BIBLIOSMIA*: Hyper-Personalized Consistent Stories for Enhanced Social Emotional Learning

Akib Shahriyar AKIB@NOOKLY.COM
Radwa Hamed RADWA@NOOKLY.COM
E. Margaret Perkoff MARGARET@NOOKLY.COM
Mostafa Aboelnaga MOSTAFA.ELSAYED@BLINK22.COM
Alya Azab ALYA.AZAB@BLINK22.COM

Abstract

Consistent visual storytelling plays a central role in how humans teach their children to understand their emotions, relationships with others and the world around them. It enhances children's cognitive and social-emotional development by providing engaging, believable narratives that help them navigate emotional complexities in a safe, imaginative context. This prevalence of visual storytelling in our lives has made it a prime application for technological advancements in artificial intelligence (AI). With AI integrations, digital stories can be created across an infinite set of topics and readily adapted to personalized contexts. As image generation algorithms advance, digital storytelling can be enhanced even further to incorporate visual elements that are unique to the author or the desired reader population. However, the burgeoning field of multi-modal story generation currently suffers from the problem of consistency - a critical element for preserving the plot lines of a story and the ability of children to relate to the characters therein. To mitigate this, we propose *Bibliosmia*, a novel framework for designing consistent digital stories. The framework encompasses three main components: a story generation module, an alignment module and an image generation and validation module that collectively preserve key narrative and visual story elements to allow expert and new authors alike to craft deeply personal developmental stories for children. We evaluated the effectiveness of the framework in the context of an online automatic story generation application. Our experimental results demonstrate *Bibliosmia*'s superior performance in prompt similarity (0.307 CLIP Score) and near-top-tier identity consistency (0.860 CLIP Score), surpassing other approaches in scalability and user satisfaction. These findings highlight *Bibliosmia*'s effectiveness in delivering high-quality, personalized storytelling experiences, setting a new standard in multi-modal digital storytelling.

Keywords: Multi-Modal Storytelling, Digital Story Generation, Character Consistency, Image-text Alignment

1. Introduction

Throughout human history, stories have played a central role in shaping society. Stories help people in everything from defining legal and social norms to sparking creativity, fostering ideals and providing explanations for controversial matters Mathews and Wacker (2008). Visual storytelling has long served as an essential tool for shaping humanity's understanding of history, culture and emotions Boyd (2009); Eisner (2008). For children, stories become even more impactful, offering a structured yet creative environment for cognitive develop-

* From the greek words "biblio" (book) and "-osmia" (smell). bibliosmia (n.) is the pleasant aroma of a new book, caused by the gradual chemical breakdown of the compounds used within the paper.

ment and social-emotional growth Morrison (2024).

Children's social-emotional learning (SEL) is influenced by a confluence of cognitive, linguistic and cultural factors. Prior research underscores how storytelling systems provide a unique opportunity to scaffold learning in these domains by enabling children to navigate emotional complexities in a safe, imaginative context Gunawardena and Koivula (2023); Heath et al. (2017). The importance of well-developed believable characters in such narratives cannot be overstated; consistent character representation enhances engagement and believability directly impacting the success of narrative Riedl and Young (2010).

The advent of artificial intelligence (AI) has introduced transformative potential for personalized storytelling by tailoring narratives to individual learning needs and contexts Luckin and Holmes (2016); Cardona et al. (2023). Research into AI-driven education has highlighted the importance of adaptive learning systems that grow alongside the learner, catering to their evolving needs and abilities Luckin and Holmes (2016); Holmes et al. (2019); Uslu and Uslu (2021). Multimodal storytelling further introduces layers of complexity, as visual, auditory and textual elements must not only align with each other but also with the learner's developmental stage. This highlights the need for a robust validation mechanism to reconcile these elements into a cohesive, engaging and personalized narrative.

