RESEARCH ARTICLE

Capturing Poetic Essence: Text Summarization and Visual Generation via Multimodal

Junaid Yousaf[1,*], Mazhar Iqbal[3], Iqra Pervaiz[2,*], Muhammad Ismail[4], Toqeer Ul Islam[5] and Khurram Khan Jadoon[1]

[1] Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi 23460, Pakistan
[2] Faculty of Computer Science, CECOS University of Information Technology and Emerging Sciences, Peshawar 25000, Pakistan
[3] Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan
[4] School of Information Technology, Deakin University, Geelong, Victoria 3220, Australia
[5] School of Computing and Digital Technology, Birmingham City University, West Midlands B5 5JU, United Kingdom

Abstract

Poetry, as a profound and creative form of human expression, presents unique challenges in interpretation and summarization due to its reliance on figurative language, symbolism, and deeper meanings. Building upon the PoemSum dataset, which introduced the task of poem summarization, we extend its scope by exploring multimodal applications. Specifically, we implement and fine-tune two state-of-the-art abstractive summarization models—BART and T5—to generate concise and meaningful interpretations of poems, focusing on figurative summarization that captures metaphorical and symbolic elements inherent in poetic language. These summaries are then transformed into visual representations using two diffusion-based generative models: Stable Diffusion for high-quality image generation. Our approach evaluates the effectiveness of abstractive summarization models in capturing the essence of poetry and demonstrates how diffusion models can translate abstract poetic themes into visually compelling images. Evaluation results show that the BART model outperforms T5 in summarization, achieving a high ROUGE score of 41.90 and a BERTScore of 85.22. For image generation, the Inception Score (IS) of 7.63 ± 0.62 reflects high visual quality and diversity, while the CLIP (Contrastive Language-Image Pre-training) Score of 29.48 indicates strong alignment between textual summaries and generated images.

Keywords: poetry, summarization, image, diffusion model, multimodal, Transformer and Bart.

Academic Editor:
Xuebo Jin

Submitted: 03 May 2025
Accepted: 05 June 2025
Published: 27 July 2025

Vol. 2, **No.** 3, 2025.
10.62762/TIS.2025.405393

***Corresponding authors:**
✉ Junaid Yousaf
Junaidyousaf432@gmail.com
✉ Iqra Pervaiz
pervaiziqra2000@gmail.com

1 Introduction

Poetry, as one of the most profound and creative forms of human expression, conveys emotions, thoughts, and ideas through intricate figurative

Citation
Yousaf, J., Iqbal, M., Pervaiz, I., Ismail, M., Islam, T. U., & Jadoon, K. K. (2025). Capturing Poetic Essence: Text Summarization and Visual Generation via Multimodal. *ICCK Transactions on Intelligent Systematics*, 2(3), 160–168.

language, vivid imagery, and metaphors. Unlike conventional prose, poetry often embodies layered meanings and abstract interpretations, making its understanding and summarization uniquely challenging. Capturing the essence of poetry requires more than a literal comprehension of its words; it demands an appreciation of its deeper themes, emotional resonance, and creative nuances. This complexity has historically posed significant challenges for computational models that attempt to engage with creative text.

Recent advances in natural language processing (NLP) have opened new avenues for tackling such creative tasks. The introduction of the PoemSum dataset marked a significant milestone in this domain by establishing a foundation for the poem summarization task. PoemSum demonstrated the potential of NLP models to engage in creative language interpretation, providing researchers with a benchmark dataset and highlighting the challenges of summarizing poetry as opposed to factual or prose-based texts [1]. However, while progress has been made in understanding poetic content, there remains an untapped opportunity to connect such textual interpretations to other modalities, such as visual representations, which can further enhance the accessibility and engagement with poetic works.

Building on this foundation, this study aims to extend the application of poem summarization to the multimodal domain, integrating textual and visual understanding. Following the implementation of the poem summarization task as outlined in the PoemSum study, this work uses the generated summaries to create visual representations through diffusion models, a cutting-edge approach in generative image synthesis [2]. By transforming the nuanced and abstract ideas conveyed in poetry into corresponding visual forms, this approach seeks to bridge the gap between textual creativity and visual imagination.

