PoemTale Diffusion Framework : Connecting Verses and Visuals in Poetic Storytelling

Anonymous ACL submission

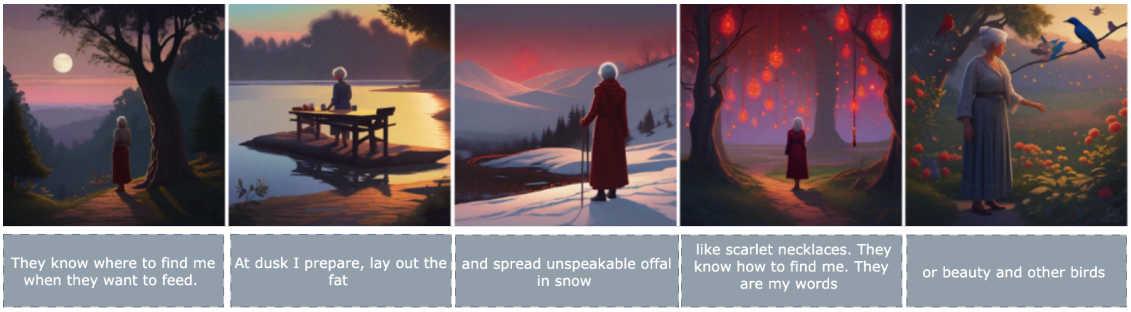

Figure 1: Here we show the images generated by our proposed framework *PoemTale Diffusion*, corresponding to various segments of the poem (shown in the grey boxes). These segments are generated from the poem via our novel Entity-Plus-Emotion algorithm. Our approach captures intricate details of segments while preserving appearance of the main character consistently.

Abstract

Vision is a common source of inspiration and comprehension for poetry. Despite significant research in the field of creative language processing in text and image generation, the intersection of poetry and visual representation has been largely overlooked. Recognizing the untapped potential of representing poetic verses with visual illustrations, we propose the *PoemTale Diffusion* Framework. This framework incorporates consistent self-attention and image generation techniques with four key modules: poem segmentation based on context shifts, generation of relevant image descriptions, main character identification, and creation of visually coherent images. We employ a consistent self-attention mechanism to preserve the poem's central character in the generated images, ensuring visual coherence. In addition, we propose a segmentation algorithm, *Entity Plus Emotion (EPE)*, which segments poems based on shifts in entities and emotions, aiding in the creation of a coherent narrative structure. Furthermore, to address the limited resources availability, we introduce *Hindi Odes and Poems of Expressive Simplicity (HOPES)*, a pioneering dataset for poems in an Indian language, comprising 3,988 Hindi poems. Utilizing the *HOPES* dataset, along with existing datasets, we conduct extensive evaluations and compare our proposed framework with state-of-the-art methods for the visual narration of segmented poems. Our image generation approach not only enhances poem narration but also fosters a more detailed understanding of poems by capturing emotional and entity shifts. *The codes and dataset are available in the supplementary material*[1].

1 Introduction

Poetry is a distinctive form of expression that evokes emotions and feeds the imagination. Every person who reads a poem undergoes distinct emotions and imaginations, which can result in a multitude of interpretations. This variety makes reading poetry particularly captivating, as it invites readers to connect with the poet's inspiration while crafting their own personal perspectives. The differing imaginations of poets and readers create a rich tapestry of interpretations and emotional responses. In the field of natural language processing (NLP), a variety of generative tasks related to poetry have been thoroughly investigated. These include topic-based poem generation (Ghazvininejad et al., 2016;

[1]We will release the codebase and entire dataset upon acceptance.

Yan et al., 2013; Zhang and Lapata, 2014), acrostic poem generation (Agarwal and Kann, 2020), rhetorically controlled poetry generation (Liu et al., 2019), and poem generation from images (Cheng et al., 2018). Additionally, the rise of large language models (LLMs) has led to extensive exploration of different prompting techniques for poem generation (Chakrabarty et al., 2022; Lo et al., 2022; Huynh and Bao, 2024). Despite these advancements, the generation of images from poetry remains a relatively under-explored area in research. *In our work, we aim to address this gap by introducing an innovative visual narration framework that generates sequential images from poems by employing consistent self-attention techniques along with our proposed segmentation modules.*

Motivation and Research Questions: Poetry, as a unique and profound form of human expression, often receives limited attention from researchers, particularly in the areas of comprehension and visualization (Thompson, 1978). Poetry, with its layered meanings and emotional depth, often presents a challenge for those unfamiliar with literary language. For example, a Shakespearean sonnet, like *"Shall I compare thee to a summer's day?,"* may captivate readers with its beauty, yet its deeper themes such as the contrast between fleeting beauty and eternal love can be missed. Visualization techniques can help make these complexities more accessible. However, poetry frequently encompasses hidden meanings, intricate narratives, and subtle emotional variations that present challenges for visualization. This presents a significant inquiry:

RQ1: *How can we effectively capture the context shifts inherent in poetry to generate images that reflect these changes while preserving the narrative structure?*

RQ2: *How can we quantitatively evaluate these images in relation to the segmented parts of the poem, ensuring that they accurately represent and preserve the essence of the preceding segments?*

To address these challenges, we propose a segmentation **Entity Plus Emotion (EPE)** algorithm designed to divide poems into distinct segments based on shifts in entities and emotional tones. This algorithm identifies changes in both the characters and the emotional landscape of the poem, enabling improved comprehension and representation of its thematic transitions. To quantitatively evaluate the images generated based on this segmentation, we introduce the **Visual Narration Score (VNS)**. This metric evaluates the degree of alignment and consistency between the segmented text of a poem and its corresponding images, providing a robust measure of how well the visuals capture and reflect the evolving narrative and emotional context of the poem. Furthermore, to advance our research towards poetry visualization, we applied our proposed framework to poems written in Hindi. To facilitate this, we curated a dataset of 3,988 Hindi poems. We evaluated its effectiveness through both qualitative analysis and quantitative metrics, comparing results with existing state-of-the-art methods and our own proposed framework. To summarize, we make the following key contributions:

Datasets	Type	BL	MM	TIG
IDN-Sum	Games	✗	✗	✗
MovieBook	Movies	✗	✓	✗
BookCorpus	Books	✗	✗	✗
PoemSum	Poems	✗	✗	✗
Movie Description	Movies	✗	✓	✗
HOPES (Ours)	Poems	✓	✓	✓

Table 1: Comparison of Datasets Based on Bilinguism (BL), Multimodality (MM) and Applications in Text-to-Image Generation (TIG) Tasks.

- *We introduce a **novel** task for poem narration using a distinct framework, **PoemTale Diffusion**, which generates multimodal outputs in the form of visual cues.*
- *We introduce a **pioneering** dataset, **HOPES: Hindi Odes and Poems of Expressive Simplicity**, a multimodal dataset comprising 3988 collections of Hindi poems.*
- *We propose a **novel** poem-segmentation algorithm, **Entity Plus Emotion (EPE)**, which segments poems based on context shifts, capturing changes in entities and emotions.*
- *We introduce a **novel** metric, **Visual Narration Score (VNS)**, which is designed to assess the effectiveness of visual narration by quantifying the alignment between poem text segments and their corresponding images.*

2 Background and Related Works

Recent advancements in Large Language Models (LLMs) and Vision-Language Models (VLMs) have sparked interest in art, design, and content creation, offering innovative tools for generating,

interpreting, and visualizing creative concepts.

