RESEARCH-ARTICLE

Personalized Generative Storytelling with AI-Visual Illustrations for the Promotion of Knowledge in Cultural Heritage Tourism

ANDREA FERRACANI, University of Florence, Florence, FI, Italy

MARCO BERTINI, University of Florence, Florence, FI, Italy

PIETRO PALA, University of Florence, Florence, FI, Italy

GABRIELE NANNOTTI, University of Florence, Florence, FI, Italy

FILIPPO PRINCIPI, University of Florence, Florence, FI, Italy

GIUSEPPE BECCHI, University of Florence, Florence, FI, Italy

Published: 28 October 2024

Citation in BibTeX format

MM '24: The 32nd ACM International Conference on Multimedia
October 28 - November 1, 2024
Melbourne VIC, Australia

Conference Sponsors:
SIGMM

SUMAC '24: Proceedings of the 6th workshop on the analySis, Understanding and proMotion of heritAge Contents (October 2024)
https://doi.org/10.1145/3689094.3689465
ISBN: 9798400712050

Personalized Generative Storytelling with AI-Visual Illustrations for the Promotion of Knowledge in Cultural Heritage Tourism

Andrea Ferracani
Marco Bertini
MICC - University of Florence
Firenze, IT
andrea.ferracani@unifi.it
marco.bertini@unifi.it

Pietro Pala
Gabriele Nannotti
MICC - University of Florence
Firenze, IT
pietro.pala@unifi.it
gabriele.nannotti@edu.unifi.it

Filippo Principi
Giuseppe Becchi
MICC - University of Florence
Firenze, IT
filippo.principi@unifi.it
giuseppe.becchi@unifi.it

Abstract

This paper presents a mobile application that exploits interactive narrative storytelling through GPT-4 and a custom image generative pipeline to improve cultural tourism experiences. The application helps tourists visiting cities to program and personalize cultural city tours creating stories with the user as the protagonist. The app guides the users to choose Point-Of-Interests (POIs) and narrative genres of the narrative while the image generation pipeline provides them with visual and coherent representations of their actions in the story contributing to a more immersive and personalized experience. Technical challenges include producing coherent stories and real-time and quality images, maintaining visual composition and person identity, including multiple concepts, through prompt engineering. We validate the effectiveness of the application and the image generative pipeline through users studies which evaluate the educational potential of our approach.

CCS Concepts

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → **Media arts**.

Keywords

AI, storytelling, cultural heritage, cultural tourism, stable diffusion, image generation, personalization, GPT-4

ACM Reference Format:

Andrea Ferracani, Marco Bertini, Pietro Pala, Gabriele Nannotti, Filippo Principi, and Giuseppe Becchi. 2024. Personalized Generative Storytelling with AI-Visual Illustrations for the Promotion of Knowledge in Cultural Heritage Tourism. In *Proceedings of the 6th workshop on the analySis, Understanding and proMotion of heritAge Contents (SUMAC '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3689094.3689465

1 Introduction

Narrative storytelling has proven to be effective in enhancing knowledge and learning [32] as it can evoke a high level of user engagement, known as 'state-of-flow' [22]. In the context of cultural tourism, storytelling serves as a means to strengthen the value proposition of locations and, consequently, enhance cultural experiences [3]. Literature preceding the advent of generative systems since 2019 has mostly focused on rule-based logical generation of narrative stories either manual or based on hand-crafted algorithms [21, 34, 37, 44]. More recently, scientific literature has concentrated on the study of systems providing storytelling through human-AI collaboration. Li *et al.* [19] conduct an exhaustive analysis of 60 articles examining the roles played by humans and AI at various levels. However, before the introduction of LLMs, even the top story generation systems, including those utilizing transformer architectures, had difficulty to produce coherent stories with well-developed characters and plots [2]. The past two years instead have seen the emergence of powerful generative systems based on LLMs (e.g., GPT-3.5, GPT-4 [1], GPT-4o[1], Llama2 [40]) and image generative models such as DALL-E [25, 26], Midjourney[2], stable diffusion [29], and GPT-4o, capable to provide a significant contribution to data storytelling due to their proficiency in natural language and text-to-image generation. These models, trained on vast datasets, have the ability to mimic the patterns typical of human writing and creativity [11, 33].

