Cinema Meowdiso: Films Co-Created by Human, Large Language Models and a Cat

Xuanyang Huang
Infomation Hub, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology(Guangzhou)
China
xhuang383@connect.hkust-gz.edu.cn

Xiaoyun Zhong
Infomation Hub, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou)
China
xzhong204@connect.hkust-gz.edu.cn

Zhihua Zhong
Infomation Hub, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou)
China
zzhong839@connect.hkust-gz.edu.cn

Theodoros Papatheodorou
Infomation Hub, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou)
China
theodoros@hkust-gz.edu.cn

Kei-Man Yip
Infomation Hub, Computational Media and Arts Thrust, The Hong Kong University of Science and Technology (Guangzhou)
China
daveyip@hkust-gz.edu.cn

Figure 1: The protagonist cat watches the Western movie it stars in. (©Xuanyang Huang and Xiaoyun Zhong)

SA Art Papers '24, December 03–06, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1133-6/24/12
https://doi.org/10.1145/3680530.3695443

Abstract

Cinema Meowdiso is a system that utilizes generative AI to transform daily footage captured by pets into short films across various genres. This project explores the potential of non-human entities as narrative agents and examines how AI can autonomously create stories with human guidance. An automated visual story generation system based on pets' perspective is proposed, aiming to provide a novel fictional narrative experience that resonates with real-life footage.

CCS Concepts

• **Applied computing** → **Media arts**;

Keywords

AI Art, Generative AI, LLMs, Narrative, Cinema

ACM Reference Format:

Xuanyang Huang, Xiaoyun Zhong, Zhihua Zhong, Theodoros Papatheodorou, and Kei-Man Yip. 2024. Cinema Meowdiso: Films Co-Created by Human, Large Language Models and a Cat. In *SIGGRAPH Asia 2024 Art Papers (SA Art Papers '24), December 03–06, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3680530.3695443

1 INTRODUCTION

Cinema Meowdiso explores the creative potential of visual narratives generated by large language models (LLMs) from non-human perspectives. This automated system collaborates with cameras, LLMs, and a cat. It transforms the cat's recorded life footage into fictional narrative experiences through multimodal transformations. Leveraging the animal's viewpoint and the combined creativity of humans and artificial intelligence (AI), it co-creates and produces short films. These synthesized short films are broadcast to pet parents, featuring their cat as the protagonist and showcasing the cat's stories across various film genres.

We found inspiration for this project last year when we placed an action camera on a domestic cat's collar, capturing videos that revealed unique perspectives. This experiment drew parallels to Dr. Julius Neubronner's pigeon camera [DenHoed 2018], an early exploration that featured unconventional perspectives and photographic methods featuring non-human entities. While reviewing the footage recorded by our cat, we developed a profound curiosity about the potential narratives and experiences from the cat's perspective. We posed the following research questions: If a cat were the protagonist of a story, what story would it experience? Can we augment or transform the pet's real-life experiences into fictional storylines comprehensible to humans? This initial inquiry led us to develop *Cinema Meowdiso.*

We propose a workflow integrating external perception and AI. Cat-recorded videos are converted to textual data, serving as references for generating fictional stories. Using prompt engineering, a LLM transforms this data into various film genres, adhering to conventional dramatic structures. A Text-to-Image tool then creates visuals based on the processed information. The result is an automatically edited short film sent to the pet parent. This method explores the boundaries between real and fictional narratives, creating novel experiences by combining pet curiosity, AI capabilities, and cinematic knowledge from humans.

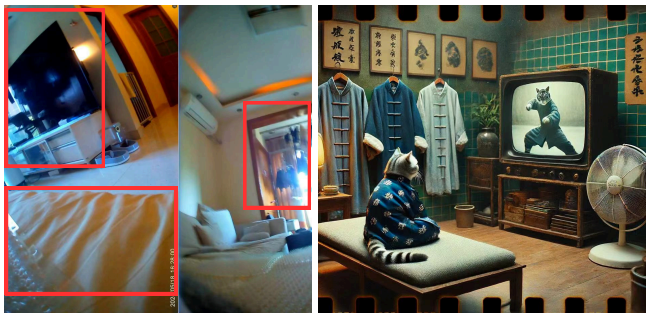

Figure 2: The visual elements of the real-world scenes documented by the cat correspond to the AI-generated elements in the shots of a Kung Fu film. (Objects labeled in the red frames, such as the television, blankets, and clothes, are retained in the AI-generated image) (©Xuanyang Huang and Xiaoyun Zhong)

2 BACKGROUND AND MOTIVATION

2.1 Non-Human Perspectives in Film-making

Recent advancements in photography technology, such as CCTV and drones, have integrated automated cameras into everyday applications and content creation. Unlike traditional cameras, these automated cameras operate independently, producing images from non-human perspectives. This shift aligns with Donna Haraway's concept of partial perspective [Haraway 1988] and Joanna Zylinska's discourse on non-human photography, which suggests a move away from human-centric visualization [Zylinska 2017]. Automated cameras represent a transformative shift in photography, challenging traditional notions of authorship and visual representation.

