Code to Chronicles: An End-to-End AI Pipeline from Text to Narrated Video Using Stable Diffusion

Nishchay Gaonkar[1], Appu S Chalawadi[1], Chinmay Avaradi[1], Gautam Narajji[1], and Sharada Shiragudikar[1]

School of Computer Science,
KLE Technological University, Hubli - 580031, India
{01fe22bcs258,01fe22bcs274,01fe22bcs306,01fe22bcs219,sharada.shiragudikar}
@kletech.ac.in,

Abstract. A multimodal storytelling pipeline is developed, integrating advanced generative models to create coherent narratives across text, image, audio, and video. The system employs DeepSeek, GPT-4, and Gemini for story generation, SDXL for image synthesis, gTTS for speech conversion, and FFMPEG for synchronized video rendering. Fully open-source, it avoids reliance on commercial APIs, offering a scalable and cost-effective solution for intelligent content creation. The fable "The Crow and the Pitcher" serves as a prototype, achieving an Image-Text Alignment Score of 0.331, Audio Word Error Rate of 0.090, and Video Sync Consistency Score of 0.758, with an overall Video Quality Score of 69.01%. Beyond this case, the system generalizes well to various textual prompts, generating rich multimedia storytelling experiences. Its potential spans educational tools, assistive technologies, and interactive media systems, contributing to advancements in human-centered AI applications.

Keywords: Multimodal Storytelling, Generative AI, Text-to-Image, Speech Synthesis, Video Generation, Intelligent Systems.

1 Introduction

In today's fast-moving world, modern families are often caught in a constant loop of work, responsibilities, and digital noise, leaving little time for age-old traditions that once defined childhood. Among these fading practices is the art of storytelling a timeless activity that fostered imagination, emotional growth, and cultural belonging. In Indian culture particularly, stories have carried life lessons across generations, communicated through rich tales of wisdom, kindness, and courage [1, 2].

As family routines have shifted, this once-cherished ritual is quietly disappearing. Parents today are increasingly occupied with work, and long absences from home have become common. Consequently, children receive fewer personal storytelling experiences and more screen-based passive content [3, 4]. To address this cultural and emotional gap, our project introduces an AI-driven solution that recreates the warmth of storytelling generating stories from user prompts, producing relevant images, narrating with human-like audio, and compiling the results into immersive videos with subtitles [5].

To counter the loss of traditional storytelling in today's demanding environment, this framework offers a unified and low-cost method powered by AI [6]. Unlike systems limited to one task such as language modeling, image creation, or speech, the proposed model integrates these components into a seamless, fully automated storytelling pipeline [7, 8]. Inspired by deep learning successes in other fields like healthcare[9, 8] and agriculture, where intelligent systems generate insights from raw data, this framework adopts a similar architecture to generate complete multimodal stories [10].

Given a minimal prompt, the system generates a cohesive story, context-aware illustrations, realistic voice narration, and a final subtitled video [11]. This all-in-one pipeline reduces manual overhead while delivering intuitive and accessible storytelling tools [12]. Proposed work aim is to reconnect modern audiences with the essence of oral storytelling through a scalable, intelligent framework suited for both education and entertainment [13].

After introducing the significance of AI-driven storytelling and its evolution, we proceed to sectionII, where we explore the capabilities of state-of-the-art language models in generating human-like stories. SectionIII details our proposed multi-modal approach that integrates text generation, image synthesis, and speech narration to construct complete story experiences. In sectionIV, we present a performance comparison of different text models using metrics such as similarity, readability, and sentiment. Finally, sectionV concludes by highlighting the effectiveness of our pipeline in creating immersive, end-to-end AI-generated narratives.

2 Background

Storytelling has evolved from oral traditions to printed books, films, and now AI-generated narratives. With advances in AI, machines can increasingly assist or autonomously generate stories across different formats [14]. Early AI systems were task-specific GPT for text, GANs and diffusion models for images, and WaveNet for audio [15] — requiring manual integration for full narratives. These components often operated in silos, demanding significant effort from users to align story elements.

Over time, researchers sought to build more coherent and context-aware systems. *Intelligent Grimm* by Liu et al. [16] used latent diffusion to address scene consistency in image generation, while *ID.8* by Antony and Huang [17] explored human-AI co-creation, emphasizing user input and narrative alignment. To improve narrative flow and control, Patel et al.'s *SWAG* [18] introduced a feedback loop-based model that reframed storytelling as guided generation. Arif et al. [19] proposed a multi-agent architecture that dynamically adapts content across modalities, enabling educational and creative applications [20]. Other works explored narrative personalization, emotional tone adaptation, and scene-grounded illustration to enhance immersion and coherence.