1.1 Interactive digital storytelling

Interactive digital storytelling has long been an important factor to the engagement, comprehension, and education of students in the field of cultural heritage and beyond, especially when provided through serious gaming [6, 28, 36] and, more recently, automated storytelling [16]. Trichopoulos [41] examines GPT's function as a digital storytelling tool, which can be trained and directed to serve as both a museum guide and a recommendation system. The author shows how GPT-4 represents a significant leap forward in AI-driven natural language processing, for its enhanced contextual understanding, improved coherence, and a broader knowledge base, and how it can be exploited to provide engaging stories that meet the specific interests and preferences of museum-goers. Colucci *et al.* [7] explore the possibilities offered by interactive conversation managed by virtual agents to enhance the user experience through personalized dialogues and contextualized information. Several automated conversational tools based on transformer models are reviewed, i.e. GPT, BERT [8], XLNet [46] and RoBERTa [20], showing how GPT (Generative Pre-trained Transformer), although it demands considerable computational power, is the best choice for the development of cultural heritage applications, ensuring versatility across different tasks [45].

SUMAC '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1205-0/24/10
https://doi.org/10.1145/3689094.3689465

[1] https://platform.openai.com/docs/models
[2] https://www.midjourney.com/home

1.2 Generative illustrations

The contribution of visual illustration of narrated stories in enhancing the comprehensibility and learnability of text is well known [5]. In this regard, the support provided by modern image generative models (IGMs) is crucial for their ability of seamlessly creating images and videos that align with the logical and temporal sequence of events described in the text [15, 39]. Consequently, these systems offer significant opportunities for automating story production, serving as illustrators where AI can orchestrate logically coherent stories, as demonstrated in [10, 14, 42]. Additionally, IGMs are now not only capable of semantically consistent illustrations but also of adapting to visual styles and user-specific preferences. However, generating consistent character representations across different contexts still remains a challenge. To fully immerse users in a story, engaging plots must be complemented by coherent and personalized visual imagery. Low-Rank Adaptation (LoRA) models have shown promising results in this area [12]. These models, trained on small, specialized image datasets, can be leveraged to maintain visual consistency throughout the narrative, promoting deeper immersion. Another important issue with IGMs is related to the implementation of real-time systems and applications. In fact, these systems can face performance bottlenecks when used at runtime. The issue stems from the extensive computing resources required, especially in custom training or fine-tuning [17, 48]. On the other hand, pre-trained IGMs, though flexible, struggle to compose multiple concepts together or to introduce personalizations based on user preferences such as inpainting a specific concept, such as the user face [24, 38]. To solve this issue, common approaches use deep generative models conditioned on classes, images, and text or exploit transfer learning for tuning whole models to single domains via either fine-tuning all the parameters as in DreamBooth [31] or introducing and optimizing a word vector through Textual Inversion [9]. Kumari et al. [18] propose a method for compositional fine-tuning of multiple concepts fine-tuning a subset of the cross-attention layer parameters, significantly decreasing the time required and outperforming DreamBooth and Textual Inversion. However, it is still difficult to generate new images with variations in poses, viewpoints, and backgrounds while maintaining stable an original concept (as in the case of face inpainting). Recently, She *et al.* [35] introduced InstantBooth, an innovative method for instantaneous, text-guided image personalization which doesn't require test-time fine-tuning. They propose an image encoder that converts input images into a global embedding to capture the general concept, and adapter layers that capture intricate identity details. The evaluation demonstrate better perceptual quality, vision-language alignment and identity preservation compared to DreamBooth, Textual Inversion, and ELITE [43]. InstantBooth achieves qualitatively comparable results with our pipeline, which, however, produces more predictable and consistent results while providing greater control over the image structure.

2 The Storytelling Application

This article introduces an innovative mobile application[3] that utilizes LLMs-driven personalized storytelling accompanied by AI-generated illustrations to enhance tourism experiences. The main objective is to demonstrate the feasibility of using generative text and AI-driven illustrative images in the context of a multimedia application that exploits interactivity, personalization and immersiveness to enhance user experience and learning. The application acts as a multimedia-enriched tour guide, leading users through selected POIs in the cities of Rome, Florence and Venice (IT) via an AI-generated narrative. Users can customize the story by choosing specific locations as the story setting, selecting preferred genres, defining the protagonist's appearance from selfies, with the narrative adapting dynamically based on their choices. Possible genres include adventure, history, romance. POIs are enriched with multimedia content: descriptions, panoramic images, image carousels, videos, maps. User interactions involve explicit AI-generated choices that suggest story progression (Fig. 1.2), so providing immersive experiences not only tailored to user preferences but which have the user as the main agent. The story is generated through a dynamic template-based prompt that takes user configurations from the app as input to the ChatGPT API. The responses are encoded in JSON. Hallucinations are limited through RAG, providing descriptions of selected POIs as context. Furthermore, the application features a subsystem for image generation and personalization, where a custom AI-powered pipeline generates images depicting the user as the protagonist within the narrative. The multi-step image generation pipeline is designed for fast image generation, high-quality composition, multiple concepts incorporation, configurability of locations and themes, user's face inpainting, structure and identity preservation with the aim of enhancing the sense of immersion. In this regard, our approach, which combines several task specific generative models, outperforms single solutions requiring high-cost resources and extensive fine-tuning, reducing randomness, ensuring consistent visual style and better alignment with the story's events, and enhancing the overall narrative cohesion.