Prior work has made strides in enhancing consistency in AI-generated text storytelling through memory-driven coherence Rahman et al. (2023), context-aware transitions Kim et al. (2024) and multimodal alignment Zang et al. (2024), but these methods often struggle with dynamic narrative shifts, misalignment with reading levels, a lack of cultural sensitivity or broader contextual coherence. These work evaluates consistency in the context of the text narrative only. As image generation models catch up with the existing text generation methods Li et al. (2025); Tewel et al. (2024); Wang et al. (2024), this opens up a new set of challenges including the need for consistency in the story visuals.

To mitigate these issue, we introduce the Bibliosmia framework for multi-modal storytelling design. Bibliosmia addresses these shortcomings by unifying character-driven micro-consistency and macro-level narrative flow, while ensuring adaptability to diverse cultural and stylistic contexts. Bibliosmia integrates robust character representations with advanced prompt engineering and diffusion models to produce coherent narratives and visually compelling imagery. Through Bibliosmia, we provide a framework that not only preserves the narrative and visual fidelity of stories but also adapts dynamically to each child's developmental needs. By doing so, we contribute to a future where AI is not just a tool but a partner in nurturing children's growth—one story at a time.

2. Related Work

2.1. Consistent AI-generated Storytelling

Efforts to enhance consistency in sequential image generation for stories can be grouped into three key approaches: memory-driven coherence, context-aware transitions and multimodal alignment Antony and Huang (2024); Rahman et al. (2023); Kim et al. (2024); Zang et al. (2024). Memory-driven methods Rahman et al. (2023) use visual memory modules to retain character appearances and settings across scenes, ensuring micro-level coherence but struggling with dynamic narrative shifts. Context-aware strategies Kim et al. (2024) employ transition maps and temporal anchoring to maintain macro-level narrative flow, though they can be computationally intensive and less adaptable to stylistic changes. Mul-

timodal alignment approaches Zang et al. (2024) unify textual and visual elements through shared embedding spaces, incorporating character-centric modules for fluency and believability. CHIRON Gurung and Lapata (2024) complements these by using structured "character sheets" to model rich character representations, capturing dialogue, traits, knowledge and goals for nuanced, evolving characters in long-form narratives.Building on CHIRON's foundation, our framework extends this methodology by pairing story caption generation with the creation of detailed, validated character sheets to enhance storytelling adaptability across educational and creative applications.

2.2. Creative Multi-Modal Story Generation using AI

Creative story generation has long struggled to balance originality and coherence. Traditional models frequently sacrificed creativity for consistency, relying heavily on pre-defined datasets Fan et al. (2018); Yang et al. (2022). While large language models (LLMs) like DOC and Re3 have improved the ability to generate intricate, longer narratives, their adaptability to new contexts remains limited. Adding another layer of complexity, the emergence of multimodal storytelling combines textual narratives with visual elements, resulting in richer and more immersive experiences. Early systems struggled with limited or unstructured outputs Yang et al. (2024), but recent advances in latent diffusion models have significantly improved text-to-image alignment and visual storytelling capabilities Rombach et al. (2022). To bridge this gap, our Bibliosmia framework incorporates CHIRON's character validation approach along with the state-of-the-art multimodal diffusion transformer models. This combination enables the generation of stories that dynamically adapt to characters' development while maintaining both creativity and coherence.

3. Methodology

Figure 1 showcases an overview of our Bibliosmia framework which is divided into three main modules: the **Generation Module**, which creates personalized narratives and detailed character representations; the **Alignment Module**, which ensures the seamless integration of textual and visual elements; and the **Image Generation and Validation Module**, responsible for producing and refining high-quality images aligned with the story. Each story in this work consists of both text and images, where the text serves as a caption paired with the generated images to provide narrative context. Together, these modules address the challenges of narrative fidelity and character consistency, setting the foundation for engaging, visually rich and personalized storytelling experiences.

3.1. Generation Module

Story Generation submodule starts with creating a detailed child profile, capturing demographics, physical traits, family structure and interests, alongside user-defined story attributes like length, style and accessibility features. Using a tailored prompt and state-of-the-art Large-Language Models (LLMs) like GPT-4o Achiam et al. (2023), it generates age-appropriate, SEL-aligned story captions with a clear beginning, middle and end.