Text-to-Image Synthesis Via GANs: Early approaches often relied on conditional generative models, which utilized textual descriptions to guide the generation of corresponding images. These methods predominantly depended on Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). With the introduction of the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) model, which has been trained on large datasets consisting of millions of pairs of text and images, is known for its ability to represent image-text pairs with semantic depth accurately, along with various GAN-based models (Abdal et al., 2022; Xia et al., 2021; Andonian et al., 2021; Gal et al., 2022b; Patashnik et al., 2021) that have significantly contributed to the progress in text-to-image synthesis. ***Limitations:*** *Despite their significance, GANs are usually difficult to train, emphasizing the need for a training-free model in the image generation process.*

Visual Narration Via Diffusion Models: The recent boom in diffusion models (Sohl-Dickstein et al., 2015) has fundamentally transformed the field of text-to-image generation. These models, such as IMAGEN (Saharia et al., 2022), DALL-E 3 (Ramesh et al., 2022), and Stable Diffusion (Rombach et al., 2022), have quickly become popular because of their exceptional capacity to generate images of high quality and variety. In the area of image editing, previous studies such as Blended Diffusion (Avrahami et al., 2022), Blended Latent Diffusion (Avrahami et al., 2023), and Paint by Example (Yang et al., 2023) have tackled the task of maintaining backgrounds by integrating masks provided by the user. However, despite these developments, the subject's identity is not consistently preserved across various images. To address this challenge, researchers proposed ID Preservation, a method that aims to generate images while retaining specific identities. It requires an initial reference image to maintain consistency across generated images. Several approaches have been developed for this purpose, such as DreamBooth (Gal et al., 2022a), Custom Diffusion (Ruiz et al., 2023), IPAdapter (Kumari et al., 2023), and PhotoMaker (Ye et al., 2023). (Jeong et al., 2023) proposed a method for story synthesis that combines a pre-trained LLM with a text-guided Latent Diffusion Model to produce coherent images. In a recent study by (Yu et al., 2024), a text-driven point cloud generation pipeline was developed to create compelling and coherent sequences of 3D scenes. Similarly, (Liu et al., 2024) introduced a method to generate coherent image sequences based on a given storyline. In another work, (He et al., 2024) employed Masked Mutual Self-Attention (MMSA) and Masked Mutual Cross-Attention (MMCA) modules to produce consistent images aligned with story narration. ***Limitations:*** *Despite their achievements, these studies primarily focus on maintaining image consistency for texts, paragraphs, or stories. None has addressed the challenge of interpreting and visualizing poetry. Poems are inherently complex literary works with layered meanings, requiring a specialized approach for accurate interpretation, which is the focus of our work.*

3 Corpus Description

To evaluate the scalability, robustness, and reliability of our framework, we employed three datasets that vary in genre, language, and size.

A) *PoemSum* Dataset: The *PoemSum* dataset (Mahbub et al., 2023) comprises 3,011 poetry samples paired with their English-language summaries. The poems in this dataset are sourced from various online platforms, while the summaries were curated from "Poem Analysis", a well-regarded website recognized for its extensive collection of detailed poem summaries. Each entry in the dataset includes the poem's title, the poet's name, the full text of the poem, the source website, and the corresponding summary.

B) *HOPES* Dataset: To the best of our knowledge, there is currently no freely available dataset dedicated to poems in Indian languages. To address this gap, we are introducing the ***Hindi Odes and Poems of Expressive Simplicity (HOPES)*** dataset. Our dataset comprises a collection of Hindi poem samples, including each poem's title, the poet's name, the poem text, and the source website. In Table 1, we present a comparison of our proposed dataset with other creative language datasets, such as those related to games (IDN-Sum (2022)), movies (MovieBook (2015) and movie descriptions (2015)), BookCorpus (2015) for books, and PoemSum (2023) for poems.

Dataset Collection: We gathered data from several open-source websites featuring Hindi poems, utilizing Python's BeautifulSoup library for web scraping. This process yielded 3,988 unique poem samples.

Dataset Cleaning: Due to the varied sources of the

collected data, extensive cleaning was necessary to ensure consistency. The cleaning process involved removing duplicate entries, correcting inaccuracies, eliminating extraneous whitespace and HTML tags, and addressing other errors. We enlisted three postgraduate students from the humanities department, all with linguistic expertise in Hindi, to assist with data verification. After data cleaning, these students carefully reviewed the dataset to ensure that all issues were resolved. *We present sample entries from the dataset in the Supplementary.*

C) *TinyRhymes Dataset:* To evaluate the robustness and scalability of our framework across different genres, we compiled a small dataset of 139 nursery rhymes, which we've named *TinyRhymes* for ease of reference. While previous datasets covered a variety of genres, they lacked content focused on children's poetry. *TinyRhymes* addresses this gap, enhancing poem analysis by including children's poems along with their titles. This dataset was curated from open-source internet websites with a primary focus on nursery rhymes and was meticulously verified by three final-year undergraduate students proficient in English. This dataset was utilized in our study to evaluate the efficacy of our framework, specifically in the domain of children's poetry.

4 Our Methodology

4.1 Problem Definition

Given a dataset where each data point is a poem represented as a textual sequence $T = \{w_1, w_2, \ldots, w_k\}$ where $w_1, w_2, \ldots, w_k$ are the individual words comprising the poem. The task is to generate a sequence of images $\{I_1, I_2, \ldots, I_n\}$ that visually narrates the story conveyed by T, while maintaining a coherent narrative structure. The textual sequence T is segmented into multiple parts $\{S_1, S_2, \ldots, S_n\}$, where each segment $S_i = \{w_{p_i}, w_{p_i+1}, \ldots, w_{q_i}\} \subseteq T$ is identified based on contextual shifts such as changes in entities, emotions, or themes. The objective is to design a framework, the *PoemTale Diffusion* framework, which generates a corresponding sequence of images $\{I_{S_1}, I_{S_2}, \ldots, I_{S_n}\}$, where each image I_{S_i} visually represents the narrative and context of its respective segment S_i. The image generation process ensures that the sequence $\{I_{S_1}, I_{S_2}, \ldots, I_{S_n}\}$ collectively captures the narrative structure of the entire poem T, reflecting the segmented parts $\{S_1, S_2, \ldots, S_n\}$ accurately and coherently.

4.2 *PoemTale Diffusion*: Visual Narration of Poem

To make poetry more understandable to readers, we present the *PoemTale diffusion* framework, which generates images for poetic stanzas, resulting in a story-like visual narration. In this section, we introduce our *PoemTale Diffusion* pipeline, which allows for training-free poem narration while maintaining the coherence of the poem's main character. For a comprehensive analysis, our approach is divided into five distinct modules: (i) Segmentation Module, (ii) Main Character Identification Module, (iii) Image Generation Module, (iv) Prompt Generator and (v) Translation Module.

4.2.1 Segmentation Unit

We employ an *Entity Plus Emotion (EPE)* approach to detect segment boundaries in a poem. Our approach incrementally processes each line of the poem, assigning NER tags and emotions to each line. When there is a change in NER tags or emotions between consecutive lines, the algorithm detects a shift in context and segments the poem accordingly. This method focuses exclusively on identifying where one segment of the poem ends, and another begins based on changes in context. We employed the NLTK library to identify entities in the poem and employed a fine-tuned version of DistilRoBERTa-base (Hartmann, 2022) for the classification of emotions. *The core concept is outlined in Appendix A.2.*

4.2.2 Main Character Identification (MCI) Unit

The primary goal of our innovative *PoemTale Diffusion* framework is to ensure the main character remains consistent throughout the poem narration. To achieve this, we have developed a Main Character Identification (MCI) unit that precisely identifies the central character in each poem, ensuring coherence across all subsequent visual narratives. In our creative approach, we employed `GPT-4o-mini` to build the MCI component, marking the first instance of leveraging LLMs to precisely identify the main characters in the verses. In poetry, main characters, often subtly hinted at or indirectly described, play a crucial role in shaping the narrative and emotional impact of the poem. As demonstrated in Figure 3, the poem is processed through the MCI, which identifies and generates the main character. However, in some verses, the poets explore abstract emotions without specifying a tangible character.