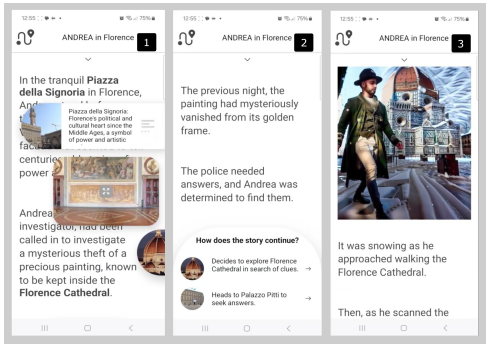

Figure 1: The app: 1) Generated story with contextual POIs: descriptions accessible through sliding cards contextual to POIs names; 2) Story progression choices; 3) An example of a generated image contextual to the narrative.

As an example let's consider a tour of Florence, IT, as shown in Fig. 1. The app suggests several POIs within the city. Next, the user chooses the story genre, e.g. a thriller. Then he selects a third-person narrative perspective. Finally, he enters his name and takes the selfies, which will be used to visually represent him in the illustrative images. Based on these inputs, the system generates a story. E.g.: the

[3]A video of the app is available at https://tinyurl.com/64fk2aur, source code is on GitHub at https://tinyurl.com/mv8swvj6

protagonist, modeled after the tourist, is depicted as a detective investigating the theft of a famous painting housed in the Florence Cathedral, in the 18th century. It is snowing, and the detective decides to visit the cathedral for an inspection. The automatically generated image, in Fig. 1.3, shows the detective/tourist, dressed in a 18th century attire, walking near the Florence Cathedral.

3 The image generation pipeline

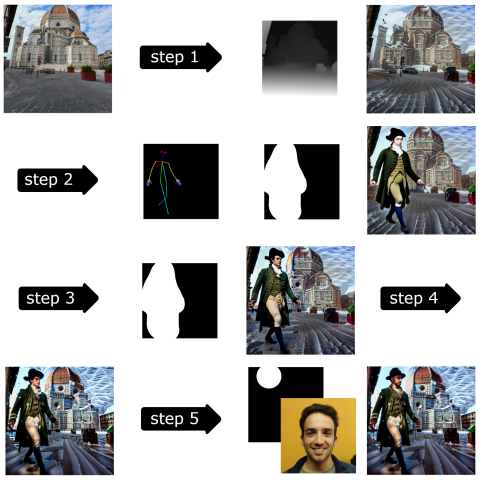

Figure 2: Text to Image generation pipeline in five steps.