Character Sheet Generation submodule uses the generated story captions and child profile to extract key attributes for each character. These attributes include physical features, age, personality traits and roles within the story. The finalized character sheets provide a comprehensive and consistent representation of all characters, forming the basis

for subsequent modules. See Figure 3 for sample *Character Sheet* JSON structure.

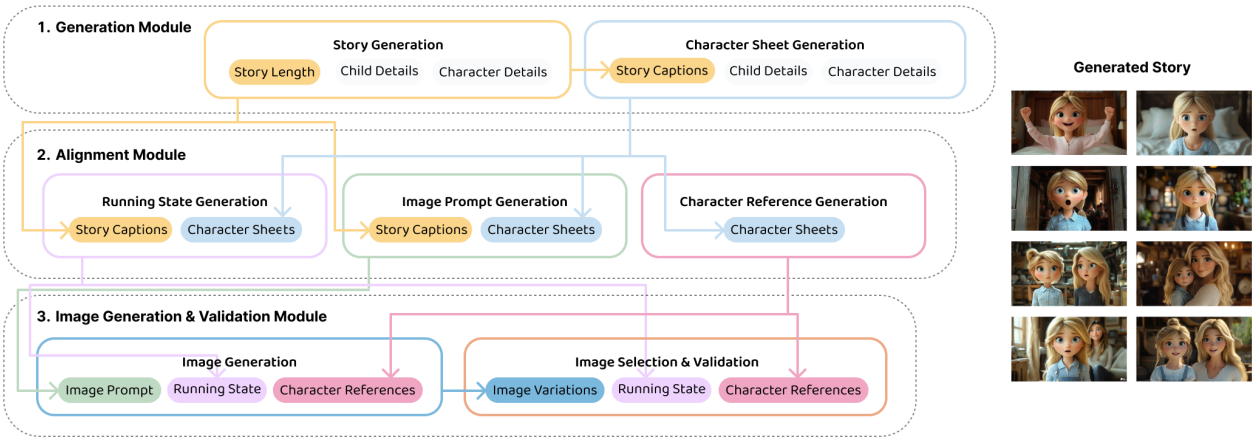

Figure 1: An overview of our Bibliosmia framework consisting of Generation Module, Alignment Module and Image Generation and Validation Module

3.2. Alignment Module

Character Reference Generation submodule focuses on using character sheet details to build image prompts used to generate Character Reference Images. These character reference images act as visual anchors to maintain consistency in character portrayal across the entire narrative.

Running State Generation submodule ensures that the story's dynamic elements, including character interactions, emotional expressions and settings, are systematically tracked. By analyzing the story captions and character sheets, this submodule generates a running state for each narrative point, outlining the appearances of the characters involved, their emotional states and the environmental context.

Image Prompt Generation submodule synthesizes the outputs of the previous steps into structured prompts for image generation. Using the story captions, running states and character sheets, it creates prompts that detail the scene's composition, character actions and settings while adhering to guidelines that ensure clarity and cultural appropriateness.

3.3. Image Generation & Validation Module

In the **Image Generation** submodule, the final image prompts are paired the detailed image prompts with their corresponding character references to ensure that the visualizations accurately reflect the narrative's context and maintain character consistency. The Multi-Modal Diffusion Transformer (MMDiT) Esser et al. (2024) model serves as the core architecture for image generation. This model leverages its ability to process both text and visual inputs, using forward diffusion and reverse denoising processes to synthesize coherent and contextually aligned images.For each image prompt, the MMDiT generates multiple candidate images, typically between 4 to 8, to provide diverse options. These images are stored in the database for subsequent evaluation and refinement.

In the **Image Selection and Validation** submodule, each image prompt, along with its corresponding candidate images and character references, is evaluated using a state-of-the-art Vision-Language Model, such as GPT-4 Achiam et al. (2023). The Vision-Language

Model assesses the images for their consistency with the narrative descriptions, character attributes and emotional themes. The model also rates the candidate images, identifying the most accurate and coherent visual representation for each story point. Once the best image for each story point is identified, the module compiles these images to form the complete illustrated story.