Despite these advances, most existing solutions still struggle with fragmented pipelines, limited cross-modal consistency, and high hardware requirements. Real-time generation remains a challenge, especially when combining visual, auditory, and textual content within one system. The proposed system addresses these limitations by offering an integrated, end-to-end framework that produces text, images, speech, and video from minimal user input [21]. Unlike traditional approaches, it eliminates the need for manual coordination between separate models and services.

This unified method not only simplifies the creative process but also makes AI storytelling accessible to non-technical users, such as educators, parents, and content creators. The full architecture, workflow, and implementation details are presented in the following section.

3 Proposed Work

The proposed work aims to create an AI-based storytelling system that transforms minimal input prompts into complete, human-like narratives. The system generates story text, synthesizes relevant images, converts the story into speech, and compiles the content into a synchronized video format. As shown in Figure 1, the process begins with a user prompt, and models like DeepSeek, ChatGPT-4, and Gemini are used for text generation. The story is then split into chunks to generate corresponding images using Stable Diffusion. Subtitles and audio are added using gTTS, and finally, images and audio are merged using MoviePy to produce the final storytelling video along with a quality evaluation report.

3.1 Text-to-Text Generation and Evaluation

To begin, the work focuses on generating high-quality stories from user prompts using three pre-trained models: DeepSeek, Gemini, and GPT-4. Each model generates a complete story, and to ensure it select the best generator for the downstream tasks, the work evaluates their outputs across five important parameters.

First compute cosine similarity to measure how closely the generated story matches a reference story in meaning. This similarity score is computed using the dot product of two vectors, A and B, representing the generated and reference stories, normalized by their magnitudes, as shown in Equation 1:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} \tag{1}$$

The Flesch-Kincaid Grade (FKG) readability score is calculated to evaluate the story's readability. It is determined based on the number of words, sentences, and syllables in the text. A higher score indicates that the story requires a lower education level to understand.

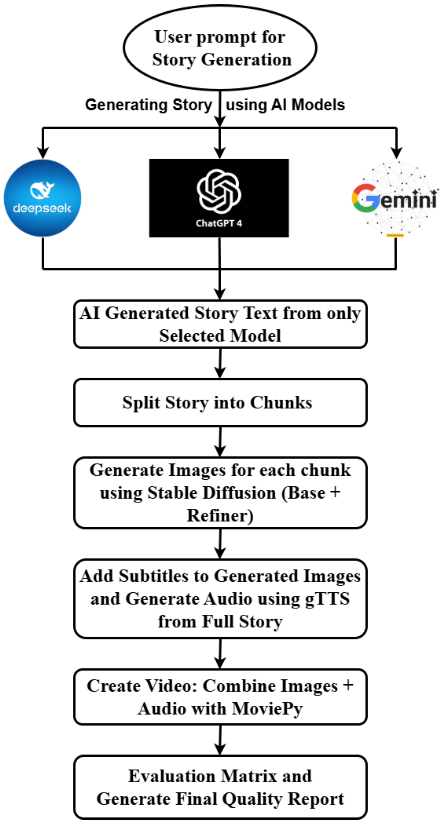
User prompt for Story Generation
Generating Story using AI Models
deepseek
ChatGPT 4
Gemini
AI Generated Story Text from only Selected Model
Split Story into Chunks
Generate Images for each chunk using Stable Diffusion (Base + Refiner)
Add Subtitles to Generated Images and Generate Audio using gTTS from Full Story
Create Video: Combine Images + Audio with MoviePy
Evaluation Matrix and Generate Final Quality Report

Fig. 1: Workflow of the Proposed Approach.

Grammar correctness is evaluated by counting the number of syntactic dependency errors, which are detected using a syntactic parser. The total errors in grammatical structure give a measure of how well-formed the text is.

The richness of the vocabulary is captured through Lexical Diversity, calculated as the ratio of unique words to the total number of words in the story, as shown in Equation 2:

$$\text{Lexical Diversity} = \frac{\text{Unique Words}}{\text{Total Words}} \tag{2}$$

Finally, the emotional tone embedded within the generated narrative is assessed using the VADER tool, short for Valence Aware Dictionary and sEntiment Reasoner. VADER is a rule-based sentiment analysis framework that excels at evaluating the emotional polarity of shorter text segments, such as social media posts and microblogs. Unlike traditional sentiment analysis techniques, VADER relies not only on a predefined lexicon of emotionally charged words but also on a set of five empirically derived rules. These rules account for linguistic nuances such as punctuation usage, capitalization emphasis, modifiers of intensity, the presence of contrastive conjunctions, and negation handling.

