embeddings, while the discriminator attempted to determine whether an image was real or synthetic, thereby improving the generator through adversarial learning. In contrast, diffusion-based models always begin with pure noise and iteratively denoise it step by step, guided by text embeddings, to synthesize an image. Transformer-based models treated images as sequences of tokens and jointly modeled image tokens with text tokens in an autoregressive sampling framework, enabling image generation conditioned on textual input. Xu et al. [1] proposed AttnGAN (Attentional Generative Adversarial Network), which generated images from text descriptions by focusing on relevant words corresponding to different regions of the image. High-resolution images were synthesized progressively across multiple stages, where attention over word-level embeddings refined the image resolution step by step. Qiao et al. [2] presented a GAN-based text-to-image model consisting of three major components: the Semantic Text Embedding Module (STEM), the Global–Local Collaborative Attentive Module (GLAM), and the Semantic Text Regeneration and Alignment Module (STREAM). The STEM module employed a recurrent neural network to obtain word- and sentence-level embeddings. The GLAM module was responsible for generating images at multiple scales, while the STREAM module, implemented with an LSTM, attempted to regenerate the text description from the generated images to ensure semantic alignment with the input text. Saharia et al. [3] introduced a diffusion-based model, Imagen, for text-to-image generation. The model incorporated a pretrained T5-based transformer to obtain text embeddings, which were then injected as conditions into an efficient U-Net backbone for image synthesis. The U-Net progressively upscales images to higher resolutions by transferring learned weights across different scales. This approach achieved an impressive FID score of 7.27. Ramesh et al. [4] proposed the DALL-E 2 model, a decoder-based diffusion framework. The model followed a two-stage pipeline: in the first stage, text was mapped into CLIP (Contrastive Language–Image Pretraining) embeddings, and in the second stage, a diffusion-based decoder generated images conditioned on these embeddings. Rombach et al. [5] proposed the Stable Diffusion model. In this framework, images were first compressed into a latent representation using a trained variational autoencoder (VAE). Diffusion-based denoising was then applied within this latent space, where CLIP-based text embeddings were injected into a U-Net to guide the denoising process. Finally, the VAE decoder reconstructed the high-resolution image from the denoised latent representation. In the category of transformer-based models, Ramesh et al. [6] presented a zero-shot text-to-image generation framework, known as DALL-E. The approach first compressed images into a 32×32 latent space using a discrete VAE. The encoded text was then concatenated with the latent representation, and an autoregressive transformer was trained to model the joint distribution. Due to its zero-shot nature, this model demonstrated the ability to handle unusual and imaginative prompts effectively.

In multilingual text-to-image generation, translation plays a crucial role. A simple text encoder alone is not sufficient, since additional pretrained models are required to handle translation from multilingual text into English. Furthermore, summarization techniques are often applied to compress sentences, ensuring they are suitable as input for the generative model. Ramesh et al. [7] from the AI4Bharat team proposed the IndicTrans model for multilingual text-to-English conversion. It supports 11 Indian languages and was trained on the publicly available Samanantar corpus. Their implementation is based on the fairseq framework, using a Transformer architecture with six encoder and six decoder layers, an input embedding dimension of 1536, and 16 attention heads. Subsequent improvements led to IndicTrans2 [8], which extends support to 22 scheduled Indian languages spanning diverse scripts. IndicTrans2 employs 18 encoder layers and 18 decoder layers, with an input embedding dimension of 1024, a feed-forward dimension of 8192, and 16 attention heads, resulting in a total of 1.1B parameters. The model achieved an average score of 62.8 on the IT2 benchmark and currently serves as the backend for the Government of India's Bhashini application. Costa-jussà et al. [9] from the NLLB (No Language Left Behind) team introduced the NLLB model, which supports translation across approximately 200 languages, with particular focus on high-resource languages such as German, French, and English. It is a Transformer-based sequence-to-sequence model with 54B parameters, although smaller variants are also available. The training incorporates a Mixture-of-Experts (MoE) strategy to improve efficiency. The reported performance reaches 44.8 on the ChrF++ metric. In addition to these research models, several commercial large language models (LLMs), such as Google's Gemini, OpenAI's GPT-4o, and Anthropic's Claude, can also be leveraged for translation tasks.