The image generation pipeline, shown in Fig. 2, exploits the `Diffusers` library[4] to generate an image through the following steps, starting with a predefined image of a location: step 1) modifying the location theme, meteorological conditions and daytime/nighttime; step 2) inpainting a person in any pose using a control image; step 3) enhancing photo-realism and image quality; 4) enhancing image homogeneity 5) inpainting the face of the app user. The pipeline is invoked asynchronously including in the prompt the text of the main event of each story section, detected through the ChatGPT API. **Step 1** of the pipeline takes a base image and a computed depth map [27] to output a transformed image with a new theme, daytime and weather conditions. It uses two models: ControlNet, which utilizes the depth map to understand the structure and layout of the base image [47], and a photorealistic image generation model[5]. An high value of strength is used to give more importance to the prompt with respect to the image in case that both a different theme and a new weather condition setting are given. In this step, a process can also be activated to transform daytime images into night-time versions, using instruct-pix2pix [4]. The theme is passed as a parameter to the pipeline based on the user's genre selection, while the weather conditions and the distinction between day and night are extracted through requests to OpenAI chat completions API. In **Step 2**, the process involves subject inpainting with a specific pose. This step uses a model specifically trained for inpainting [30], which is crucial because not all models are designed for this task without significantly impacting the background where the subject is inserted. The diffuser is conditioned by a mask image to insert the subject in a specified area, enabling flexibility in determining the field of view, whether medium or long. Additionally, a ControlNet is again used to define the subject's pose or action, providing further guidance for the inpainting process. The pose and actions performed by the character are again extracted through the OpenAI API, while the masks fed to the pipeline as a constraint are images that represent the most common types of atomic and behavioural actions such as smiling, crying, waving, running, walking [23]. In **Step 3**, the objective is to enhance the photorealism of the inpainted subject. In fact, in **Step 2**, while the initial inpainting model is effective at preserving the background, it often lacks the ability to produce high-quality, realistic subjects. To tackle this, this step uses a different approach with a focus on photorealistic content generation. This involves applying a mask to define the specific area where the subject will be refined. The mask ensures that the process remains focused on the subject, minimizing unintended alterations to other parts of the image. Careful control of the generation strength is required to avoid significant impact on the surrounding background and to maintain the original pose. **Step 4** aims to smooth the transition between the inpainted subject and the original background exploiting the same model. After enhancing the photorealism of the subject in previous step, a noticeable boundary can appear between the altered area and the untouched background. In this step, no mask is used. Instead, the whole image is processed to blend the subject with the background. A low strength setting is applied to gently harmonize the entire image without affecting its overall content, reducing the visibility of any inpainting edges. **Step 5** focuses on inpainting the user's face onto the previously generated character. This step involves combining a LoRA (Low-Rank Adaptation) model [13] trained on the user's face with a stable diffusion model [30] to achieve a realistic integration of the user's features. A specific mask around the character's head is provided to the model. The images used to train the LoRA model with the user's facial features are captured during the initial configuration of the story in the app, which prompts the user to takes five selfies from different viewpoints (enough to focus model's learning on the particular characteristics of the user).

4 App evaluation

To test the quality and effectiveness of the application, we conducted three user tests: the first, in Subsec. 4.1 evaluates the perceptual quality of the generated images; the second evaluates the degree of satisfaction, engagement, and immersiveness of the app through a comparative study among user groups (Subsec. 4.2); the third, in Subsec. 4.3, assesses the differences in learning outcomes among the three groups regarding the acquisition of cultural knowledge about the POIs presented in generated stories.

4.1 Image quality assessment

We conducted a user study to perceptually evaluate the results of our image generation method. For each evaluation, each user was shown two input images (the user face and the POI's image that will be used as the background), the prompt, and the image generated by our pipeline. The user ranked the generated image on a Likert scale from 1 (worst) to 5 (best) based on its visual quality, vision-language alignment (adherence to the prompt), and identity preservation. We selected 10 identities from the "person" category, with each identity personalized using 10 prompts resulting in 100 unique

[4]https://huggingface.co/docs/diffusers/
[5]https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0

evaluation samples (some examples for an identity are shown in Fig. 3). The study was deployed via Amazon Mechanical Turk, with at least 10 samples evaluated by 20 users. The results, shown in Table 1, indicate the following average scores: The relatively

Aspect	Average Score
Visual quality	2.8
Vision-Language alignment	4.2
Identity preservation	4.7

Table 1: Evaluation results for the image generation and personalization pipeline.

low score in visual quality can partially be attributed to the use of a low mid-range GPU which prompted us to opt for a rather low resolution of the input images, balancing performance with available hardware resources. However, the method demonstrated a notable capability in achieving good vision-language alignment and preserving identity, highlighting its potential.