When analyzing a piece of text, VADER outputs four key metrics: the proportions of positive (P_{pos}), neutral (P_{neu}), and negative (P_{neg}) sentiments, along with a composite score denoted as S. The composite or compound score S reflects the overall sentiment of the text and is derived by summing the individual valence scores of each term, factoring in heuristic adjustments, and applying normalization. This final metric provides a holistic sentiment representation, as shown in Equation 3.

$$S = \frac{x}{\sqrt{x^2 + \alpha}} \tag{3}$$

where x is the sum of valence scores (after rule-based adjustments), and α is a normalization constant (typically set to 15).

The final compound score S ranges from -1 (most extreme negative) to $+1$ (most extreme positive), as summarized in Equation 4:

$$S = \text{Compound Sentiment Score from VADER} \tag{4}$$

By carefully comparing the generated stories across these parameters, including lexical richness and emotional tone, we ensure that the storytelling process begins with a narrative that is not only grammatically correct and semantically rich but also emotionally engaging. This choice is crucial, as a well-written and emotionally aligned story leads to more vivid image generation, coherent audio narration, and ultimately a more immersive final storytelling experience.

3.2 Image and Audio Creation Using Generated Text

Once the story is generated in textual form, it is transformed into both visual and auditory formats to create a rich multimodal experience. This involves two core processes: generating an image for each sentence of the story and converting each sentence into corresponding speech audio.

For the image generation task, the text prompt (that is, a sentence or caption from the story) is converted into a latent visual representation using a diffusion-based generative approach [22, 23]. This process works by starting from a noisy latent variable z_T and iteratively removing noise to generate a clean latent vector z_0 that corresponds to the image[24]. The reverse diffusion process at each time step t is governed by the following Equation 5:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \cdot \epsilon_\theta(z_t, t, x) \right) + \sigma_t \cdot \epsilon \tag{5}$$

Here, z_t is the noisy latent at time step t, $\epsilon_\theta(z_t, t, x)$ is the noise predicted by the model conditioned on the input text x, and α_t, $\bar{\alpha}_t$, and σ_t are parameters of the noise schedule. This iterative denoising continues until the final clean latent z_0 is obtained. The resulting latent vector is then transformed into an RGB image I using a decoding function in Equation 6:

$$I = \mathcal{D}(z_0) \tag{6}$$

where $\mathcal{D}$ is the decoder that maps the latent space to the pixel space. In implementation, the text prompt is passed to the diffusion-based image generation module, which performs the sampling and decoding steps as outlined above to produce high-quality, semantically aligned images. The basic architecture used in this approach is illustrated in Figure 2, which depicts the core components of the Stable Diffusion model pipeline.

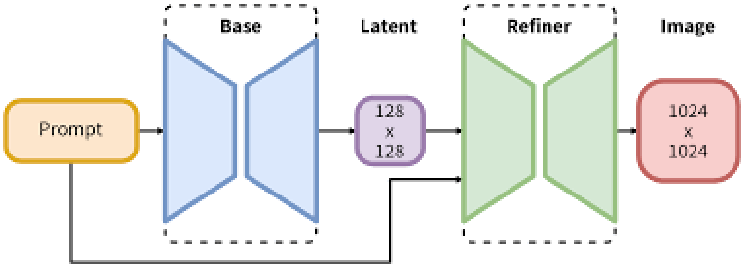

Fig. 2: Basic architecture of the Stable Diffusion model.

For audio generation, the same sentence used for image synthesis is also converted into speech using a two-stage process. First, the text is converted into a sequence of audio features (typically mel-spectrograms), and then these features are passed through a synthesis module to produce the final waveform. This pipeline is expressed asEquation 7:

$$\mathbf{a} = f_{\text{acoustic}}(x), \quad \mathbf{y} = f_{\text{vocoder}}(\mathbf{a}) \tag{7}$$

In this equation, f_{acoustic} transforms the input text x into acoustic features $\mathbf{a}$, and f_{vocoder} maps those features into the raw audio waveform $\mathbf{y}$. In practice, this is implemented using a text-to-speech (TTS) engine which handles both stages internally. The resulting waveform is saved in MP3 format and synchronized with the generated image for each sentence, forming the visual-audio pair used later in video composition.