In the summarization domain, two primary approaches exist: extractive and abstractive summarization. For story generation tasks, abstractive summarization is particularly important, as it enables the creation of abstracted narratives that can subsequently be fed into image generation models. Lewis et al. [10] from Facebook proposed BART, a denoising sequence-to-sequence model. BART integrates two powerful pretrained components: a bidirectional Transformer encoder (similar to BERT) and an autoregressive Transformer decoder (similar to GPT). It has been widely applied to tasks such as summarization, text reconstruction, and sentence generation. A fine-tuned variant, BART-Large-CNN, demonstrates strong performance on abstractive summarization, achieving robust ROUGE scores on the CNN/DailyMail dataset. Zhang et al. [11] from Google introduced PEGASUS, a Transformer-based encoder–decoder architecture explicitly designed for summarization tasks. The model is pretrained using two key strategies: gap-sentence generation and masked language modeling. Pretraining is performed on large-scale corpora, including the HugeNews dataset and the C4 dataset, followed by fine-tuning on summarization benchmarks such as CNN/DailyMail and others. Raffel et al. [12] from Google proposed the T5 (Text-to-Text Transfer Transformer) model, a versatile encoder–decoder Transformer framework. T5 is capable of handling diverse downstream tasks, including translation, question answering, summarization, and classification, by casting all tasks into a unified text-to-text format. The largest variant, with 11B parameters, was trained on the C4 corpus and achieved a score of 83.28 on the GLUE benchmark. Beltagy et al. [13] from AllenAI presented the Longformer model, which introduces an efficient attention mechanism called the attention pattern module to reduce the quadratic time complexity of standard self-attention in Transformers. Trained in an autoregressive manner, Longformer is particularly well-suited for long-document summarization, such as scientific papers and legal documents. On summarization benchmarks, the model achieved a ROUGE score of 44.4.

From the above discussion, it can be observed that, apart from encoder–decoder based models, a wide range of large language models (LLMs) exist, available in both open-source and proprietary forms. These models often demonstrate remarkable performance; however, their deployment on consumer-grade GPUs is generally infeasible due to high computational requirements. Some open-source variants can be executed locally through projects such as ollama.cpp, but this comes at the cost of significantly slower inference speeds. Alternatively, several LLMs can be accessed via API calls. While this approach simplifies usage, it introduces practical limitations, such as restricted request quotas and potential latency, making it unsuitable for continuous large-scale tasks. Furthermore, the direct use of pretrained models for summarization or translation is constrained by strict token-length limits. As a result, handling long-form narratives requires specialized methods for segmenting, preprocessing, and reassembling the text before feeding it into such models, which are described in below sections.

3. Dataset Description

The dataset provided by the MUSIA Shared Task [14, 15] comprises stories in both English and Hindi. For the English training dataset, there are two primary directories: one containing the story text files and another containing the corresponding images. The story directory for English includes 360 text files, each containing a multi-paragraph story. The image directory holds the related images, named using the pattern eng_story_XXX_01, where eng_story_XXX corresponds to the respective story file in the stories directory. The Hindi training dataset follows the same structure, containing 185 story text files along with their corresponding images. For the testing datasets, the English set includes 40 stories, while the Hindi set includes 30 stories. Unlike the training data, the test sets do not contain any images. Instead, each test set includes a JSON file — EN_story_image_counts.json for English and its equivalent for Hindi — which specifies the number of images that should be generated for each story. This setup enables the evaluation of a model's ability to generate a visually consistent and contextually appropriate number of images corresponding to each story during testing.