This integration of NLP with diffusion models for text-to-image generation represents a novel effort to capture the interplay of linguistic and visual creativity [3]. Through this work, poetic expressions are not only summarized, but also visually interpreted, enabling a deeper and more holistic engagement with poetry. For example, themes of love, nature, despair, or triumph encapsulated in poetic language can be vividly brought to life in images, offering readers a new way to connect with the underlying narratives and emotions.

In this paper, we propose a novel research direction that resolves significant limitations in earlier poetic summarization studies by integrating natural language processing with visual generation to construct a multimodal representation of poetry. While earlier work Mahbub et al. [1] (2023) introduced the PoemSum dataset and established baselines for text-only poem summarization, no earlier work has addressed how the poetic meaning, particularly its metaphorical and symbolic richness, can be expressed and perceived in visual modalities.

We bridge this gap by suggesting a two-stage pipeline: first, we employ transformer-based models (T5 and BART) to generate abstractive, metaphor-aware summaries of poetry; second, we use Stable Diffusion to translate these summaries into high-quality visualizations. As far as we know, this is the first instance of abstractive poetry summarization and diffusion-based image synthesis being paired, thereby enabling a new mode of AI-assisted figurative interpretation. As opposed to prior work, which lacked either multimodal understanding or preserved abstract meaning, our method demonstrates empirically that BART outperforms T5 and prior baselines at capturing semantic detail with a ROUGE-L score of 41.90 and a BERTScore of 85.22. To further rigorously test the text-image coherence, we also employ the Inception Score (7.63 ± 0.62) and CLIP Score (29.48), providing quantitative guarantees of visual-textual alignment a facet hitherto uncharted in aesthetic AI. Cumulatively, our contributions significantly advance the frontiers of multimodal AI by introducing a framework that not only innovates in textual summarization but also opens new vistas for visual interpretation of poetic themes, with material implications for pedagogy, digital humanities, and human-AI collaboration.

2 Problem Statement

Interpreting poetry requires a deep understanding of figurative language, metaphors, and abstract ideas. Despite this advance, NLP models are still unimpressive when extracting abstract information from actual text, especially poetry. Furthermore, converting the poetic interpretations into useful and appealing images is difficult, thus, reduces the interaction with the poetic contents. This research tackles two challenges: First inputting poetry into natural language processing models to produce contextually rich summaries of poetry, and Second generating visual representations of these summaries using diffusion models. Solving

these problems further promotes the development of creative understanding of language and allows for new uses in education, art, and multimodal AI systems.

Our work solves two key problems:

1. Summarizing poems while preserving metaphors (using NLP models like BART)
2. Converting summaries into matching images (using Stable Diffusion)

This connects text and visuals for better poetry understanding.

3 Related Work

Text summarization has been a prominent area of research in natural language processing (NLP), with approaches broadly categorized into extractive and abstractive methods. Extractive methods focus on identifying and extracting key points directly from the source text, while abstractive methods rephrase the content using novel word sequences to convey the main ideas. Abstractive summarization is particularly suited for tasks like poetry summarization, where direct extraction of words fails to capture the deeper, figurative meanings embedded in the text [4].

Mahbub et al. [1] introduced the PoemSum dataset and marks a significant step forward in addressing this gap. PoemSum is the first dataset specifically designed for poem summarization, comprising 3,011 poems paired with human-written summaries that capture the creative and abstract meanings of the texts. The authors benchmarked several state-of-the-art summarization models, including T5, BART, Pegasus, and mT5, and demonstrated the limitations of these models in capturing the deeper essence of poetry. Their findings highlighted the challenges of summarizing creative language and the need for specialized approaches to improve interpretive summarization tasks [1].

While the PoemSum study focused on textual summarization, the multimodal potential of poetry remains largely unexplored. Recent advancements in text-to-image generation using diffusion models have shown promise in transforming abstract textual descriptions into visual representations. Models such as DALL-E and Stable Diffusion have demonstrated the ability to generate high-quality images from text, but their application to poetic content, which requires a nuanced understanding of abstract and metaphorical language, remains underexplored [5].