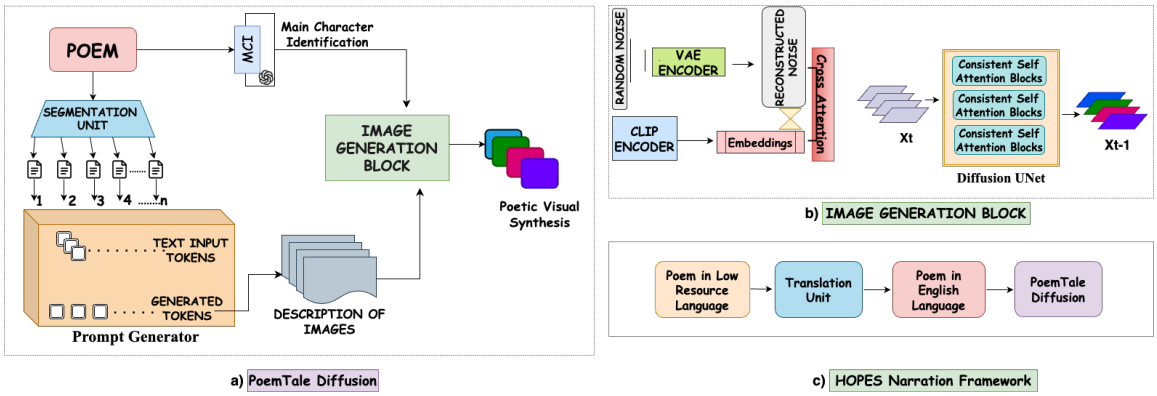

Figure 2: a) The *PoemTale Diffusion* pipeline for generating images that preserve the main character in the visual narration of poems. b) Image generation block takes image descriptions as input. c) Pipeline for Image Generation for the Hindi Language Poems *HOPES* by incorporating a Translation module.

To address these abstract instances, the MCI generates the central theme of the poem, which is then fed into the image generation block, resulting in descriptive visuals.

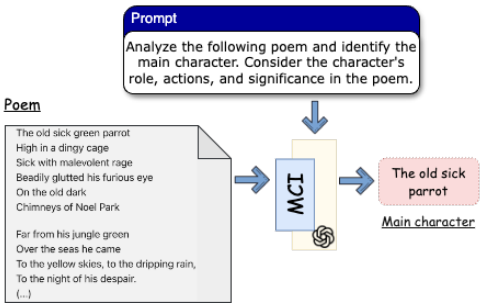

Figure 3: An instance of a poem passing through *Main Character Identification unit (MCI)* to generate main character of the poem.

4.2.3 Prompt Generator

The segmented parts of the poems are then processed through a prompt generator to achieve a detailed understanding of the verses. The direct input of poem verses into the image generation module resulted in irrelevant images due to the brevity and lack of descriptive content in poems. To produce relevant and coherent visuals, we developed a prompt generator that converts the poem segments into high-quality image descriptions in the form of prompts. Our prompt generator is inspired by the work of (Li et al., 2024), which converts user instructions into high-quality prompts for creating context-relevant images. This approach is illustrated in advanced image generation frameworks such as DALLE 3 (Betker et al., 2023), which uses VLMs' generation and understanding capabilities to achieve higher-quality text conditions for their generation tasks. As demonstrated in Figure 4, these detailed descriptions are then input into the image generation module, which further produces correlated and relevant images.

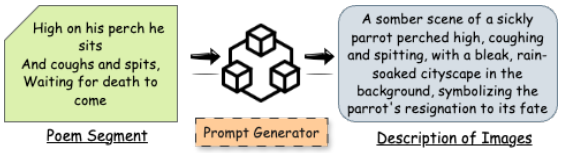

Figure 4: An instance of a poem segment passing through *Prompt Generator* to generate description of images.

4.2.4 Image Generation Block

This component ensures main character consistency by transforming textual descriptions into a series of images. Multiple textual prompts $\mathcal{T} = [\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n]$ are fed into a CLIP encoder to generate text embeddings, $E = [e_1, e_2, \ldots, e_n]$. A variational auto-encoder (VAE) then generates an initial latent representation from random noise $z_0 \sim \mathcal{N}(0, I)$. The cross-attention mechanism combines these text embeddings with the VAE encoder output:

$$Q = W_Q z_0, \quad K = W_K E, \quad V = W_V E$$

where W_Q, W_K, and W_V are learnable weight matrices. Here, Q is the query derived from the VAE encoder output, while K and V are the key and value derived from the text embeddings. The cross-attended output is then passed through a Diffusion UNet to generate the final images. The image generation block processes image descriptions to produce embeddings X_t, which are subsequently fed

into the Diffusion UNet pipeline to generate coherent images X_{t-1}. Our Diffusion UNet is inspired by (Zhou et al., 2024), which introduced the Consistent Self-Attention mechanism to maintain image consistency. Motivated by the approach in (Zhou et al., 2024), we replaced the self-attention layer in the existing UNet image generation model with Consistent Self-Attention, leveraging the original self-attention weights to ensure a training-free and pluggable implementation.

4.2.5 Translation Module

The *PoemTale Diffusion* model has proven its efficacy in generating visual narrations for English language poems. To extend the model's capabilities to the *HOPES* dataset, which includes Hindi poems, we introduced a translation unit pipeline, as illustrated in Figure 2(c). We employed IndicTrans2 (Gala et al., 2023), a translation model capable of supporting 22 diverse Indian languages. This adaptation was essential because the original *PoemTale Diffusion* components are not designed to process languages such as Hindi.

5 Experiments Results and Analysis

5.1 Implementation Details

To generate multiple images, we applied our methodology across various baseline models, including SDXL Base (Podell et al., 2023), SDXL Turbo (Sauer et al., 2023), and Story Diffusion (Zhou et al., 2024). We benchmarked our introduced *HOPES* dataset on these baseline models and *PoemTale Diffusion* framework. We identified two established quantitative metrics for image-text matching in this evaluation: Image-Text Matching Loss (ITM) (Li et al., 2022) and Image-Text Contrastive Loss (ITC) (Li et al., 2022), along with our newly introduced metric, the *Visual Narration Score (VNS)*.

While Image-Text Matching Loss (ITM) and Image-Text Contrastive Loss (ITC) are crucial for evaluating feature alignment and classification of image-text pairs, they have limitations regarding narrative coherence. ITC focuses on the similarity of feature vectors but does not address narrative continuity or visual consistency. ITM evaluates the correctness of image-text pairs individually without considering how well the images support the broader narrative. Both metrics primarily assess individual pairs and feature alignment but fail to ensure that images consistently enhance the narrative across multiple segments. *The VNS addresses these limitations by integrating ITM scores with pixel ratio consistency into a unified metric, providing a comprehensive evaluation of narrative coherence and visual alignment across the entire text as demonstrated in Algorithm 1. The average combined pixel ratio serves as a comprehensive indicator, measuring both the visual coherence of individual images and their adherence to the corresponding text prompts. A detailed description of the algorithm is included in the Appendix C.*

Algorithm 1: Visual Narration Score (VNS)

```
for i, image in enumerate(images_llm_prompt):
    over_all_combined_ratios = []
    for i in range(len(itm_scores)):
        combined_pixel_ratio = itm_scores[i] / (1 + abs(average_ratio - ratio_of_foreground_background[i]))
        over_all_combined_ratios.append(combined_pixel_ratio)
    average_combined_ratio = sum(over_all_combined_ratios) / len(over_all_combined_ratios)
    print(f"Average combined pixel ratio : {average_combined_ratio}")
```

5.2 Qualitative Evaluation:

Figure 5 presents a detailed comparison of images produced by different models across two datasets, focusing on three key aspects:

(1) Character Consistency: The *PoemTale Diffusion* model demonstrates exceptional ability in maintaining the main character's consistency across all image segments. Similarly, the Story Diffusion model also preserves character consistency. However, SDXL Base and SDXL Turbo face difficulties in preserving the main character, resulting in the inclusion of irrelevant characters across image segments.