4. Methodology

In this section, we present our approach for generating visual illustrations from multilingual story data. To ensure clarity and maintain scope, our work focuses on two languages: Hindi and English. For Hindi stories, we first employ the pretrained translation model `facebook/nllb-200-distilled-600M` [9] to translate text from Hindi (Devanagari script) into English (Latin script). Both the translated stories and the original English texts are then standardized through a preprocessing stage, which removes redundant spaces, newline characters, and other formatting inconsistencies. Next, each story is segmented into semantically coherent sub-stories, corresponding to the number of images to be generated per story. To achieve this segmentation, sentences are first encoded using a TF–IDF vectorizer [16], and cosine similarity scores are computed between them. Sentences exhibiting higher similarity values are grouped together, while low-similarity sentences delineate the boundaries between segments, ensuring that each segment captures a coherent narrative idea. Since abstractive summarization models impose sequence length constraints, these segments are subsequently rebalanced to achieve uniformity. Each segment is initially tokenized using NLTK's `word_tokenize` function, and token counts are adjusted to ensure near-equal distribution across segments. The tokens are then reconstructed into sentences, and each token-limited segmented sentence or chunk is validated against a maximum sequence length of 512 tokens using the `facebook/bart-large-cnn` tokenizer [10]. In rare cases, particularly long sentences are truncated, which may cause minor degradation in summarization quality. After balancing, each chunk is passed through the `facebook/bart-large-cnn` abstractive summarization model [10] to generate concise summaries that encapsulate the essence of each story segment. These summaries serve as textual prompts for the `stable-diffusion-X1-base-10` model [5], which produces corresponding visual illustrations. This pipeline effectively bridges the gap between multilingual storytelling and automated visual generation. A detailed workflow of the proposed method is illustrated in figure 1.

5. Result

During the evaluation of the proposed methodology, our focus was not on assessing the quality of the pretrained models themselves, as these models are already well-established and widely validated. Instead, the primary objective was to evaluate the overall effectiveness of the proposed framework—specifically, its ability to generate images from multilingual stories with varying numbers of narrative segments. For this purpose, we utilized the test dataset provided by the MUSIA Shared Task. It is important to note that we did not fine-tune any of the pretrained models used in this framework due to the significant GPU resource requirements associated with such training. Instead, all models were employed in their off-the-shelf pretrained configurations, ensuring that the evaluation focuses purely on the integration

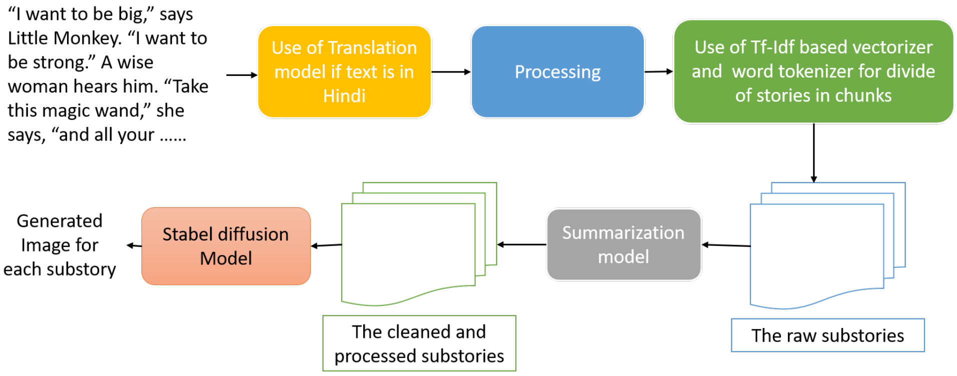

"I want to be big," says Little Monkey. "I want to be strong." A wise woman hears him. "Take this magic wand," she says, "and all your ......
Use of Translation model if text is in Hindi
Processing
Use of Tf-Idf based vectorizer and word tokenizer for divide of stories in chunks
The raw substories
Summarization model
The cleaned and processed substories
Stabel diffusion Model
Generated Image for each substory