This work builds upon the foundation laid by PoemSum by extending poem summarization to a multimodal domain. After implementing abstractive summarization models for poetry, we leverage the generated summaries to produce visual representations using a diffusion model. This integration of text summarization and image generation addresses the challenge of bridging textual creativity with visual imagination, contributing to the expanding research at the intersection of language and vision. By doing so, this study not only advances the understanding of creative language summarization but also demonstrates the potential of multimodal AI systems in engaging with and interpreting abstract artistic expressions.

4 Methodology

In this paper, we utilized the PoemSum dataset and applied preprocessing techniques to prepare the data for analysis.

Figure 1 illustrates our multimodal pipeline for poetic summarization and visualization. The process begins with the PoemSum dataset, where poems undergo preprocessing to remove noise and standardize the text. These cleaned poems are then fed into abstractive summarization models (BART and T5), which generate concise, metaphor-rich summaries evaluated using ROUGE and BERTScore metrics. The highest-quality summaries (from BART) serve as prompts for the Stable Diffusion model, which synthesizes corresponding images. Finally, the generated images are assessed for visual quality (Inception Score) and text-image alignment (CLIP Score). This pipeline bridges textual and visual creativity, ensuring both the abstract essence of poetry and its aesthetic translation are rigorously quantified.

4.1 Dataset

To enrich research on the task of poem summarisation, we have employed a dataset known as 'PoemSum' that is available on GitHub Dataset poemsum [1]. This dataset includes poems along with their summaries, where each entry contains the poem's title, author, text, source website, and a brief summary. The overall data statistics for the PoemSum are provided in Table 1, with all length values measured in word counts. "Max" and "Avg." represent the maximum and average lengths, respectively, while "Number of Poets" indicates the number of unique authors in the dataset.

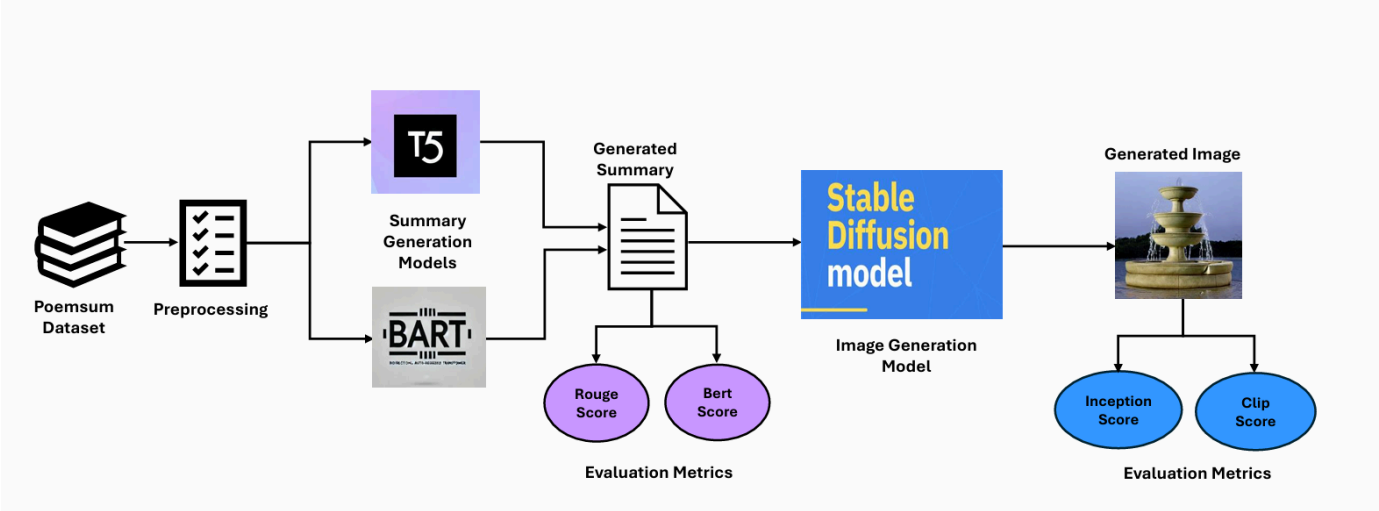


Figure 1. Overview of the poetry summarization and image generation pipeline.

Table 1. Statistics of the PoemSum dataset.