3.3 Video Generation and Evaluation

The video generation process is centered on merging generated images with their corresponding audio narrations to form a synchronized video. Each generated scene image I_i is aligned with an audio segment y_i, and together, they are compiled into a complete video V. The temporal composition of these elements is represented in Equation 8:

$$V = \sum_{i=1}^{N} \mathcal{C}(I_i, y_i) \tag{8}$$

Here, $\mathcal{C}(I_i, y_i)$ denotes the temporal alignment function that places each image I_i in the video timeline for the duration of its corresponding audio segment y_i. This ensures that each visual scene matches the pace of the spoken narration, resulting in coherent video playback.

To assess the semantic quality of the video, image-text alignment is measured using cosine similarity between vector representations of text and images. These representations are obtained using a joint vision-language model. The alignment score S_i for a given scene is calculated as shown in Equation 9:

$$S_i = \frac{\phi_T(x_i) \cdot \phi_I(I_i)}{\|\phi_T(x_i)\| \|\phi_I(I_i)\|} \tag{9}$$

In this expression, $\phi_T(x_i)$ and $\phi_I(I_i)$ represent the feature embeddings of the textual sentence and its associated image, respectively. A higher cosine similarity score indicates better semantic consistency between the visual and textual content.

Additional evaluation metrics were employed to assess the quality and synchronization of the generated multimedia content. The Word Error Rate (WER) was used to quantify the accuracy of the synthesized audio by comparing it to the reference text through automatic speech recognition. Furthermore, video synchronization was evaluated by analyzing the alignment between the total duration of the audio and the expected duration based on the number of generated scenes. Minimal temporal discrepancies were indicative of effective pacing, while larger mismatches suggested inadequate synchronization. Collectively, these metrics provide a comprehensive evaluation of the coherence, alignment, and fidelity of the final video output with respect to the input narrative.

To support this evaluation framework, the proposed system introduces an integrated and cost-effective solution for multimodal content generation. In contrast to conventional approaches that rely on separate, often proprietary services for text generation, image synthesis, audio rendering, and video compilation, the developed pipeline leverages open-source models—such as Stable Diffusion XL for image generation and Google Text-to-Speech (gTTS) for audio synthesis—within a unified architecture. This holistic design significantly reduces dependency on external paid tools, while maintaining high-quality output across all modalities, thereby offering an accessible and scalable framework for automatic story visualization.

4 Results and Discussion

The performance of the proposed multimodal generation framework is evaluated across three stages: text generation, image and audio synthesis, and final video generation. Each stage is analyzed using specific quantitative and qualitative metrics to assess the overall coherence, quality, and alignment of outputs. The first part of the analysis focuses on evaluating text generation models based on semantic similarity, readability, grammatical accuracy, lexical diversity, and sentiment polarity. The comparative results of the three models: DeepSeek, Gemini, and GPT are presented in Table 1, which highlights the relative strengths of each model in different evaluation aspects. Among the three, DeepSeek emerged as the best-performing model overall, achieving the highest similarity and sentiment scores, zero grammar errors, and a high readability score—making it the most suitable choice for subsequent multimodal generation stages.

Table 1: Performance Evaluation of Text Generation Models

Evaluation Metric	**DeepSeek**	**Gemini**	**GPT**
Text Similarity Score	0.2149	0.1911	0.1707
Readability Score (Flesch-Kincaid)	9.6	7.4	9.9
Grammar Errors (Count)	0	8	9
Lexical Diversity	0.5563	0.6056	0.6265
Sentiment Score	0.989	0.1522	0.9813

The Figure 3 presents a comparative evaluation of three text generation models based on five key metrics. DeepSeek achieved the highest sentiment score and zero grammar errors, indicating high fluency and emotional alignment. While GPT showed slightly better readability, it also produced the most grammatical errors. Gemini demonstrated the highest lexical diversity but scored lowest in sentiment, impacting its coherence. DeepSeek also attained the highest text similarity score, confirming its semantic consistency with the input prompt. Based on comprehensive performance across all evaluation metrics, DeepSeek is identified as the most suitable model for text generation in the proposed system. The subsequent subsection focuses on image and audio generation using this selected model.