Figure 1: The proposed method of multilingual story illustration

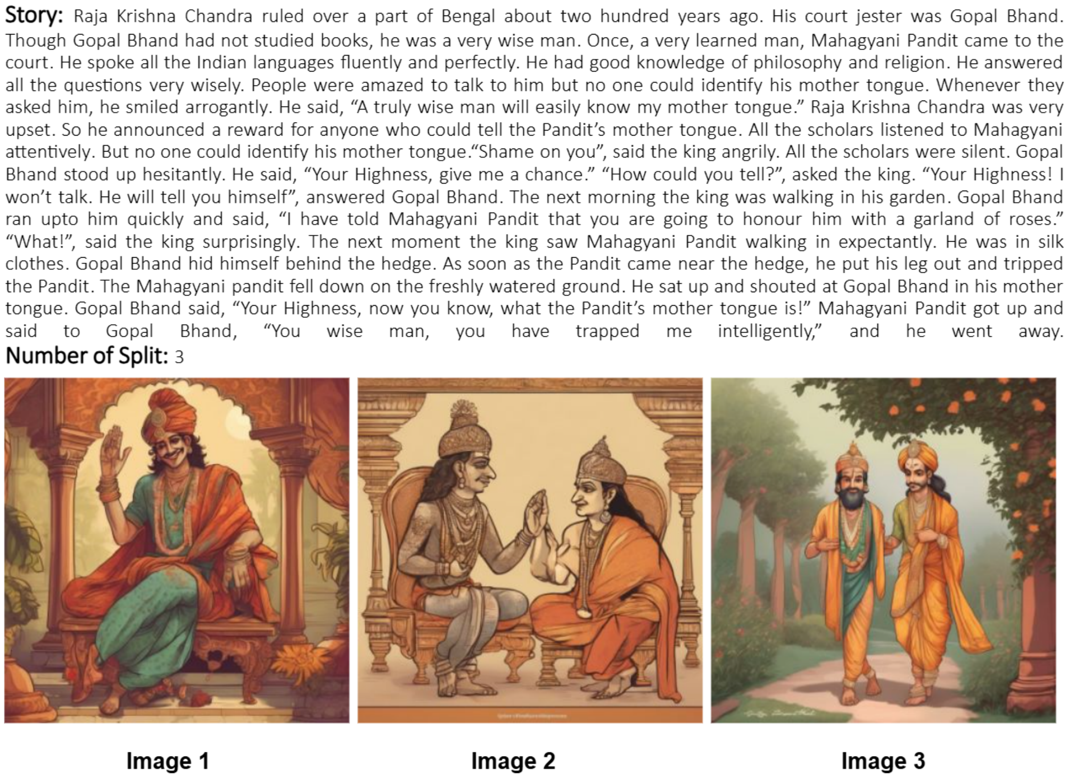
Story Language	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
English (ours)	15	22	2	3	10	26	3	8	28
English (Retriever)	39	0	0	34	5	0	24	14	1
English (Nandini_Divya)	36	2	0	5	21	12	8	18	12
English (NLPFusion)	0	0	39	0	6	33	0	2	37
English (JU Team)	0	0	39	0	2	37	0	0	39

Table 1
Evaluation of generated story illustrations across three criteria (Visual Quality, Relevance, Consistency) for English stories.

Story Language	Visual Quality			Relevance			Consistency		
	Good	Moderate	Fair	Good	Moderate	Fair	Good	Moderate	Fair
Hindi (ours)	19	11	0	0	0	30	0	3	27
Hindi (Retriever)	30	0	0	30	0	0	27	3	0
Hindi (Nandini_Divya)	30	0	0	10	15	5	15	10	5
Hindi (NLPFusion)	0	0	30	2	4	24	0	2	28

Table 2
Evaluation of generated story illustrations across three criteria (Visual Quality, Relevance, Consistency) for Hindi stories.