(2) Context Shifts: The *PoemTale Diffusion* model effectively captures context changes within poem segments, as illustrated in Figure 5(a). For instance, it successfully represents the transition in emotion from sadness in segment 2 to hopefulness in segment 3. In contrast, other models, except for SDXL Turbo, fail to adapt to these context changes, producing images that do not align with the evolving emotional tone of the poem's narrative.

(3) Image Coherence: Figure 5(a) and Figure 5(b) show that the *PoemTale Diffusion* model maintains strong image coherence by consistently aligning

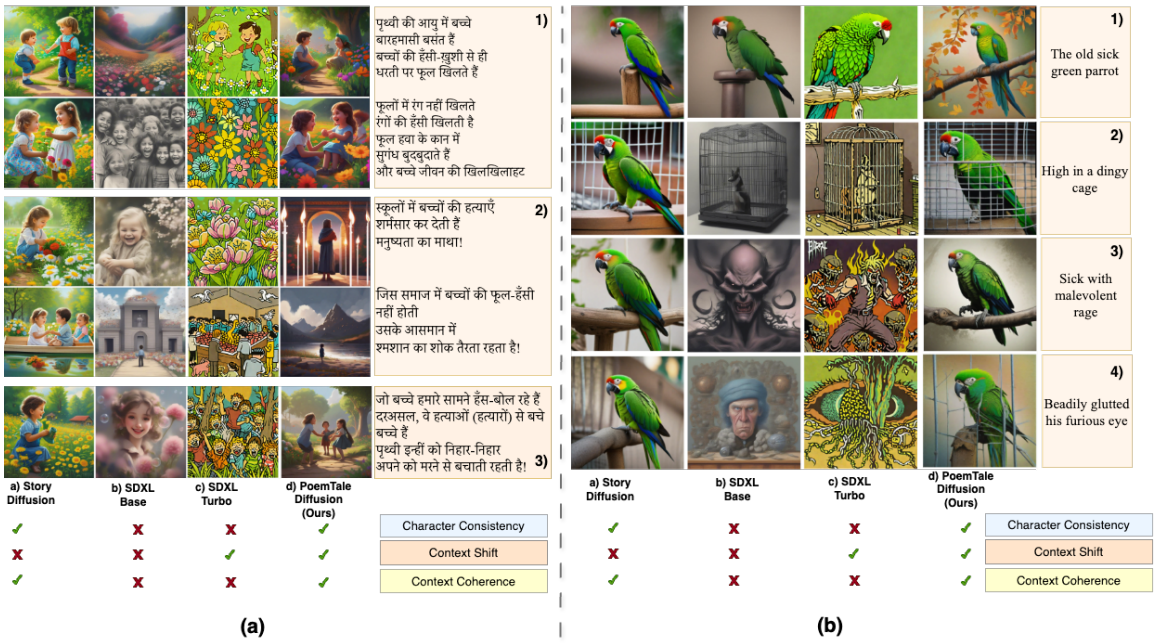

Figure 5: A comparison of *PoemTale Diffusion* with recent state-of-the-art image generation models on (a) *HOPES* dataset (a collection of poems in Indian languages) and (b) *PoemSum* dataset.

background and foreground elements across all segments. In contrast, SDXL Base and SDXL Turbo demonstrate a lack of coherence, characterized by inconsistent background and foreground elements in each segment, resulting in fragmented visual storytelling. Furthermore, for the *TinyRhymes dataset*, *PoemTale Diffusion* performs consistently well in character consistency, context shifts, and image coherence. As demonstrated in Figure 6, the images generated for nursery rhymes maintain character consistency throughout, capture context shifts effectively, and ensure visual coherence across all segments. This evaluation was conducted to assess our framework's performance across different genres of poems.

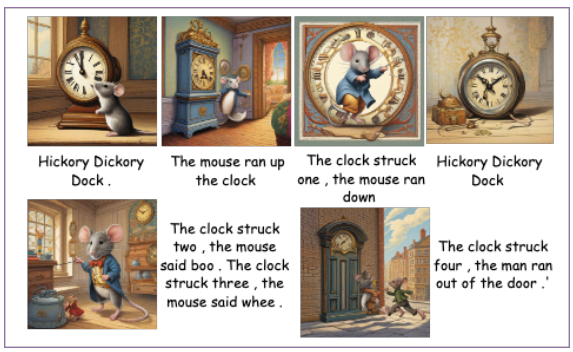

Figure 6: Visual Narration of Poems for *TinyRhymes* Dataset (Collection of Nursery Rhymes).

5.3 Human Evaluation

Recognizing that existing metrics may not fully capture the quality of the generated images, particularly aspects like coherence, relevance, and aesthetic appeal, and with no standardized metric available for assessing consistency within visual stories, we incorporated human evaluations. This evaluation focuses on capturing context shifts, the interpretability of the poem's meaning, and character consistency. To conduct this evaluation, we collaborated with four area experts from the Poetry Society of the Nation, including one renowned poet, two university professors, and a PhD in visual and literary arts (none of whom are affiliated with the authors). We selected 5% of the samples from the *PoemSum* dataset. The experts reviewed the images produced by our *PoemTale Diffusion* model and other baseline models. For each selected visual story, participants were asked to rate the samples on a scale from 1 to 5, considering the criteria of image-poem alignment, interpretability of meaning, and character consistency, with higher scores reflecting better quality. The human evaluation results, as shown in Table 2, indicate that our *PoemTale Diffusion* model achieves superior performance in overall scores, particularly in terms of consistency and quality. This demonstrates the model's capability to generate coherent image sequences that are well-aligned with the given poem segments.

Baselines	*Character Consistency*	*Context Shifts*	*Context Coherence*
SDXL Base	1.75	1.23	2.4
SDXL Turbo	1.07	1.09	2.21
Story Diffusion	3.7	3.87	4.11
Intelligent Grimm	3.65	3.7	3.92
PoemTale Diffusion	**3.89**	**3.92**	**4.31**

Table 2: Comparison results of human evaluation.

5.4 Quantitative Evaluation

The results detailed in Table 3 offer important insights into the performance of various image generation models, highlighting the effectiveness of different approaches within the context of our task. The following key findings and trends are observed: **(1) *PoemTale Diffusion* Performance:** Across all image generation models, our *PoemTale Diffusion* framework consistently outperforms others on two datasets, demonstrating its robustness in generating images that align closely with the descriptions provided by our prompt generator. The superior ITM and ITC scores highlight *PoemTale Diffusion*'s ability to accurately interpret textual descriptions and translate them into corresponding visual illustrations. Additionally, the strong VNS scores indicate that the framework maintains consistent pixel ratios and consistently high ITM scores across all poem segments. This results in visually coherent illustrations that effectively narrate the entire poetic content. **(2) Performance Comparison of Different Models** Among the evaluated image generation models, SDXL Turbo outperforms both SDXL Base and Story Diffusion, achieving high ITM and ITC scores across most datasets. This suggests that SDXL Turbo is highly effective at aligning individual images with their corresponding text descriptions. However, its VNS score indicates challenges in maintaining visual consistency across narrative segments. In contrast, Story Diffusion consistently underperforms across ITM, ITC, and VNS metrics. This underperformance is likely due to its focus on character consistency, which, while beneficial, often neglects other textual details critical for ensuring coherence across image segments. This aspect is particularly crucial for poetry, which often involves hidden and abstract meanings that are vital for ensuring overall narrative consistency. **(3) Dataset-Based Performance** We extensively evaluated various image generation models across three datasets of varying genres, languages, and lengths. Our analysis portrays that the *PoemTale Diffusion* framework performs well with the *PoemSum* and *HOPES datasets*, demonstrating its advanced ability to generate images for both English poems and Hindi-to-English translations. For the *TinyRhymes* dataset, which consists of nursery rhymes, the *PoemTale Diffusion* framework achieves comparable results to the SDXL Base model, which is the highest-performing model in terms of ITM and ITC scores, with values of 0.643 and 0.370, respectively. Despite this, *PoemTale Diffusion* demonstrates exceptional performance in VNS scores across all datasets, highlighting its ability to guarantee visual coherence. The overall improved performance of models on the TinyRhymes dataset is due to its simplicity and the absence of complex, hidden meanings in the poems, which eases the image generation process. In contrast, the *PoemSum* and *HOPES* datasets, with their rich emotional and detailed content, pose significant challenges that other models fail to address effectively. *PoemTale Diffusion*, however, showcases its robust capability to align images with text and ensure visual consistency across these intricate datasets.