Type	Size
Number of Poems	3011
Number of Poets	930
Max Poem Length	6830
Max Summary Length	1104
Avg. Poem Length	209
Avg. Summary Length	141
Avg. No. of Poems per Poet	3.24

4.2 Data Cleaning

Data cleaning plays a crucial role in enhancing performance by ensuring the input data is consistent, noise-free, and well-structured for effective learning [12, 13]. In the data cleaning process, we remove special characters from poems, ensuring alphanumeric characters and white space are retained. There are multiple coherent spaces, which we reduced to a single space to maintain uniformity, and any leading or training white space is stripped. This cleaning ensures that the text is free of unnecessary symbols, extra spaces, and formatting inconsistencies, resulting in a standardized and clean dataset ready for further preprocessing and analysis using regular expressions.

4.3 Models

We utilized the models T5-base, T5-small, BART for text summarization and Stable Diffusion for image generation on the PoemSum dataset.

BART: BART [4] is a denoising sequence-to-sequence transformer model that combines a bidirectional encoder (like BERT) with an autoregressive decoder (like GPT). It is pre-trained by corrupting text inputs through noising functions (e.g., token masking, deletion) and learning to reconstruct the original text. This pretraining enables BART to effectively generate coherent and contextually rich abstractive summaries, making it well-suited for capturing the nuanced and figurative language in poetry.

T5: A versatile transformer-based model that uses a unified text-to text transfer learning framework that can be used for diverse tasks like translation, summarization, question answering etc [6].

Stable Diffusion: Stable Diffusion is a state-of-the-art latent diffusion model designed for generating high-quality images from textual descriptions [7].

In this process, the generated summary is used as a prompt for the diffusion model. The model's encoder converts the input text into an embedding vector, which is then passed through a latent U-Net architecture. This architecture, coupled with a variational autoencoder, iteratively refines the image, generating a final visual representation that aligns with the semantic content of the poem. This workflow demonstrates how the integration of textual summaries and diffusion models can effectively translate abstract poetic themes into visually compelling images, showcasing the synergy between text and image generation, as shown in Figure 2.

The pre-trained models were fine-tuned with a learning rate of 0.00003, and AdamW optimizer with a batch size of 16.

4.4 Evaluation Metrics

Rouge: It measures the overlap between the actual and generated summaries by comparing unigrams,

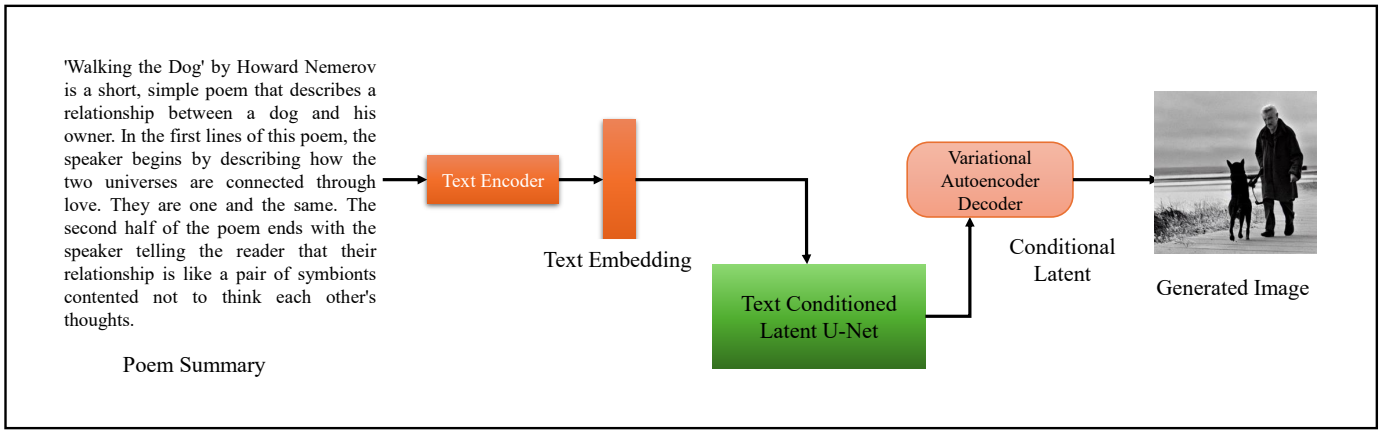


Figure 2. Working of diffusion models.