and performance of the proposed pipeline rather than the optimization of individual components. Several automatic evaluation metrics are commonly employed to assess the performance of generative models, including Fréchet Inception Distance (FID), Inception Score (IS), Learned Perceptual Image Patch Similarity (LPIPS), CLIPScore, and Diversity Score. These metrics evaluate different aspects of generative performance, such as visual coherence, semantic alignment, and output diversity. FID and IS are traditionally used to assess the visual realism and fidelity of generated images by comparing them with real images, and are particularly valuable during the training phase of models such as diffusion-based generators. For text-to-image generation, however, CLIPScore is more suitable, as it directly measures the semantic consistency between the input text and the generated image. The score ranges from 0 to



Story: 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 had good knowledge of philosophy and religion. 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.

Number of Split: 3

Image 1 **Image 2** **Image 3**

Figure 2: Visual Illustration of English Story.

1, with higher values indicating better alignment and higher-quality outputs. Additionally, diversity can be evaluated using an LPIPS-based Diversity Score, which computes perceptual distances between multiple images generated from the same text prompt. This reflects the model's capacity to produce varied yet semantically related outputs—where a lower score indicates reduced diversity and a higher score suggests greater variability. Since our framework employs a pretrained text-to-image generation model, metrics such as FID and IS are less informative, as they primarily assess image realism rather than text–image correspondence. Therefore, we report CLIPScore as our primary evaluation metric. On the MUSIA test dataset, our approach achieved an average CLIPScore of 0.31, indicating a moderate level of semantic alignment between the generated images and their corresponding story segments. Nonetheless, we recognize that CLIPScore is still a developing and limited metric, underscoring the need for more comprehensive evaluation approaches tailored to multilingual text-to-image generation.

Hence, we relied on human evaluation conducted by a team of experts from the MUSIA Shared Task. The evaluators assessed the generated illustrations based on three key criteria: Relevance, Visual Quality, and Consistency. Relevance measures how effectively the illustrations capture the key moments and semantics of the story; Visual Quality assesses the originality, aesthetic appeal, and overall artistic presentation; and Consistency evaluates whether the sequential images from different story segments form a coherent and continuous narrative. Among these, Consistency is given the highest priority during rating, followed by Relevance and Visual Quality. To quantify inter-rater reliability, Cohen's Kappa coefficient was computed, and the final ratings were categorized into three levels—Good, Moderate, and Fair. Table 1 presents the aggregated evaluation results for English stories, while Table 2 reports the corresponding results for Hindi stories. Both tables include a comparative analysis between our proposed framework and other participating teams whose submissions were shortlisted in the MUSIA Shared Task. From our results, it can be observed that our proposed method performs strongly in terms of Visual Quality for both English and Hindi stories. In terms of Relevance, the English stories achieve comparatively better performance, while the Hindi stories exhibit slightly lower scores. Regarding Consistency, both language categories achieve moderate performance, which can be considered satisfactory given the experimental constraints; this is expected, as pretrained models without

Story: 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 कंघी करती है। रीना माँ के बनाए पराठे और सब्ज़ी आनं द 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-दूस रे 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 मुस्कुती हुई नमस्ते करती हैं। रीना स्कूल में मन लगाकर पढ़ाई करती है। वह अपनी सहेलियों के साथ खेलती है और थोड़ी शरारत भी करती हैं। 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 भीखेलती है।रीना को रात को जल्दी ही नींद आ जाती है। दादी प्यार से रीना को शभुरात्रि कहकर सुला देती हैं.