		Image Generation Models			
Datasets ↓	**Metrics ↓**	*SDXL Base*	*SDXL Turbo*	*Story Diffusion*	*PoemTale Diffusion*
PoemSum	ITM	0.212	0.501	0.135	**0.504**
	ITC	0.361	0.363	0.237	**0.369**
	VNS	0.145	0.279	0.111	**0.399**
HOPES	ITM	0.678	0.741	0.121	**0.986**
	ITC	0.342	0.376	0.263	**0.413**
	VNS	0.511	0.359	0.099	**0.838**
TinyRhymes	ITM	**0.643**	0.593	0.557	0.612
	ITC	**0.370**	0.388	0.355	0.360
	VNS	0.489	0.431	0.449	**0.493**

Table 3: Quantitative Comparisons of Image Generation Models for Poem Narration

6 Conclusion and Future Work

To sum it all up, *PoemTale Diffusion* represents a significant advancement in the domain of visual narration for poetry. Our framework demonstrates the ability to comprehend the inherent and complex meanings within poems and generate multiple images that form a cohesive visual illustration. Additionally, our Entity and Emotion Segmentation algorithm effectively captures context shifts within poems, both in terms of entities and emotions. To evaluate our framework's performance across different languages, we introduced a new dataset, *HOPES*, in Hindi language and conducted the image generation process. Moreover, our proposed VNS metric offers a valuable tool for assessing the coherence and narrative quality of the images generated for the segmented parts of poems. Looking ahead, we aim to develop a language-agnostic visual model capable of performing visual narrations across a variety of languages in poetry.

7 Limitations

While *PoemTale Diffusion* effectively narrates poems visually, it currently lacks language-agnostic capabilities. Despite the inclusion of a translation module designed to assess the framework's robustness, it still generates images exclusively for English poems and does not support other languages.

References

Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. 2022. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9.

Rajat Agarwal and Katharina Kann. 2020. Acrostic poem generation. *arXiv preprint arXiv:2010.02239*.

Alex Andonian, Sabrina Osmany, Audrey Cui, Yeon-Hwan Park, Ali Jahanian, Antonio Torralba, and David Bau. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

Omri Avrahami, Ohad Fried, and Dani Lischinski. 2023. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*.

Wen-Feng Cheng, Chao-Chung Wu, Ruihua Song, Jianlong Fu, Xing Xie, and Jian-Yun Nie. 2018. Image inspired poetry generation in xiaoice. *arXiv preprint arXiv:1808.03090*.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022a. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022b. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1183–1191.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. 2024. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899*.

Triet Minh Huynh and Quan Le Bao. 2024. Vietnamese poem generation & the prospect of cross-language poem-to-poem translation. *arXiv preprint arXiv:2401.01078*.

Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. 2023. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint arXiv:2302.03900*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *Preprint*, arXiv:2403.18814.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6190–6200.

Zhiqiang Liu, Zuohui Fu, Jie Cao, Gerard De Melo, Yik-Cheung Tam, Cheng Niu, and Jie Zhou. 2019. Rhetorically controlled encoder-decoder for modern chinese poetry generation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 1992–2001.

Kai-Ling Lo, Rami Ariss, and Philipp Kurz. 2022. Gpoet-2: A gpt-2 based poem generator. *arXiv preprint arXiv:2205.08847*.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14878–14886, Singapore. Association for Computational Linguistics.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. *Preprint*, arXiv:1501.02530.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Ashwathy T Revi, Stuart Middleton, and David Millard. 2022. Idn-sum: A new dataset for interactive digital narrative extractive text summarisation.

Denys Thompson. 1978. *The uses of poetry*. Cambridge University Press.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265.

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. 2013. I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2197–2203. AAAI Press.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. 2024. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 670–680.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*.

Yukun Zhu. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *arXiv preprint arXiv:1506.06724*.

A Image Generation Workflow for English Poems in the *PoemSum* Dataset via *PoemTale Diffusion*

In this section, we outline the complete image generation process, illustrating how the poem is processed through various components of the *PoemTale Diffusion* framework, ultimately resulting in the creation of multiple image segments. Our framework is composed of four key components: Main Character Identification, Segmentation Unit, Prompt Generation, and Image Generation Block.

A.1 Main Character Identification (MCI)

We are processing the poems through the MCI Block to identify the main character, ensuring consistency in the subsequent image generation. As shown in Table 4, the main characters have been identified for the English and Hindi poems, respectively.

Poems	*Main Character Identification(MCI)*
A fortnight before Christmas Gypsies were everywhere: Vans were drawn up on wastes, women trailed to the fair. 'My gentleman,' said one, 'you've got a lucky face.' 'And you've a luckier,' I thought, 'if such a grace And impudence in rags are lucky.' 'Give a penny For the poor baby's sake.' 'Indeed I have not any Unless you can give change for a sovereign, my dear.' 'Then just half a pipeful of tobacco can you spare?' I gave it. With that much victory she laughed content. I should have given more, but off and away she went With her baby and her pink sham flowers to rejoin The rest before I could translate to its proper coin Gratitude for her grace. And I paid nothing then, As I pay nothing now with the dipping of my pen For her brother's music when he drummed the tambourine And stamped his feet, which made the workmen passing grin, While his mouth-organ changed to a rascally Bacchanal dance 'Over the hills and far away.' This and his glance Outlasted all the fair, farmer, and auctioneer, Cheap-jack, balloon-man, drover with crooked stick, and steer, Pig, turkey, goose, and duck, Christmas Corpses to be. Not even the kneeling ox had eyes like the Romany. That night he peopled for me the hollow wooded land, More dark and wild than stormiest heavens, that I searched and scanned Like a ghost new-arrived. The gradations of the dark Were like an underworld of death, but for the spark In the Gypsy boy's black eyes as he played and stamped his tune, 'Over the hills and far away', and a crescent moon.	Gypsy boy

Table 4: Main Character Identification for Poems in the *PoemSum* Dataset Using the MCI Unit

A.2 *Entity Plus Emotion* Segmentation algorithm

The algorithm segments a poem into distinct parts by identifying changes in detected entities (such as people, organizations, or locations) and shifts in emotions throughout the text. It starts by breaking the poem down into individual sentences. For each sentence, entities are extracted and the emotion conveyed is classified. Segmentation occurs whenever there is a noticeable change in emotion or entities. If a sentence contains no new entities or if the entities remain consistent with the previous ones, it is added to the current segment. The algorithm finalizes and stores each segment whenever a change is detected. After processing the entire poem, it outputs the segmented sections along with the corresponding entities and emotions. This method allows for a nuanced analysis that captures both the thematic elements and the emotional dynamics of the poem, offering a richer understanding of its structure and meaning. The entire algorithm is demonstrated in Algorithm 2. As shown in Table 5, the results highlight the segmented parts of the poem, along with a detailed description of the entities and emotions that aided the segmentation.