3.4. Evaluation Metrics

To evaluate the alignment between the generated visuals and their corresponding textual prompts, as well as the consistency of character representations across different story contexts, we employ the CLIP-based Radford et al. (2021) metrics for prompt similarity and identity consistency which are commonly used in the personalization literature Avrahami et al. (2023); Gal et al. (2023); Ruiz et al. (2023).

Prompt Similarity metric evaluates how well the generated images align with the textual descriptions in the prompts. Using CLIP, we compute the normalized cosine similarity between the CLIP text embedding of the input prompt and the CLIP image embedding of the generated image. A higher similarity score indicates a closer alignment between the visual output and its corresponding textual input, ensuring that the generated images accurately reflect the narrative elements.

Identity Consistency calculates the pairwise similarity between the CLIP image embeddings of the same character generated under different prompts to ensure character consistency across varying contexts. This approach assesses whether the visual representation of characters remains consistent throughout the story, regardless of changes in narrative scenarios or textual variations.

4. Experiment Results and Key Insights

This section compares the pre-launch and post-launch phases of the Bibliosmia framework, during which our platform was publicly accessible to a diverse user base. In the pre-launch phase (May 6–June 6, 2024), our platform used GPT-4 for story and image prompts with MMDiT for image generation. Post-launch (October 8–November 8, 2024), the full Bibliosmia framework was implemented. User privacy was maintained throughout the study. Our platform is now open for public access at www.nookly.com.

Prompt and Identity Consistency We have randomly sampled 10 generated stories each from the pre-launch and post-launch period of Bibliosmia framework where each story contained the image prompts and their corresponding generated images. Figure 2(a) highlights the superior performance of Bibliosmia in achieving a balanced trade-off between prompt similarity and identity consistency, as evidenced by its leading scores in prompt similarity (0.307) and competitive identity consistency (0.860). Unlike methods such as LoRA-DB, which exhibits the highest identity consistency (0.877) but requires subject-wise individual model training, Bibliosmia achieves its performance without any additional training, making it more scalable and efficient. Similarly, while approaches like Textual Inversion and BLIP-Diffusion excel in prompt similarity, they compromise on identity consistency. Notably, even our Pre-Launch setup, which pairs GPT-4 without advanced prompt engineering with MMDiT, demonstrates competitive identity consistency (0.818) and solidifies Bibliosmia's foundation by outperforming methods like ELITE and IP-Adapter. The integration of prompt engineering and multimodal capabilities in Bibliosmia further elevates

prompt similarity while maintaining near-top-tier identity consistency, striking a critical bal-

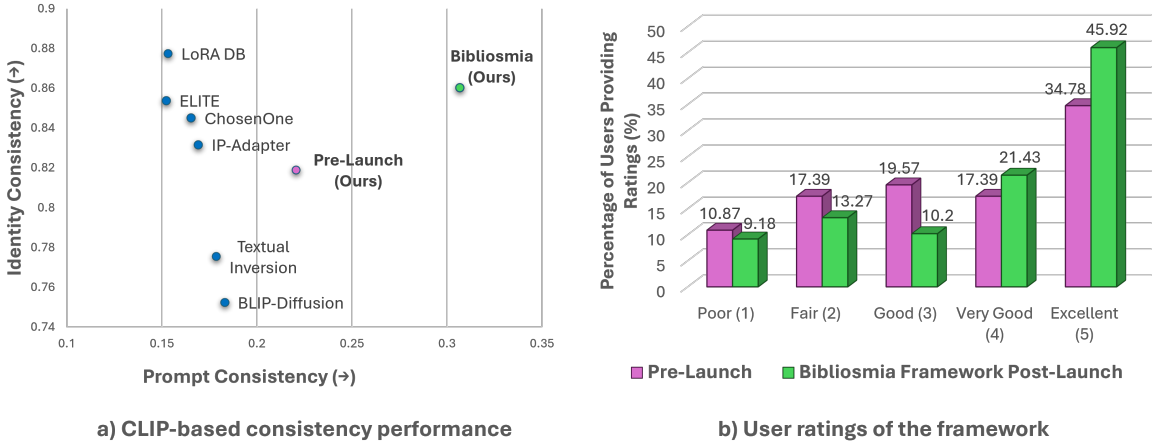

Figure 2: Quantitative comparison & User ratings: (a) An overview of Bibliosmia's prompt consistency and identity consistency performance compared to other methods and (b) User rating of Bibliosmia compared to pre-launch of the framework