Number of Split: 4

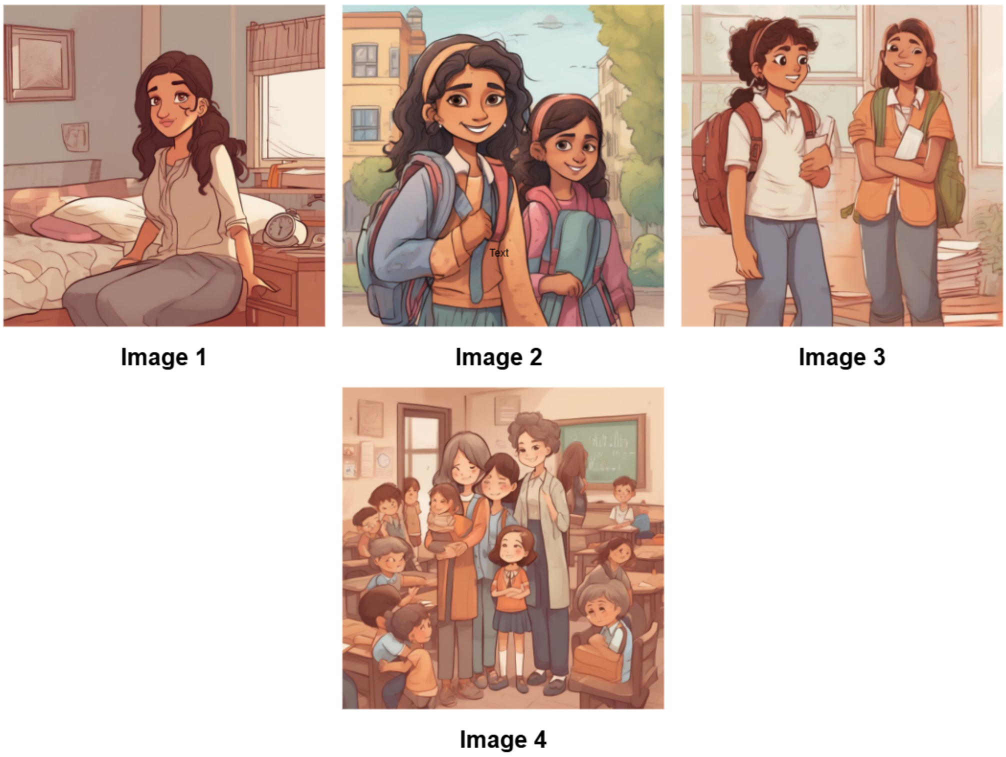


Image 1 **Image 2** **Image 3**

Image 4

Figure 3: Visual Illustration of Hidi Story.

domain-specific fine-tuning often struggle to maintain strong relevance and narrative consistency across all generated outputs. Figure 2 illustrates the application of the proposed methodology on an English story, where the narrative is divided into multiple segments, and for each segment, a corresponding image is generated to visually depict the key events and preserve the story's progression. Similarly, Figure 3 demonstrates the application of the method to a Hindi story, showcasing the framework's multilingual capability and its effectiveness in producing coherent visual illustrations across different languages.

6. Discussion

Multilingual story illustration is an important and challenging problem, yet there is currently no proven method that can produce perfect illustrations directly from text. The accuracy of story-to-image generation largely depends on how the story is provided to the model, as well as the length and complexity of the narrative being fed. From the methodology and results discussed above, it can be observed that different pretrained models can be integrated at various stages of the proposed pipeline. However, the final quality of illustrations ultimately depends on diffusion-based image generation models and the way in which the script is structured to guide the model in understanding the intended scene. For multilingual stories, it is our assumption that the IndicTrans model developed by AI4Bharat could potentially yield better results in translation tasks compared to the pretrained models we have used, since it is specifically trained and validated on native Indian languages. One limitation we have observed is that diffusion-based models are computationally heavy, and their training requires complex fine-tuning procedures, which are sometimes impossible in consumer-grade GPUs. In this work, we have employed a pretrained diffusion model without additional fine-tuning on the given training and validation datasets, which occasionally led to inconsistencies and mismatches in visualization. The development of lightweight diffusion models tailored specifically for this type of illustration task could significantly reduce the dependency on heavy, server-grade GPUs. Such models would also allow easier fine-tuning, thereby facilitating domain adaptation with minimal effort and further enhancing the quality and effectiveness of the generated illustrations.