A.3 Prompt Generator

Next, we use the Prompt Generation module to create segments descriptions for the same poem where the main character has been identified. As demonstrated in Table 6, these prompts are generated for each segment and then passed to the Image Generation module to produce multiple images corresponding to each segment.

A.4 Image Generation Block

The Image Generation block now utilizes the main character identified by the MCI unit (as detailed in Table 4) and the segment descriptions provided by the Prompt Generator (shown in Table 6) as inputs. This process results in the creation of images for various poem segments, forming a visually coherent narrative as demonstrated in Figure 7. As illustrated in Figure 7, these images correspond to the poem segments presented in Table 5.

B Image Generation Workflow for *Hindi(Indian Language)* Poems in the *HOPES* Dataset via *PoemTale Diffusion*

In this section, we provide a detailed overview of the image generation process for Indian language poems after they have been translated into English as demonstrated in Figure 8 and processed through the *PoemTale Diffusion* framework. This process results in the creation of multiple image segments. Our framework consists of four key components: Main Character Identification, Segmentation Unit, Prompt Generation, and the Image Generation Block.

B.1 Main Character Identification (MCI)

We then process the English-translated poems through the MCI Block to identify the main character. As illustrated in Figure 9, the main characters

Figure 7: A sample of images generated for Poems present in *PoemSum dataset*

Segments of Poems	*Entities Identified*	*Emotions Identified*
Segment1: "A fortnight before Christmas Gypsies were everywhere: Vans were drawn up on wastes, women trailed to the fair. 'My gentleman,' said one, 'you've got a lucky face.' 'And you've a luckier,' I thought, 'if such a grace And impudence in rags are lucky.'",	'PERSON': 'gypsy woman', 'ORGANIZATION': set(), 'LOCATION': set(),	'neutral',
Segment 2: "'Give a penny For the poor baby's sake.'",	'PERSON': 'poor baby', 'ORGANIZATION': set(), 'LOCATION': set(),	'sadness',
Segment 3: "'Indeed I have not any Unless you can give change for a sovereign, my dear.'",	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'neutral',
Segment 4: "'Then just half a pipeful of tobacco can you spare?' I gave it.",	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'joy',
Segment 5: 'With that much victory she laughed content.',	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'neutral',
Segment 6: 'I should have given more, but off and away she went With her baby and her pink sham flowers to rejoin The rest before I could translate to its proper coin Gratitude for her grace.',	'PERSON': 'she', 'ORGANIZATION': set(), 'LOCATION': set(),	'anger',
Segment 7: "And I paid nothing then, As I pay nothing now with the dipping of my pen For her brother's music when he drummed the tambourine And stamped his feet, which made the workmen passing grin, While his mouth-organ changed to a rascally Bacchanal dance "Over the hills and far away.'",	'PERSON': 'her brother', 'ORGANIZATION': set(), 'LOCATION': set(),	'sadness',
Segment 8: 'This and his glance Outlasted all the fair, farmer, and auctioneer, Cheap-jack, balloon-man, drover with crooked stick, and steer, Pig, turkey, goose, and duck, Christmas Corpses to be.',	'PERSON': 'farmer', 'auctioneer', 'cheap-jack', 'balloon-man', 'drover', 'ORGANIZATION': set(), 'LOCATION': set(),	'neutral',
Segment 9: 'Not even the kneeling ox had eyes like the Romany.',	'PERSON': 'Romany', 'ORGANIZATION': set(), 'LOCATION': set(),	
Segment 10: "That night he peopled for me the hollow wooded land, More dark and wild than stormiest heavens, that I searched and scanned Like a ghost new-arrived. The gradations of the darklike an underworld of death, but for the spark In the Gypsy boy's black eyes as he played and stamped his tune, 'Over the hills and far away', and a crescent moon.	'PERSON': 'Gypsy boy', 'ORGANIZATION': set(), 'LOCATION': 'hollow wooded land', 'stormiest heavens'	'fear', 'sadness'

Table 5: Poems Segmented on basis of shift in entities and shift of emotion using our *Entity Plus Emotion* Segmentation Algorithm for English Language Poems present in *Poemsum* dataset

Hindi(Indian Language) Poems	*Translated to English Poems*
"हर सुबह हज़ारों मील दूर से चली आती हैं तुम्हारी बौद्ध प्रार्थनाएँ हवा में तैरते हुए तुम्हारी हँसी महानगर के भयानक शोर में भी आत्मा की आवाज़ की तरह साफ़-साफ़ सुनाई देती है तुम मुस्कुराते हो तो तुम्हारे चेहरों के प्रतिबिंब चाँद के दर्पण में पूरी पूरी रात दमकते हैं जब भी छूते हो कहीं तुम कोई दरख़्त पृथ्वी पर पेड़, पानी और मनुष्य बढ़ जाते हैं। ड्राम्त्से गटसल चिल्ड्रेंस कम्यनिटी, लमला, तवांग के बच्चों के लिए"	Every morning comes from thousands of miles away, your Buddhist prayers float in the air, your laughter is as clear as the sound of the soul even in the terrible noise of the metropolis. You smile, then the reflections of your faces glow all night in the mirror of the moon. Whenever you touch a tree, trees, water and humans grow on the earth. For the children of Zamtse Gatsal Children's Community, Lumla, Tawang.
(दस-ग्यारह साल की एक लड़की) गोल हैं ख़ूब मगर आप तिरछे नज़र आते हैं ज़रा। आप पहने हुए हैं कुल आकाशतारों - जड़ा; सिर्फ़ मुँह खोले हुए हैं अपनागोरा - चिट्टागोल - मटोल, अपनी पोशाक को फैलाए हुए चारों सिम्त। आप कुछ तिरछे नज़र आते हैं जाने कैसे—ख़ूब हैं गोकि! वाह जी, वाह! हमको बुद्ध ही निरा समझा है! हम समझते ही नहीं जैसे कि आपको बीमारी है : आप घटते हैं तो घटते ही चले जाते हैं, और बढ़ते हैं तो बस यानी कि बढ़ते ही चले जाते हैं— हम नहीं लेते हैं जब तक बिलकुल ही गोल न हो जाएँ, बिलकुल गोल।यह मर्ज़ आपका अच्छा ही नहीं होने में	(A girl of ten-eleven years) You are round, but you look skewed. You are wearing a total star-studded dress; only your mouth is open. Apangora-Chittagol-Matol, spreading your dress around. You look a little skewed. You know how nice! Wow, wow! We have understood the Buddha! We don't understand as if you have a disease: if you decrease, you decrease, and if you increase, you only increase. We don't take goals at all. This is not good for you.