ance between narrative fidelity and character coherence. This dual optimization aligns with user ratings (Figure 2(b)), which highlight the improvement in user satisfaction following the implementation of the Bibliosmia framework.

User Experience Rating The feedback for this study was collected from a user base composed of 60% professionals and 40% parents, who work with a diverse population of children aged 4 to 8 years. This population includes 63% male and 37% female children, with 46% identifying as White, 23% as Black, 18% as Hispanic, 4% as Arab and 9% from other backgrounds. Figure 2(b) captures feedback on the users' overall experience on the website through an in-app survey. Here, we can observe a growing preference for higher-rated experiences after the framework deployment, which underscores the importance of visual-textual harmony and consistent character portrayal for engaging storytelling. Before the framework's implementation (from May 6 to June 6, 2024), we gathered 46 ratings with an average of 3.5. Following the launch of the framework (from October 8 to November 8, 2024), 98 ratings were collected, with the average increasing to 3.8. This improvement represents a notable increase of 8.6% in user satisfaction. Additionally, the proportion of "good" ratings (4 and 5) rose from 52% to 67%, reflecting a 15% increase. These results indicate that the Bibliosmia framework substantially improved the consistency of generated images and the overall user experience. The decrease in lower ratings and increased proportion of positive feedback suggest that users found the updated system more reliable, engaging and effective in delivering high-quality AI-generated storytelling experiences.

Regeneration of Images We measured the frequency of image regenerations, a feature that allowed users to refine visual outputs by requesting specific images to be regenerated if the initial result did not meet their expectations - mostly driven by inconsistencies. To assess this feature, we focused on the Average Consumed Regenerations rate which is calculated as:

$$\text{Avg. Consumed Regenerations} = \left(\frac{\text{Total Regenerations Used}}{\text{Maximum Possible Regenerations}} \right) 100$$

Pre-Bibliosmia (Captions)	Pre-Bibliosmia (JSON)	Post-Bibliosmia (Captions)	Post-Bibliosmia (JSON)
Alistair is playing with blocks when his dad says it's almost time for his nap	{ "name": "Alistair", "age": 2, "gender": "male", "ethnicity": "hispanic", "accessories": [], "needs": [], }	Ruby woke up with excitement. It was Thanksgiving, and she couldn't wait to see all her family.	{ "character_name": "Ruby", "age": 7, "type": "main", "human": true, "nature": "", "gender": "female", "ethnicity": "white", "hair_color": "blonde", "hair_style":"straight", "hair_length":"below-shoulder", "hair_do": "in a bun", "eye_color": "blue", "accessories": "", "complexion_tone":"fair, "face_shape": "round", "facial_hair": "none", "body_type": "average" }
Alistair feels frustrated that play time is ending		But as the day went on, Ruby started to feel a little nervous. There would be so many people at the dinner. What if it was too loud?	
Dad reminds Alistair naps help him grow and gives him 5 more minutes to play		When they arrived at her aunt's house, Ruby saw the big table, the busy kitchen, and everyone talking and laughing. The noise felt overwhelming.	
When time is up, dad helps Alistair clean up the blocks		Ruby's heart began to race, and she felt a wave of worry. It was just too much. She wanted to find a quiet place to hide.	
Alistair starts to cry and says he wants to keep playing		"I don't know if I can do this," Ruby whispered to her mom, her voice trembling. "There are too many people."	
Dad gives Alistair a hug and reads him a story as they rock in the chair		Her mom gave her a warm hug. "It's okay to feel nervous, Ruby. Let's take a deep breath together and find a quiet spot for a few minutes."	
Alistair's eyes grow heavy as he drifts off to sleep, hugging his stuffed bear		Ruby and her mom found a cozy corner in the living room. They sat together, breathing slowly. Ruby started to feel a little calmer.	
When Alistair wakes, he feels refreshed and ready to play again		After a while, Ruby felt ready to go back. "I think I can do it now," she said, smiling at her mom.	