7. Conclusion

This work explored the problem of multilingual story illustration, with a focus on English and Hindi narratives. By combining pretrained language and vision models within a story segmentation based framework, we demonstrated that it is possible to generate illustrations that are semantically aligned with story segments while maintaining reasonable narrative coherence. Human evaluation showed that the method performs well in terms of visual quality for both languages, achieves strong Relevance in English but weaker in Hindi, and provides moderate consistency across stories. Despite these promising results, the reliance on heavy diffusion models without fine-tuning introduced occasional mismatches in visualization and consistency. Future directions include developing lightweight diffusion models for easier adaptation and leveraging stronger multilingual translation systems, such as IndicTran, to improve performance on native languages. Overall, this study lays the groundwork for culturally rooted, coherent, and scalable multilingual story illustration systems with applications in education, entertainment, and creative media.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4 and Grammarly to check grammar and spelling. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, 2017. URL: http://arxiv.org/abs/1711.10485. doi:10.48550/arXiv.1711.10485, arXiv:1711.10485 [cs].

[2] T. Qiao, J. Zhang, D. Xu, D. Tao, MirrorGAN: Learning Text-to-image Generation by Redescription, 2019. URL: http://arxiv.org/abs/1903.05854. doi:10.48550/arXiv.1903.05854, arXiv:1903.05854 [cs].

[3] 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: http://arxiv.org/abs/2205.11487. doi:10.48550/arXiv.2205.11487, arXiv:2205.11487 [cs].

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022. URL: http://arxiv.org/abs/2204.06125. doi:10.48550/arXiv.2204.06125, arXiv:2204.06125 [cs].

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, 2022. URL: http://arxiv.org/abs/2112.10752. doi:10.48550/arXiv.2112.10752, arXiv:2112.10752 [cs].

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, 2021. URL: http://arxiv.org/abs/2102.12092. doi:10.48550/arXiv.2102.12092, arXiv:2102.12092 [cs].

[7] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, M. J, D. Kakwani, N. Kumar, A. Pradeep, S. Nagaraj, K. Deepak, V. Raghavan, A. Kunchukuttan, P. Kumar, M. S. Khapra, Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages, 2023. URL: http://arxiv.org/abs/2104.05596. doi:10.48550/arXiv.2104.05596, arXiv:2104.05596 [cs].

[8] J. Gala, P. A. Chitale, R. AK, V. Gumma, S. Doddapaneni, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, P. Kumar, M. M. Khapra, R. Dabre, A. Kunchukuttan, IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages, 2023. URL: http://arxiv.org/abs/2305.16307. doi:10.48550/arXiv.2305.16307, arXiv:2305.16307 [cs].

[9] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No Language Left Behind: Scaling Human-Centered Machine Translation, 2022. URL: http://arxiv.org/abs/2207.04672. doi:10.48550/arXiv.2207.04672, arXiv:2207.04672 [cs].

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2019. URL: http://arxiv.org/abs/1910.13461. doi:10.48550/arXiv.1910.13461, arXiv:1910.13461 [cs].

[11] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, 2020. URL: http://arxiv.org/abs/1912.08777. doi:10.48550/arXiv.1912.08777, arXiv:1912.08777 [cs].

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2023. URL: http://arxiv.org/abs/1910.10683. doi:10.48550/arXiv.1910.10683, arXiv:1910.10683 [cs].

[13] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, 2020. URL: http://arxiv.org/abs/2004.05150. doi:10.48550/arXiv.2004.05150, arXiv:2004.05150 [cs].

[14] 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.

[15] 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.

[16] M. Sadhukhan, P. Bhattacherjee, T. Mondal, S. Dasgupta, I. Bhattacharya, Opinion classification at subtopic level from COVID vaccination-related tweets, Innovations in Systems and Software Engineering 21 (2025) 215–226. URL: https://doi.org/10.1007/s11334-022-00516-9. doi:10.1007/s11334-022-00516-9.