Figure 8: Hindi Poems from the *HOPES* Dataset Translated to English using *IndicTrans2* Translation Module

English Translated Poems	*Main Character Identification(MCI)*
Every morning comes from thousands of miles away, your Buddhist prayers float in the air, your laughter is as clear as the sound of the soul even in the terrible noise of the metropolis. You smile, then the reflections of your faces glow all night in the mirror of the moon. Whenever you touch a tree, trees, water and humans grow on the earth. For the children of Zamtse Gatsal Children's Community, Lumla, Tawang.	Buddhist Monk
(A girl of ten-eleven years) You are round, but you look skewed. You are wearing a total star-studded dress; only your mouth is open. Apangora-Chittagol-Matol, spreading your dress around. You look a little skewed. You know how nice! Wow, wow! We have understood the Buddha! We don't understand as if you have a disease: if you decrease, you decrease, and if you increase, you only increase. We don't take goals at all. This is not good for you.	girl (ten-eleven years)

Figure 9: Main Character identified for Poems present in *HOPES* dataset

Segments of Poems	*Prompts Generated for Segments*
Segment1: "A fortnight before Christmas Gypsies were everywhere: Vans were drawn up on wastes, women trailed to the fair. 'My gentleman,' said one, 'you've got a lucky face.' 'And you've a luckier,' I thought, 'if such a grace And impudence in rags are lucky.'",	Prompt 1: ' A rustic scene with vintage caravans parked on a rural wasteland, women in period clothing milling about, evoking a sense of festive anticipation.
Segment 2: "'Give a penny For the poor baby's sake.'",	Prompt 2:' A heartwarming illustration of a child reaching out to a passerby with a small, hopeful gesture, set against a backdrop of a bustling city street.
Segment 3: "'Indeed I have not any Unless you can give change for a sovereign, my dear.'",	Prompt 3:' A vintage British coin, a sovereign, prominently displayed on a richly textured background, evoking a sense of history and nostalgia,
Segment 4: "'Then just half a pipeful of tobacco can you spare?' I gave it.",	Prompt 4: ' A vintage pipe filled with rich, dark tobacco, resting on a textured surface, surrounded by soft, warm lighting.
Segment 5: 'With that much victory she laughed content.',	Prompt 5:' A joyful woman with a radiant smile, laughing heartily in a warm, inviting setting.
Segment 6: 'I should have given more, but off and away she went With her baby and her pink sham flowers to rejoin The rest before I could translate to its proper coin Gratitude for her grace.',	Prompt 6:' A whimsical scene of a woman with a baby and a bouquet of vibrant, pink flowers, set against a backdrop of a bustling marketplace.
Segment 7: "And I paid nothing then, As I pay nothing now with the dipping of my pen For her brother's music when he drummed the tambourine And stamped his feet, which made the workmen passing grin, While his mouth-organ changed to a rascally Bacchanal dance "Over the hills and far away.'",	Prompt 7:' A lively scene of a musician energetically playing a tambourine and stamping his feet, surrounded by a crowd of amused onlookers.
Segment 8: 'This and his glance Outlasted all the fair, farmer, and auctioneer, Cheap-jack, balloon-man, drover with crooked stick, and steer, Pig, turkey, goose, and duck, Christmas Corpses to be.',	Prompt 8:' A whimsical scene of a festive fair, filled with a variety of characters and activities, set against a backdrop of a snowy landscape.
Segment 9: 'Not even the kneeling ox had eyes like the Romany.',	Prompt 9:' A serene and majestic ox kneeling in a snowy field,
Segment 10: "That night he peopled for me the hollow wooded land, More dark and wild than stormiest heavens, that I searched and scanned Like a ghost new-arrived. The gradations of the darklike an underworld of death, but for the spark In the Gypsy boy's black eyes as he played and stamped his tune, 'Over the hills and far away', and a crescent moon.	its eyes reflecting a sense of calm and strength. Prompt 10:' A mysterious and haunting forest at night, filled with dark and ominous trees, illuminated by a faint crescent moon.

Table 6: Description of Segments Generated by the Prompt Generator for Poems in the *PoemSum* Dataset

for the Hindi poems have been successfully identified.

B.2 *Entity Plus Emotion* Segmentation algorithm

We similarly apply the *Entity plus Emotion* Segmentation algorithm to the English-translated poems. As shown in Table 7 the results highlight the segmented parts of the poem, along with a detailed description of the entities and emotions that aided the segmentation.

B.3 Prompt Generator

Next, we use the Prompt Generation module to create segment descriptions for the same Indian Language poem where the main character has been identified. As demonstrated in Table 8, these prompts are generated for each segment and then passed to the Image Generation module.

B.4 Image Generation Block for *HOPES* dataset

The Image Generation block now utilizes the main character identified by the MCI unit (as detailed in Figure 9) and the segment descriptions provided by the Prompt Generator (shown in Table 8) as inputs. This process results in the creation of images for various poem segments, forming a visually coherent narrative as demonstrated in Figure 10. As illustrated in Figure 10, these images correspond to the poem segments presented in Table 7.

C Visual Narration Score(VNS) Explanation

This algorithm processes a series of images by first loading them and then evaluating their content using a combination of image analysis and machine learning techniques. It starts by translating each image into grayscale and applies binarization to determine the foreground-to-background ratio. This metric, alongside an average ratio computed across all images, provides insight into the relative prominence of foreground elements. The algorithm then computes an Image-Text Matching (ITM) score for each image, which measures how well the image aligns with a corresponding text prompt. By integrating these ITM scores with the foreground-background ratios, the algorithm calculates a combined pixel ratio for each image, offering a comprehensive evaluation. The detailed process is highlighted in Algorithm 3.

Segments of Poems	*Entities Identified*	*Emotions Identified*
[Segment 1: 'Every morning comes from thousands of miles away, your Buddhist prayers float in the air, your laughter is as clear as the sound of the soul even in the terrible noise of the metropolis.',	['PERSON': set(), 'ORGANIZATION': 'Buddhist', 'LOCATION': set(),	'fear',
Segment 2: 'You smile, then the reflections of your faces glow all night in the mirror of the moon.',	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'joy',
Segment 3: 'Whenever you touch a tree, trees, water and humans grow on the earth.',	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'disgust',
Segment 4: "For the children of Zamtse Gatsal Children's Community, Lumla, Tawang."]	'PERSON': set(), 'ORGANIZATION': 'Zamtse', 'Children', 'Community', 'Gatsal', "'s", 'LOCATION': 'Lumla', 'Tawang']	'neutral'
Segment1: '(A girl of ten-eleven years) You are round, but you look skewed. You are wearing a total star-studded dress; only your mouth is open.	['PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'neutral', 'disgust',
Segment 2: Apangora-Chittagol-Matol, spreading your dress around. You look a little skewed.	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'joy',
Segment 3: You know how nice!',	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	
Segment 4: 'Wow, wow!', '	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	'surprise',
Segment 5: We have understood the Buddha!	'PERSON': 'Buddha', 'ORGANIZATION': set(), 'LOCATION': set(),	'anger', '
Segment 6: We don't understand as if you have a disease: if you decrease, you decrease, and if you increase, you only increase. We don't take goals at all	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set(),	neutral',
Segment 7: This is not good for you.	'PERSON': set(), 'ORGANIZATION': set(), 'LOCATION': set()]	'sadness'

Table 7: Poems Segmented on basis of shift in Entities and shift of Emotion using our *Entity Plus Emotion* Segmentation Algorithm for Indian Language Poems present in *HOPES* dataset

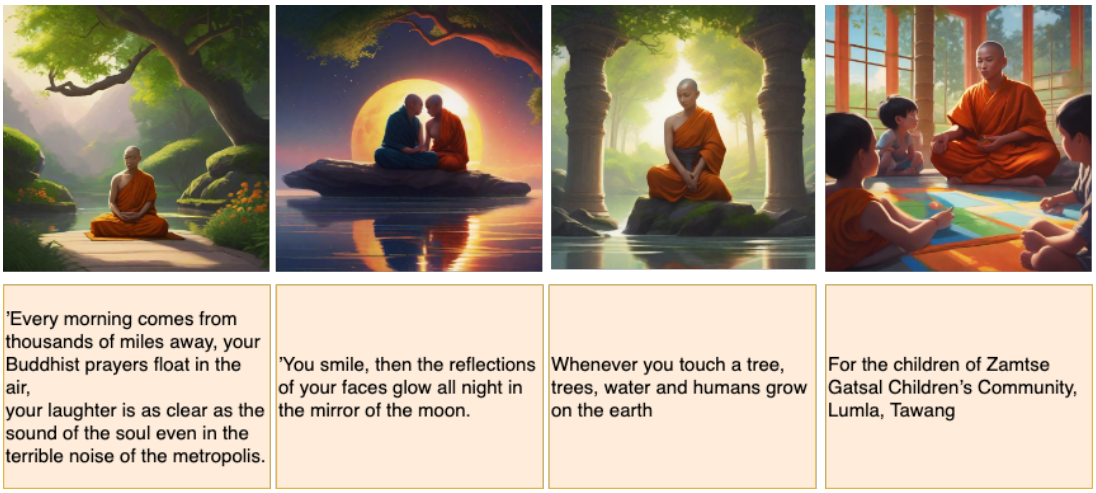

Figure 10: A sample of images generated for Indian Language Poems present in *HOPES* dataset

D Additional Resources added in the Supplementary Files

- Source Code for the proposed Visual Narration Score (VNS) evaluation metric.
- Source Code for the proposed Entity Plus Emotion Segmentation Algorithm.
- Sample dataset from the *PoemSum* dataset, including image paths with associated image segments.
- Sample dataset from the *HOPES* dataset, including image paths with associated image segments.