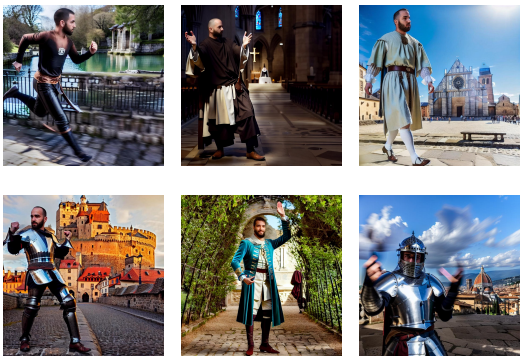

Figure 3: Some personalized images generated for a story

4.2 Satisfaction, engagement, immersion

Three groups of 10 participants each were recruited for the study. Group 1 (G1, Printed) used only printed-like descriptions and contextual images of the points of interest. Group 2 (G2, App Non-Interactive) used the app without interactive AI-generated story progression choices and image personalization. Group 3 (G3, App Interactive) used the app with interactive AI-generated choices for story progression and personalized images from the generation pipeline depicting users as the protagonists. App users in G2 and G3 generated a story session, while users in G1 followed an itinerary with a sequential description of the points of interest along a route. For both G1 and G2, we provided photographic images of the places or contextual images generated by Chat-GPT using story textual information as the prompt. Participants were evenly distributed across demographic variables to ensure a balanced representation. After the session, participants completed a questionnaire evaluating their satisfaction, engagement and immersion. Responses were recorded using a 1 to 10 point scale. **Results**: the data were analyzed to compare the three groups' experiences. The results are summarized in Table 2. G3 (App Interactive) reported significantly higher immersion (M=7.8) compared to G1 (Printed) (M=2.9) and G2

Measure	G1 (P)	G2 (ANI)	G3 (AI)
Satisfaction	6.4	5.5	5.8
Engagement	3.4	5.5	8.3
Immersion	2.9	6.5	7.8

Table 2: Comparison of satisfaction, engagement and immersion between the Printed material (G1), App Non-Interactive (G2) and App Interactive groups (G3).

(App Non-Interactive) (M=6.5), indicating that interactive elements and personalized images greatly enhance user self-identification in the story. Satisfaction scores were relatively similar across the groups: G1 (M=6.4), G2 (M=5.5), and G3 (M=5.8). Engagement was reported highest in G3 (M=8.3), followed by G2 (M=5.5) and G1 (M=3.4). To determine the statistical significance of the differences between the groups, an ANOVA test was performed which showed a statistic relevance for engagement ($F(2, 27) = 28.63$, $p < 0.001$) and immersion ($F(2, 27) = 27.61$, $p < 0.001$). These findings indicate that incorporating interactivity and AI-generated visual elements significantly enhances user immersion and engagement. However, satisfaction did not show a statistically significant difference across the groups, suggesting that the observed scores might be due to random variation rather than a systematic effect of interactivity and AI-generated illustrations.

4.3 Learning outcomes assessment

ANOVA analysis was also used to compare the scores obtained by the three groups (5 participants selected from groups in Subsec. 4.2) in a cultural heritage multi-choices questionnaire (three choices) of 10 questions on the history of 10 POIs present in stories generated by users in Group 3 (G3, AI). The 5 questionnaires with the highest average score in each group were selected in order to exclude users who had not fully understood the task and had not read all the information about the POIs in the app. The provided mean scores were Group 1 (G1, P): $n = 5, M = 5.8$, Group 2 (G2, ANI): $n = 5, M = 4.5$, and Group 3 (G3, AI): $n = 5, M = 7.8$. We obtained an F-value of 36.5 and a p-value less than 0.05. The statistically significant difference between the three groups indicates that the interactive and visual elements enhance educational aspects more effectively than both printed materials (G1) and the non-interactive app (G2). Further investigation should be dedicated to the relatively low score of Group 2. Nonetheless, it should be noted that the groups are small, and the app should be evaluated in future research with a larger number of users.

5 Conclusions

The article presented a mobile app for cultural tourism proposing an innovative approach to enhancing cultural heritage experiences and learning through personalized AI-based storytelling. As indicated by the user studies we carried out, the system achieves a good level of immersion and engagement through AI-generated narratives and images visually representing users in the context of the stories. As demonstrated by an initial evaluation, the solution has the potential to promote learning in cultural heritage tourism and to offer a novel, interactive way for users to engage with their surroundings.
Acknowledgments: this work was partially supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01 00034 23154 SMST.

References

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

[2] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic story generation: Challenges and attempts. *arXiv preprint arXiv:2102.12634* (2021).

[3] Clara Bassano, Sergio Barile, Paolo Piciocchi, James C Spohrer, Francesca Iandolo, and Raymond Fisk. 2019. Storytelling about places: Tourism marketing in the digital age. *Cities* 87 (2019), 10–20.

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv:2211.09800 [cs.CV]

[5] Russell N Carney and Joel R Levin. 2002. Pictorial illustrations still improve students' learning from text. *Educational psychology review* 14 (2002), 5–26.

[6] Vanessa Cesário, Sandra Olim, and Valentina Nisi. 2020. A natural history museum experience: memories of carvalhal's palace–turning point. In *International Conference on Interactive Digital Storytelling*. Springer, 339–343.