bigrams, and the longest common subsequence, corresponding to ROUGE-1, ROUGE-2, and ROUGE-L [8]. The formulas for Recall, Precision, and F1-Score are defined as follows:

Recall:

$$\text{Recall} = \frac{\text{Overlapping n-grams}}{\text{Total n-grams in the reference}}$$

Precision:

$$\text{Precision} = \frac{\text{Overlapping n-grams}}{\text{Total n-grams in the generated summary}}$$

F1-Score:

$$\text{F1-Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

BertScore: Instead of considering exact matches, Bertscore analyzes token similarities using contextual embeddings [9].

$$F_1(y, x) = 2 \cdot \frac{P(y, x) \cdot R(y, x)}{P(y, x) + R(y, x)}$$

where:

- $P(y, x)$: Precision.
- $R(y, x)$: Recall.

$$P(y, x) = \frac{1}{m} \sum_{y_i \in y} \max_{x_j \in x} \cos(E(y_i), E(x_j))$$

$$R(y, x) = \frac{1}{n} \sum_{x_i \in x} \max_{y_j \in y} \cos(E(x_i), E(y_j))$$

where:

- x: Tokens in the reference text.
- y: Tokens in the generated text.
- $E(x_i)$: Embedding of token x_i.
- cos: Cosine similarity function.

Inception Score: The Inception Score (IS) is a metric used to evaluate the quality and diversity of images generated by models like GANs or diffusion models [10]. The formula for the Inception Score (IS) is given as:

$$\text{IS} = \exp\left(\mathbb{E}_x \left[\text{KL}\left(p(y|x) \parallel p(y)\right)\right]\right)$$

where:

- $p(y|x)$: The conditional label distribution of an image x.
- $p(y)$: The marginal distribution of labels across all images.
- KL: Kullback–Leibler divergence measures how informative the prediction is (i.e., low entropy means the image clearly resembles one class).

Clip Score: The CLIP Score evaluates the alignment between a text description and an image by computing the cosine similarity between their embeddings [11]. The formula for the CLIP Score is given by:

$$\text{CLIP Score} = \cos(E_{\text{text}}, E_{\text{image}})$$

where:

- E_{text}: Embedding of the text description.
- E_{image}: Embedding of the image.

- cos: Cosine similarity between the embeddings, defined as:

$$\cos(E_{\text{text}}, E_{\text{image}}) = \frac{E_{\text{text}} \cdot E_{\text{image}}}{\|E_{\text{text}}\| \cdot \|E_{\text{image}}\|}$$

5 Results and Discussions

Our primary focus was on generating images from textual summaries derived using Stable Diffusion, leveraging summaries generated by T5 and BART models. The evaluation was performed in two stages:

- **Evaluation of Generated Summaries**

The BART and T5 models were evaluated on the PoemSum dataset using standard metrics such as ROUGE and BERTScore. The BART model outperformed T5, achieving a ROUGE score of 41.90 and a BERTScore of 85.22, compared to T5's ROUGE score of 25.80 and BERTScore of 81.56, as shown in Table 2 and Figure 3. The superior performance of BART can be attributed to its hybrid architecture, which combines bidirectional (like BERT) and autoregressive (like GPT) transformers. This design allows BART to better capture long-range dependencies and contextual relationships, making it well-suited for processing the complex and figurative language of poetry. Additionally, BART's denoising pretraining enhances its ability to generalize across diverse linguistic patterns. As a result, the summaries generated by BART were concise, accurate, and maintained a strong semantic alignment with reference summaries, effectively representing the essence of the poems.

Table 2. Comparison table of evaluation metrices.

Models	ROUGE Score	BERT Score
Baseline	41.18%	84.76
Bart_base	**41.90%**	**85.22**
T5_base	25.80%	81.56

When comparing BART with the baseline model from the base paper, BART also shows improved performance. Mahbub et al. [1] reported a ROUGE-L score of 41.18 and a BERTScore of 84.76 for their baseline model evaluated on the PoemSum dataset. Our BART model achieves slightly higher scores of ROUGE-L 41.90 and BERTScore 85.22 on the same dataset, demonstrating improved summarization performance. Although the improvement in ROUGE-L is modest, the increased BERTScore indicates that