Segments of Poems	*Prompts Generated for Segments*
[Segment 1: 'Every morning comes from thousands of miles away, your Buddhist prayers float in the air, your laughter is as clear as the sound of the soul even in the terrible noise of the metropolis.',	Prompt 1: A serene monk meditating in a tranquil garden, surrounded by nature, embodying peace and tranquility, digital painting, art by greg rutkowski ',
Segment 2: 'You smile, then the reflections of your faces glow all night in the mirror of the moon.',	Prompt 2: ' A couple sharing a romantic moment under the moonlight, their reflections shimmering on the water,'
Segment 3: 'Whenever you touch a tree, trees, water and humans grow on the earth.',	Prompt 3: ' A person with a gentle touch, surrounded by a lush forest, with trees and water sources flourishing ',
Segment 4: "For the children of Zamtse Gatsal Children's Community, Lumla, Tawang."]	Prompt 4:' A group of children playing and learning together in a vibrant community center, surrounded by nature.']
Segment1: '(A girl of ten-eleven years) You are round, but you look skewed. You are wearing a total star-studded dress; only your mouth is open.	Prompt 1:' A whimsical illustration of a young girl in a star-studded dress, with a curious expression and her mouth slightly open, set against a dreamy backdrop ',
Segment 2: Apangora-Chittagol-Matol, spreading your dress around. You look a little skewed.	Prompt 2:' A surreal painting of a girl with a flowing dress, surrounded by abstract shapes and colors, evoking a sense of mystery and enchantment ',
Segment 3: You know how nice!',	Prompt 3:"A close-up of a girl's face, with a warm and inviting smile, set against a soft-focus background, conveying a sense of friendliness and approachability ",
Segment 4: 'Wow, wow!', '	Prompt 4:' A dynamic illustration of a girl with a star-studded dress, her mouth open in surprise, surrounded by swirling colors and shapes, creating a sense of excitement and wonder ',
Segment 5: We have understood the Buddha!	Prompt 5:'A serene painting of a girl in a star-studded dress, sitting in a meditative pose with a peaceful expression, set against a tranquil background ',
Segment 6: We don't understand as if you have a disease: if you decrease, you decrease, and if you increase, you only increase. We don't take goals at all	Prompt 6:'A thoughtful illustration of a girl in a star-studded dress, her expression pensive and introspective, set against a backdrop of swirling, dreamlike patterns ',
Segment 7: This is not good for you.	Prompt 7:' A poignant painting of a girl in a star-studded dress, her expression somber and reflective, set against a dark, moody background ']

Table 8: Description of Segments generated using Prompt Generator for Poems present in *HOPES* dataset

Algorithm 2: Algorithm for Segmenting Poem by Entities and Emotion

```python
def segment_poem_by_entities_and_emotion(poem):
    sentences = sent_tokenize(poem)

    segments = []
    segment_entities = []
    segment_emotions = []
    current_segment = []
    prev_entities = {"PERSON": set(), "ORGANIZATION": set(), "LOCATION": set()}
    prev_emotion = None

    for sentence in sentences:
        sentence = truncate_sentence(sentence)
        tagged_sentence, current_entities = extract_entities(sentence)
        emotion = emotion_classifier(sentence)[0]
        current_emotion = emotion['label']

        print(f"Sentence: {sentence}")
        print(f"NER Tags: {tagged_sentence}")
        print(f"Emotion: {current_emotion} (Score: {emotion['score']:.2f})")
        print()

        if current_emotion != prev_emotion:
            if current_segment:
                segments.append(" ".join(current_segment))
                segment_entities.append(prev_entities)
                segment_emotions.append(prev_emotion)
            current_segment = [sentence]
            prev_entities = current_entities
            prev_emotion = current_emotion
        else:
            if (not current_entities["PERSON"] and not current_entities["ORGANIZATION"] and not current_entities["LOCATION"]):
                current_segment.append(sentence)
            elif ((current_entities["PERSON"] and current_entities["PERSON"] != prev_entities["PERSON"]) or
                  (current_entities["ORGANIZATION"] and current_entities["ORGANIZATION"] != prev_entities["ORGANIZATION"]) or
                  (current_entities["LOCATION"] and current_entities["LOCATION"] != prev_entities["LOCATION"])):
                if current_segment:
                    segments.append(" ".join(current_segment))
                    segment_entities.append(prev_entities)
                    segment_emotions.append(prev_emotion)
                current_segment = [sentence]
                prev_entities = current_entities
                prev_emotion = current_emotion
            else:
                current_segment.append(sentence)

    if current_segment:
        segments.append(" ".join(current_segment))
        segment_entities.append(prev_entities)
        segment_emotions.append(prev_emotion)

    print(segments)
    print()
    print(segment_entities)
    print()
    print(segment_emotions)

    return segments, segment_entities, segment_emotions
```

Algorithm 3: Entire Algorithm for Visual Narration Score

```
for i, image in enumerate(images_llm_prompt):
    image = np.array(image)
    image = cv2.cvtColor(image, cv2.COLOR_RGB2GRAY)

    _, thresholded = cv2.threshold(image, 0, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)

    # Calculate the foreground and background ratios
    foreground_ratio = np.sum(thresholded == 255) / thresholded.size
    background_ratio = np.sum(thresholded == 0) / thresholded.size

    print(f"Foreground Ratio of image {i+1}: {foreground_ratio}")
    print(f"Background Ratio of image {i+1}: {background_ratio}")
    k = foreground_ratio / background_ratio
    print(f"Foreground/Background ratio of image {i+1}: {k}")
    ratio_of_foreground_background.append(k)

average_ratio = sum(ratio_of_foreground_background) / len(ratio_of_foreground_background)
print(f"Average Foreground/Background ratio: {average_ratio}")

itm_scores = []
for i in range(len(images_llm_prompt)):
    ans = model(
        {"image": vis_processors["eval"](images_llm_prompt[i]).unsqueeze(0).to(device),
         "text_input": text_processors["eval"](llm_prompts[i])},
        match_head="itm"
    )
    ans = torch.nn.functional.softmax(ans, dim=1)
    ans = ans[:, 1].item()
    itm_scores.append(ans)
    print(f"ITM score of image {i+1}: {ans}")

over_all_combined_ratios = []
for i in range(len(itm_scores)):
    combined_pixel_ratio = itm_scores[i] / (1 + abs(average_ratio - ratio_of_foreground_background[i]))
    over_all_combined_ratios.append(combined_pixel_ratio)
    print(f"Combined pixel ratio of image {i+1}: {combined_pixel_ratio}")

average_combined_ratio = sum(over_all_combined_ratios) / len(over_all_combined_ratios)
print(f"Average combined pixel ratio: {average_combined_ratio}")
```