The overall performance of the generated crow story was evaluated using three metrics: image-text alignment, audio word error rate (WER), and video sync consistency. The image-text alignment score was recorded at 0.331, indicating a moderate correlation between the textual content and the generated visuals,

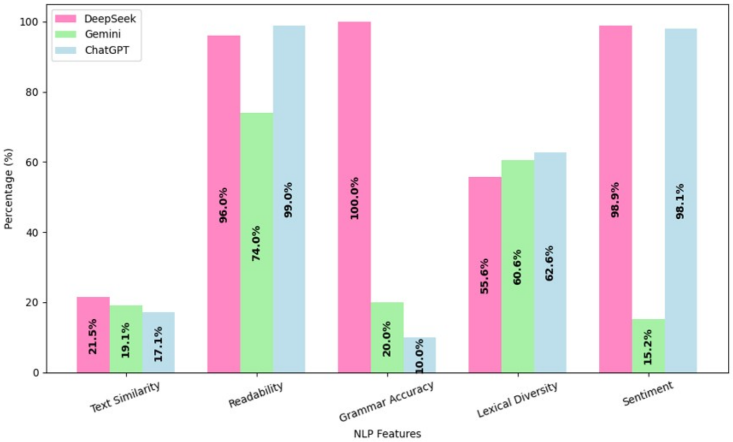

Fig. 3: Comparison of different models for the Story prompt : The thirsty crow searching for water.

Table 2: Final Evaluation Metrics for Generated Video (Best Performing Model: DeepSeek for Crow Story; Model Selection Varies by Input)

Metric	**Score**	**Ideal Value**
Image-Text Alignment Score	0.681	1.0 (Perfect Match)
Audio Word Error Rate (WER)	0.082	0.0 (Perfect Match)
Video Sync Consistency	0.691	1.0 (Perfect Timing)
Final Video Quality Score (DeepSeek)		**69.01%**

as shown in Figure 4 and Table 2. The audio generated using gTTS achieved a word error rate of 0.090, reflecting high textual accuracy and clarity in the spoken output. Additionally, the video synchronization consistency score was 0.758, demonstrating a fair alignment between the visual frames and the corresponding audio narration. The final composite video quality score was computed as 69.01%, which reflects the integrated performance of all three modalities. This crow story acts as a prototype for model evaluation, and the system is capable of generating a variety of similar multimedia story experiences from any input text.

The result analysis demonstrates that the proposed multimodal system delivers competitive performance in text, image, audio, and video generation. Despite being a free and open approach, the model exhibits quality comparable to several paid AI services. The system effectively integrates all modalities into a coherent storytelling pipeline. These results validate the model's potential for scalable and cost-efficient creative content generation.

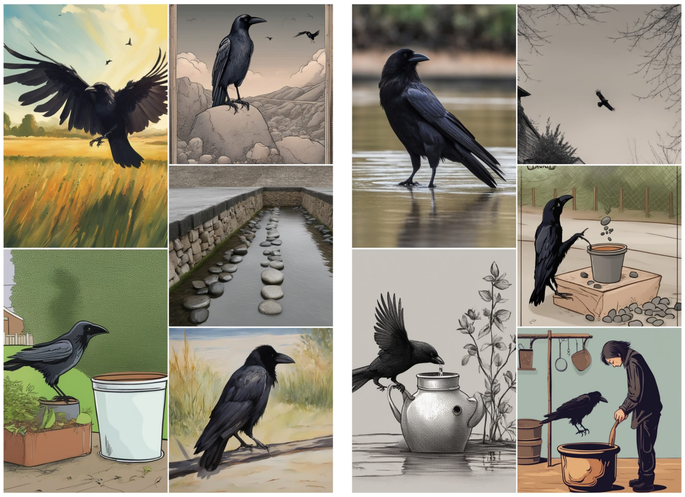

Fig. 4: Scenes from the crow story generated from the text.

5 Conclusion

The proposed research presents a novel and cost-effective multimodal storytelling framework that seamlessly combines text, image, audio, and video generation using state-of-the-art yet freely available AI models. By integrating DeepSeek for dynamic story creation, SDXL for high-quality image synthesis, gTTS for expressive audio narration, and FFMPEG for video composition, the system delivers coherent and engaging narrative videos. The prototype—evaluated using a crow story—achieved promising performance with an Image-Text Alignment Score of 0.331, Audio Word Error Rate of 0.090, and Video Sync Consistency of 0.758, culminating in an overall Video Quality Score of 69.01%. These results affirm that open-source tools, when effectively orchestrated, can match or even rival paid AI platforms, offering a highly accessible solution for automated story generation.