[7] Luigi Colucci Cante, Beniamino Di Martino, Mariangela Graziano, Dario Branco, and Gennaro Junior Pezzullo. 2024. Automated Storytelling Technologies for Cultural Heritage. In *International Conference on Emerging Internet, Data & Web Technologies*. Springer, 597–606.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

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

[10] Elia Gatti, Daniele Giunchi, Nels Numan, and Anthony Steed. 2024. AIsop: Exploring Immersive VR Storytelling Leveraging Generative AI. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 865–866.

[11] Yi He, Shixiong Cao, Yang Shi, Qing Chen, Ke Xu, and Nan Cao. 2024. Leveraging large models for crafting narrative visualization: a survey. *arXiv preprint arXiv:2401.14010* (2024).

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. http://arxiv.org/pdf/2106.09685 cite arxiv:2106.09685Comment: Draft V2 includes better baselines, experiments on GLUE, and more on adapter latency.

[14] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. 2024. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems* 36 (2024).

[15] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on visualization and computer graphics* 19, 12 (2013), 2406–2415.

[16] Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex'Sandy' Pentland, Yoon Kim, Jad Kabbara, et al. 2024. Leveraging Large Language Models for Learning Complex Legal Concepts through Storytelling. *arXiv preprint arXiv:2402.17019* (2024).

[17] Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez. 2023. Chatbot-based natural language interfaces for data visualisation: A scoping review. *Applied Sciences* 13, 12 (2023), 7025.

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

[19] Haotian Li, Yun Wang, and Huamin Qu. 2024. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *preprint arXiv:1907.11692* (2019).

[21] Lara J Martin, Brent Harrison, and Mark O Riedl. 2016. Improvisational computational storytelling in open worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*. Springer, 73–84.

[22] Sean McKenna, Nathalie Henry Riche, Bongshin Lee, Jeremy Boy, and Miriah Meyer. 2017. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 377–387.

[23] Md Golam Morshed, Tangina Sultana, Aftab Alam, and Young-Koo Lee. 2023. Human action recognition: A taxonomy-based survey, updates, and opportunities. *Sensors* 23, 4 (2023), 2182.

[24] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620* (2023).

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. Pmlr, 8821–8831.

[27] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12179–12188.

[28] Selma Rizvic, Dusanka Boskovic, Vensada Okanovic, Sanda Sljivo, and Merima Zukic. 2019. Interactive digital storytelling: bringing cultural heritage in a classroom. *Journal of Computers in Education* 6 (2019), 143–166.

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

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

[32] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1139–1148.

[33] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. 2021. Saga: Collaborative storytelling with gpt-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 163–166.

[34] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2020. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 453–463.

[35] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8543–8552.

[36] Andy Smith, Bradford Mott, Sandra Taylor, Aleata Hubbard-Cheuoua, James Minogue, Kevin Oliver, and Cathy Ringstaff. 2020. Toward a block-based programming approach to interactive storytelling for upper elementary students. In *Interactive Storytelling: 13th International Conference on Interactive Digital Storytelling, ICIDS 2020, Bournemouth, UK, November 3–6, 2020, Proceedings 13*. Springer, 111–119.

[37] Ingibergur Stefnisson and David Thue. 2018. Mimisbrunnur: AI-assisted authoring for interactive storytelling. In *Proceedings of the AAAI Conference on AI and Interactive Digital entertainment*, Vol. 14. 236–242.

[38] Andreas Stöckl. 2022. Natural language interface for data visualization with deep learning based language models. In *2022 26th International Conference Information Visualisation (IV)*. IEEE, 142–148.

[39] Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952* (2023).

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[41] Georgios Trichopoulos. 2023. Large language models for cultural heritage. In *Proc. of the 2nd International Conference of the ACM Greek SIGCHI Chapter*. 1–5.

[42] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. 2023. Autostory: Generating diverse storytelling images with minimal human effort. *arXiv preprint arXiv:2311.11243* (2023).

[43] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15943–15953.

[44] Jon Womack and William Freeman. 2019. Interactive narrative generation using location and genre specific context. In *International Conference on Interactive Digital Storytelling*. Springer, 343–347.

[45] Xinran Yang and Ilaria Tiddi. 2020. Creative Storytelling with Language Models and Knowledge Graphs.. In *CIKM (Workshops)*.

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[47] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]

[48] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239* (2023).