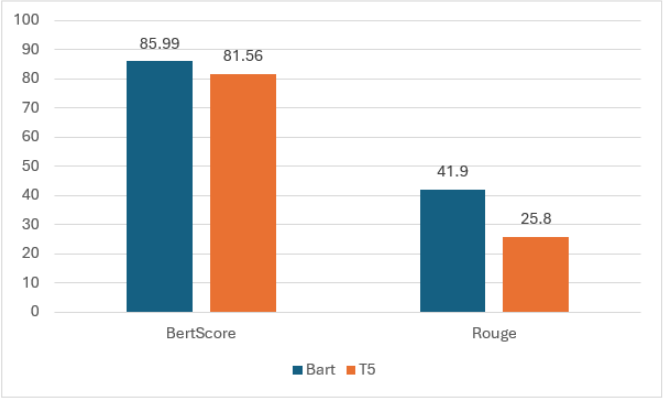


Figure 3. Comparison of metric scores between BART and T5 models.

BART better captures semantic nuances in poetry summaries compared to the baseline model. This suggests that BART is more effective in handling the nuanced and abstract nature of poetic content. Overall, BART's ability to produce high-quality summaries surpasses both the T5 and baseline models, validating its suitability for abstractive summarization tasks, particularly in the challenging domain of poetry.

- **Evaluation of Generated Images**

The textual summaries produced by the BART model were used as prompts for the Stable Diffusion model to generate images. The Inception Score of 7.63 $\pm$ 0.62 and CLIP Score of 29.48, measured on images generated from BART summaries, indicate high visual quality and strong semantic alignment with textual prompts [5]. Inception Score (IS, 1-10 scale) measuring quality/diversity (higher is better), and CLIP Score evaluating text-image alignment (higher scores indicate better match). These results align with findings from related diffusion-based text-to-image models shown in Table 3. The overall performance of the BART-Stable Diffusion pipeline was competitive, as the concise and focused nature of BART's summaries led to visually coherent images that effectively captured the main ideas of the poems.

Table 3. Evaluation of generated images (Inception Score and CLIP Score).

Models	Clip Score	Inception Score
Stable Diffusion	29.48	7.63$\pm$0.602

The final results, including the generated summaries and their corresponding images, are presented in Figures 4 and 5. They showcase the transformation of the abstractive summaries, produced by the BART and T5 models, into visual representations using

Poem	Generated Summary	Generated Image
Let it disturb no more at first Than the hint of a pool predicted far in a forest, Or a sea so far away that you have to open Your window to hear it. Think of it then as elemental, as being Necessity, Not for a cup to be taken to it and not For lips to linger or eye to receive itself Back in reflection, simply As water the patient moon persuades and stirs. And then step closer, Imagine rivers you might indeed embark on, Waterfalls where you could Silence an afternoon by staring but never See the same tumult twice. Yes come out of the street and enter The full piazza. Come where the noise compels. Statues are bowing down to the breaking air. Observe it there - the fountain, too fast for shadows, Too wild for the lights which illuminate it to hold, Even a moment, an ounce of water back; Stare at such prodigality and consider It is the elegance here, the taming, The keeping of a thousand flowering sprays, That builds this energy up but lets watchers See in that stress an image of utter calm, A stillness there. It is how we must have felt Once at the edge of some perpetual stream, Fearful of touching, bringing no thirst at all, Panicked by no perception of ourselves But drawing the water down to the deepest wonder.	Fountain' by Elizabeth Jennings is a poem about the power of water and how it can bring peace to people. In the first lines of this poem, the speaker begins by saying that water should not disturb more than a pool in a forest or a sea far away. She then describes how it can be used as a metaphor for all things. The fountain is too fast for shadows, too wild for the lights that illuminate it to hold even an ounce of water back. This is what makes it so powerful.	

Figure 4. Generated summary and image from Poem I.