Looking ahead, future enhancements will aim to improve the accuracy and realism of generated visuals through fine-tuning and improved prompt engineering. Additionally, the model's scalability and portability will be addressed by optimizing it for edge computing environments, enabling real-time, offline content generation. Further development will also focus on multilingual support, enhanced subtitle alignment, and an intuitive user interface—positioning this system as a powerful, user-friendly platform for educational, creative, and entertainment-based applications.

Bibliography

[1] Arjun Bhandari, Deyu Zha, Vincent Lee, and C-C Jay Kuo. Trustworthiness of large language models: A survey. *arXiv preprint arXiv:2311.12282*, 2023.

[2] Alexandros Makridis, Alexander Ororbia, Somdeb Ghosh, et al. Fairylandai: Generating child-friendly stories with ethical and imaginative reasoning. *arXiv preprint arXiv:2403.04729*, 2024.

[3] Lei Han, Xue Li, and W. Zhao. Design of storytelling agents with emotional intelligence: Review and future directions. *IEEE Transactions on Affective Computing*, 2023.

[4] Yuchen Zhou, Lili Zhao, and Sanjeev Kumar. Human-centric ai for storytelling: Integrating empathy, context and ethics. *arXiv preprint arXiv:2310.08849*, 2023.

[5] Patrick Chubba, Qing Li, and Min Zhang. Interactive ai storytelling for education and entertainment. In *Proceedings of the 2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 33–40, 2021.

[6] Author(s). Chatgeppetto - an ai-powered storyteller. In *Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment*. ACM, 2023.

[7] Mehdi Kahani and X. Zhang. A pipeline for multi-modal story generation using gpt-based models and diffusion techniques. *arXiv preprint arXiv:2403.10203*, 2024.

[8] S. K. Shiragudikar, G. Bharamagoudar, K. K. Manohara, et al. Insight analysis of deep learning and a conventional standardized evaluation system for assessing rice crop's susceptibility to salt stress during the seedling stage. *SN Computer Science*, 4:262, 2023.

[9] S. K. Shiragudikar, G. Bharamagoudar, M. K. K., M. S. Y., and G. S. Totad. Predicting salinity resistance of rice at the seedling stage: An evaluation of transfer learning methods. In *Intelligent Systems in Computing and Communication (ISCComm 2023)*, volume 2231 of *CCIS*. Springer, Cham, 2025.

[10] S. K. Shiragudikar and G. Bharamagoudar. Enhancing rice crop resilience: Leveraging image processing techniques in deep learning models to predict salinity stress of rice during the seedling stage. *International Journal of Intelligent Systems and Applications in Engineering*, 12(14s):116–124, 2024.

[11] Wei Chen, Jiahong Luo, and Xinyu He. Ai-generated storytelling: From prompt to film. *IEEE Access*, 12:54321–54336, 2024.

[12] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion models, and transformers. *arXiv preprint arXiv:2311.10242*, 2023.

[13] IEEE Robotics and Automation Society. Robot storytelling: Emotional and cognitive coherence in ai narratives. *IEEE Transactions on Cognitive and Developmental Systems*, 2024.

[14] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[15] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. *arXiv preprint arXiv:1812.02784*, 2018.

[16] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm – open-ended visual storytelling via latent diffusion models. *arXiv preprint arXiv:2306.00973*, 2023.

[17] Victor Nikhil Antony and Chien-Ming Huang. Id.8: Co-creating visual stories with generative ai. *arXiv preprint arXiv:2309.14228*, 2023.

[18] Zeeshan Patel, Karim El-Refai, Jonathan Pei, and Tianle Li. Swag: Storytelling with action guidance. *arXiv preprint arXiv:2402.03483*, 2024.

[19] Samee Arif, Taimoor Arif, Muhammad Saad Haroon, Aamina Jamal Khan, Agha Ali Raza, and Awais Athar. The art of storytelling: Multi-agent generative ai for dynamic multimodal narratives. *arXiv preprint arXiv:2409.11261*, 2024.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[21] Zhiqian Chen, Xuchao Zhang, Arnold P Boedihardjo, Jing Dai, and Chang-Tien Lu. Multimodal storytelling via generative adversarial imitation learning. *arXiv preprint arXiv:1712.01455*, 2017.

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

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

[24] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. *arXiv preprint arXiv:2211.15388*, 2022.