Figure 3: Story captions and JSON structure of children details within stories generated during pre-launch and post-launch of Bibliosmia framework

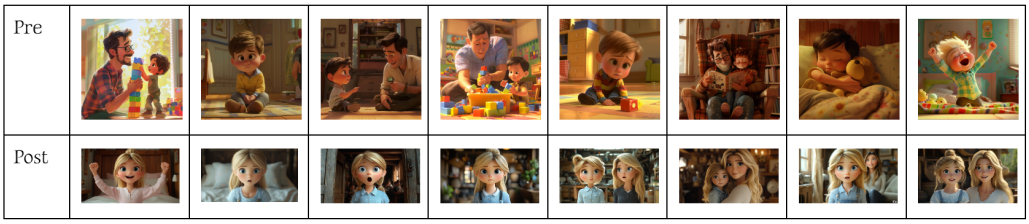

Figure 4: Qualitative comparison of stories generated during pre-launch and post-launch of Bibliosmia framework

where the Maximum Possible Regenerations is determined by multiplying the number of stories by the regeneration limit for the respective period. This metric provides a clear view of how users utilized the available regeneration capacity and allows direct comparisons between periods with different limits.

In the pre-launch period, when regenerations were capped at 4 per story, 2254 stories were created, allowing a maximum of 2254 $\times$ 4= 9016 possible regenerations. Users performed 250 regenerations during this time, resulting in an Avg Consumed Regenerations Rate of 2.77%. In the post-launch period, with an increased limit of 20 regenerations per story, 645 stories were created, allowing a maximum of 645 $\times$ 20 = 12,900 regenerations. During this period, users performed 262 regenerations, resulting in an Avg Consumed Regenerations rate of 2.03%. This decline occurred despite the significantly expanded re-

generation limit in the post-launch period, highlighting the effectiveness of Bibliosmia's improved visual generation framework. The enhanced architecture delivered higher-quality outputs on the first attempt, reducing regenerations and streamlining the creative process for a seamless user experience.

Qualitative Evaluation of Generated Stories To evaluate the effectiveness of our framework in generating consistent and personalized stories, we present two qualitative case studies, focusing on the progression from pre-Bibliosmia to post-Bibliosmia. As shown in Figure 3 and 4, each case includes a series of 8 images depicting key moments in the story, accompanied by the corresponding image captions and JSON objects describing the child for whom the story was tailored. In the pre-Bibliosmia framework, inconsistencies were evident as attributes such as the child's hair color, hairstyle, outfit and even age fluctuated between images, disrupting the narrative's coherence and diminishing personalization which is evident in Figure 4. In contrast, the post-Bibliosmia framework ensures that these attributes remain consistent across all images, aligning closely with the image captions and JSON specifications that encode demographic details, personality traits, preferences and developmental needs. This progression highlights how the Bibliosmia framework provides visually cohesive and contextually accurate storytelling, maintaining fidelity to the child's characteristics and enhancing alignment with developmental and cultural contexts.

5. Conclusion and Future Direction

In this paper, we introduced *Bibliosmia*, a comprehensive framework designed to enhance the consistency, coherence and of personalized AI-driven multimodal storytelling. By integrating advanced prompt engineering, validated character representations, state-of-the-art multimodal diffusion models and visual question and answering, the framework ensures consistency across narrative and visual components. Experimental results demonstrate Bibliosmia's ability to achieve a superior balance between prompt similarity and identity consistency, outperforming existing methods. Additionally, metrics such as reduced regeneration rates and increased user satisfaction validate its effectiveness in streamlining the storytelling process and delivering visually compelling narratives tailored to children's developmental needs. Bibliosmia's success highlights the transformative potential of AI in crafting deeply personal and emotionally resonant stories.