Poem	Generated Summary	Generated Image
Two universes mosey down the street Connected by love and a leash and nothing else. Mostly I look at lamplight through the leaves While he mooches along with tail up and snout down, Getting a secret knowledge through the nose Almost entirely hidden from my sight. We stand while he's enraptured by a bush Till I can't stand our standing any more And haul him off; for our relationship Is patience balancing to this side tug And that side drag; a pair of symbionts Contented not to think each other's thoughts. What else we have in common's what he taught, Our interest in shit. We know its every state From steaming fresh through stink to nature's way Of sluicing it downstreet dissolved in rain Or drying it to dust that blows away. We move along the street inspecting shit. His sense of it is keener far than mine, And only when he finds the place precise He signifies by sniffing urgently And circles thrice about, and squats, and shits, Whereon we both with dignity walk home And just to show who's master I write the poem.	'Walking the Dog' by Howard Nemerov is a short, simple poem that describes a relationship between a dog and his owner. In the first lines of this poem, the speaker begins by describing how the two universes are connected through love. They are one and the same. The second half of the poem ends with the speaker telling the reader that their relationship is like a pair of symbionts contented not to think each other's thoughts.	

Figure 5. Generated summary and image from Poem II.

the diffusion model. They effectively highlight how the semantic content of the poetry, captured in the summaries, is accurately reflected in the generated images, demonstrating the strong alignment between the textual summaries and their visual counterparts.

6 Conclusion

This work presents a novel approach for realizing poetic narratives through both interpretation and visualization by combining text summarization and image generation models. By leveraging BART and Stable Diffusion, we introduce an efficient and innovative application of multimodal AI for poetry summarization and image generation. This creative solution bridges textual and visual content while aligning with the latest advancements in natural language processing and generative models. We use BART and T5 for abstractive summarization, with BART outperforming T5 in capturing key content and maintaining semantic alignment with reference summaries. These summaries are then used as textual prompts for the Stable Diffusion model to generate visual representations of the poems. Evaluation through quantitative metrics such as Inception Score and CLIP Score indicates that the concise, focused summaries produced by BART translate effectively into coherent visuals, maintaining strong semantic

alignment. The proposed pipeline, combining BART with Stable Diffusion, demonstrates significant potential for bridging the gap between textual and visual creativity, paving the way for new applications in multimodal understanding and artistic explorations of literature.

7 Limitation and Future Work

Future work can focus on several promising directions to enhance the current methodology. One limitation of this study is the evaluation of only two summarization models—BART and T5—due to resource constraints. In future work, we plan to include more recent models such as FLAN-T5, LLaMA-2, T5-large, and PEGASUS to provide a broader performance baseline. Will try other emerging transformer-based architectures, which could improve the generated summaries. With their larger parameters and sophisticated pretraining, these models are better equipped to capture the abstract and metaphorical nuances inherent in poetry. Another exciting avenue involves extending the methodology to generate videos based on poetic summaries. Video generation can enable dynamic and immersive storytelling, bringing poetic narratives to life through motion, transitions, and evolving visual elements that reflect the temporal and thematic essence of poetry. Expanding the approach to include multilingual capabilities is another crucial area for exploration. Adapting the summarization and generation pipelines to handle poetry in multiple languages could significantly broaden the system's applicability, allowing for the visualization of poetry from diverse cultural and linguistic contexts. Furthermore, integrating additional modalities, such as audio narrations of the poems, alongside text and visuals, could create a richer and more immersive multi modal experience, enhancing the appreciation and accessibility of poetic works for a wider audience. These advancements can push the boundaries of creativity and technology, offering innovative ways to engage with poetry and other forms of literary expression.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mahbub, R., Khan, I., Anuva, S., Shahriar, M. S., Laskar, M. T. R., & Ahmed, S. (2023, December). 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* (pp. 14878-14886). [CrossRef]

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* (*NeurIPS*), *33*, 6840–6851. [CrossRef]

[3] Li, B., Qi, X., Lukasiewicz, T., & Torr, P. (2019). Controllable text-to-image generation. *Advances in neural information processing systems*, 32. [CrossRef]

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 7871). Association for Computational Linguistics. [CrossRef]

[5] Virmani, M., Pathak, M., Pai, K. S., & Prasad, V. B. (2023, May). Image synthesis from themes captured in poems using latent diffusion models. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing* (*ICAAIC*) (pp. 655-660). IEEE. [CrossRef]

[6] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

[7] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., ... & Rombach, R. (2023). Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

[8] Chin-Yew Lin. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81.