Overall, Bibliosmia demonstrates the responsible use of generative AI to enhance personalized learning experiences in education. Consistent storytelling plays a crucial role in maintaining engagement and fostering comprehension, making Bibliosmia a valuable tool for educational content. By ensuring both narrative and visual consistency in AI-generated stories, it supports SEL, enhancing cognitive and emotional development in children. Furthermore, by integrating multimodal AI models with the ethical and responsible use of language models for educational tasks, our framework advances the broader discourse on effective and responsible GenAI applications in education.

Future directions include expanding Bibliosmia's adaptability to diverse storytelling genres and cultural contexts, enabling real-time feedback for dynamic personalization and exploring advanced generative models to further enhance narrative depth and visual fidelity. This work paves the way for future innovations where AI becomes an even more integral partner in fostering social-emotional growth and creative engagement through personalized stories.

Acknowledgments

We sincerely thank Rex Duval for his generous support in providing resources and guidance, and Radha Manisha and Melanie Subbiah for their invaluable insights during the ideation phase. Their thoughtful discussions played a crucial role in shaping this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Victor Nikhil Antony and Chien-Ming Huang. Id. 8: Co-creating visual stories with generative ai. *ACM Transactions on Interactive Intelligent Systems*, 14(3):1–29, 2024.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

Brian Boyd. *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press, 2009.

Miguel A Cardona, Roberto J Rodríguez, Kristina Ishmael, et al. Artificial intelligence and the future of teaching and learning: Insights and recommendations. 2023.

Will Eisner. *Graphic storytelling and visual narrative*. WW Norton & Company, 2008.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Maya Gunawardena and Merja Koivula. Children's social–emotional development: The power of pedagogical storytelling. *International Journal of Early Childhood*, pages 1–22, 2023.

Alexander Gurung and Mirella Lapata. CHIRON: Rich character representations in long-form narratives. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8523–8547,

Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.499.

Melissa Allen Heath, Kathryn Smith, and Ellie L Young. Using children's literature to strengthen social and emotional learning. *School Psychology International*, 38(5):541–561, 2017.

Wayne Holmes, Maya Bialik, and Charles Fadel. *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign, 2019.

Hyunjo Kim, Jae-Ho Choi, and Jin-Young Choi. A novel scheme for managing multiple context transitions while ensuring consistency in text-to-image generative artificial intelligence. *IEEE Access*, 2024.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet$$++$$: Improving conditional controls with efficient consistency feedback. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 129–147, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72667-5.

Rose Luckin and Wayne Holmes. Intelligence unleashed: An argument for ai in education. *UCL Knowledge*, 2016.

Ryan Mathews and Watts Wacker. *What's your story?: storytelling to move markets, audiences, people, and brands*. FT Press, 2008.

Kimona Morrison. The impact of digital storytelling on the socio-emotional development of early elementary children. *Available at SSRN 4958816*, 2024.

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

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502, 2023.

Mark O Riedl and Robert Michael Young. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268, 2010.

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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, 2023.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Trans. Graph.*, 43(4), July 2024. ISSN 0730-0301. doi: 10.1145/3658157. URL https://doi.org/10.1145/3658157.

Ali Uslu and Nilüfer Atman Uslu. Improving primary school students' creative writing and social-emotional learning skills through collaborative digital storytelling. *Acta Educationis Generalis*, 11(2):1–18, 2021.

Yilin Wang, Haiyang Xu, Xiang Zhang, Zeyuan Chen, Zhizhou Sha, Zirui Wang, and Zhuowen Tu. Omnicontrolnet: Dual-stage integration for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7436–7448, June 2024.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.296. URL https://aclanthology.org/2022.emnlp-main.296.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024.

Chuanqi Zang, Jiji Tang, Rongsheng Zhang, Zeng Zhao, Tangjie Lv, Mingtao Pei, and Wei Liang. Let storytelling tell vivid stories: An expressive and fluent multimodal storyteller. *arXiv preprint arXiv:2403.07301*, 2024.