[9] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

[10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695). [CrossRef]

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning

transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

[12] Nasfi, R., De Tré, G., & Bronselaer, A. (2025). Improving data cleaning by learning from unstructured textual data. *IEEE Access*. [CrossRef]

[13] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186). [CrossRef]

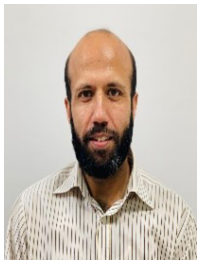
Junaid Yousaf received his MS degree in Computer Science from Ghulam Ishaq Khan Institute (GIKI), He received his Bachelor's degree in Software Engineering from Islamia College University Peshawar, in 2021 Pakistan. He has been awarded the Gradaute Assistantship (GA1) scholarship from GIKI and Frontier Education Foundation (FEF) undergraduate Scholarship by Islamia College University Peshawar, These merit-based scholarships covered full tuition with stipends His research interests include Natural Language Processing (NLP), Computer Vision, Multimodality, Sentiment Analysis, Healthcare and Artificial Intelligence applications. His work focuses on advancing computational techniques to solve real-world problems, with a particular emphasis on bridging the gap between theoretical research and practical implementations. (Email: Junaidyousaf432@gmail.com)

Mazhar Iqbal is currently pursuing a Master's degree in Information Systems Engineering at the Graduate School of Science and Technology, Osaka University, Japan. His expertise includes mobile application development, with a research focus on Artificial intelligence (AI), computer vision, 3D vision, and the Internet of Things (IoT). His work is driven by a passion for creating transformative technologies that address real-world challenges, seamlessly merging cutting-edge research with practical, scalable solutions. (Email: mazharicp786@gmail.com)

Iqra Pervaiz is currently a Master's student in Computer Science from Ghulam Ishaq Khan Institute (GIKI), Pakistan. She is working as a Research Assistant on an HEC-funded NRPU project focused on intelligent water resource management and crop monitoring. Her research interests lie in Data Science, Deep Learning, NLP and real-time AI applications. She aims to develop practical, data-driven solutions that address real-world challenges through cutting-edge computational methods. (Email: pervaiziqra2000@gmail.com)

Muhammad Ismail is a PhD student in the Department of Information Technology at Deakin University, Australia. He received his Bachelor's degree in Electrical Engineering from the University of Engineering and Technology (UET), Peshawar, in 2021, and his Master's degree from UET Mardan in 2023. He has been awarded the DUPR Postgraduate Scholarship by Deakin University, the Ehsaas Undergraduate Scholarship by the Higher Education Commission of Pakistan, and the Stori da Pakhtunkhwa Intermediate Scholarship by the Board of Intermediate and Secondary Education Peshawar. These merit-based scholarships covered full tuition with stipends. His research interests include false data injection attacks (FDIA) in smart grids, operations research, V2X communications, and unmanned aerial vehicles (UAVs). (Email: ismail.muhammad@deakin.edu.au)

Toqeer Ul Islam received his MS degree in Computer Science from the School of Computing and Engineering, Birmingham City University, UK. He completed his Bachelor's degree in Software Engineering from Islamia College University Peshawar, Pakistan in 2021. His research interests lie in Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning, with a particular focus on developing and applying computational techniques to solve real-world problems. His work aims to bridge the gap between theoretical research and practical implementations, driving innovation in AI technologies. (Email: touqeerulislam3988@gmail.com)



Dr. Khurram Khan Jadoon is an Assistant Professor at the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), Pakistan, in the School of Computer Science and Engineering. He holds a Bachelor's degree in Computer Engineering from COMSATS Institute of Information Technology (CIIT), Abbottabad, Pakistan, and an MS in Electronics Engineering from Hanyang University, South Korea. Dr. Jadoon completed his Ph.D. in Computer Science from a prestigious university in South Korea. His academic journey has shaped him into an expert in the fields of Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP), and Computer Vision. He is dedicated to bridging the gap between theoretical research and practical implementations, particularly in developing AI models and computational techniques that address real-world problems.

Dr. Jadoon's research interests primarily lie in the areas of AI, ML, and NLP, with a focus on advancing deep learning methods for both text and image-based data. He is passionate about improving machine learning algorithms to enhance natural language understanding and image processing capabilities. (Email: Khurram.jadoon@giki.edu.pk)