Detaching from anthropocentric perspectives involves considering the viewpoints of other species. Filmmakers have used miniature cameras on animals to capture their experiences. The BBC's *Spy in the Wild* [Downer 2017] uses cameras disguised as robotic animals to document wildlife behavior. Jana Sterbak's video art installation *Waiting for High Water* [Sterbak 2024] captures Venice during a flood from a dog's perspective. The documentary *Leviathan* [Castaing-Taylor and Paravel 2020] by Lucien Castaing-Taylor and Véréna Paravel employs multiple non-human camera perspectives to immerse viewers in the fishing process. These approaches reveal novel narrative techniques, similar to our project's methodology of imagining from pets' perspectives to uncover new narrative potentials through computational media.

2.2 Automated Narratives Based on Real-world Data

Computational cinema has driven our exploration of automated audiovisual content creation, viewing modern films as editable data streams that algorithms can analyze and deconstruct [Manovich 2002]. Artists use this database-driven approach in generative cinema, employing algorithms to remix film content and alter narrative structures or generate images [Grba 2017].

Artificial intelligence, particularly LLMs, accelerates story and video production, enabling users to generate storylines, images, and videos from natural language prompts. This enhances the creation of new multimodal content, enriching visual content and storytelling [Audry 2021]. Recent research on LLMs in generative storytelling explores applications like engaging adolescents with fantasy games [Steenstra et al. 2024] and supporting creative writing with chatbots [Qin et al. 2024]. These uses highlight LLMs' narrative capabilities and broad applicability in automated storytelling.

Although LLMs can generate diverse content from extensive databases, this content often fails to accurately represent real-life objects. When creators use natural language in text-to-image tools to depict real-world scenes, the resulting images lack visual details that correlate with the real world. This limitation arises from LLMs' reliance on natural language descriptions and the compression of textual information. LLMs typically grasp the literal meaning of prompts but miss deeper implicit meanings, leading to a lack of indexicality in the generated content [McCormack et al. 2023]. Consequently, we are exploring ways to incorporate real-world data as references for LLMs content creations.

We drew inspiration from digital artists who integrate real-world data into fictional narratives. James Coupe's *Jalousie Room* [Coupe 2014] uses computer vision algorithms to analyze pedestrians, incorporating this information into the storytelling. Ross Goodwin's *1 the Road* [Miller 2019] inputs data from a car's sensors into a neural network to automate novel writing. Similarly, *Narratron* [Zhao and Bao 2023] uses hand gesture recognition to trigger LLMs for constructing shadowplay visual stories. These works blend real-world perceptual media with fictional narratives, transforming data into stories. By combining real-world input with narrative techniques, these projects explore machine storytelling's potential and emphasize the emotional impact of contrasting fiction with real-world data, aligning with our creative objectives.

3 TECHNICAL RESEARCH

The overall framework of *Cinema Meowdiso* is divided into three modules: Perception, Production, and Play (see Figure 3). The Perception and Production modules rely on multimodal conversions managed by multiple prompt sequences, each handling specific tasks. We use different prompts to guide LLM to implement complex functions. Once these two modules are completed, the Play module uses an automated workflow to finalize the composition of visual and story content and distribute the video episode.

3.1 Perception

We used a compact action camera with a maximum resolution of 1920 x 1080 pixels and enlisted our cat, Whisker, as the cinematographer. The action camera was attached to a pet collar provided by the camera manufacturer. We ensured that both the camera setup and collar were appropriately sized for the cat's comfort and safety.

After the cat roamed around the house with the camera, FFmpeg was used to extract the video stored on the SD card into groups of frame segments. Each segment lasted five seconds, capturing one frame per second, resulting in a set of five frames. This system ensured continuity and coherence of instantaneous content and events. Over an hour of recording, 12 independent segments were randomly selected, constituting the raw material for the video output. In total, 12 to 15 sets of images were extracted, amounting to 60 to 75 images, serving as references for subsequent image analysis.

To analyze and summarize the captured video content, two GPT-4o prompt sequences were employed: Recognition and Selection. The Recognition sequence focused on analyzing and extracting information from video frames, identifying and categorizing potential events and objects within the video. It categorized video content into three types: background description, item description, and overall description of each keyframe, recording details such as lighting, atmosphere, name, position, materials, sizes, and possible actions. These elements were compiled in real-time into a JSON file.

The JSON file, along with the original images, was then sent to the Selection. The Selection filtered and compressed the information, linked temporal sequences, and inferred potential events. It refined the background details, object information, and the potential movement trajectory of the protagonist cat within each set of five frames. Furthermore, it summarized the content of each set of five frames into factual textual descriptions. Including the original images as supplementary information ensures comprehensive and effective utilization of the information and connection from the original frames.

3.2 Production

In the production phase, the primary objective is to guide AI in creating stories and visual content. This phase employs three prompt sequences and an Image-to-Video generation model, referred to as Screenplay, Storyboard, Text-to-Image, and Image-to-Video. The Screenplay is responsible for writing plots and generating intertitles. Subsequently, the Storyboard realizes scenes and props. The Text-to-Image tool then transforms these stories into images, which are finally converted into videos by the Image-to-Video tool.

The Screenplay performs three primary functions: crafting story outlines, linking inter-shot content, and expanding the data from the Perception stage into genre-specific visual and plot components. Six visually distinct genres were selected: science fiction, kung fu, western, detective, gangster, and horror, each with a corresponding prompt template. These genres have visually identifiable features, making them more suitable for generating stories with identifiable protagonists, settings, and objects. GPT-4o was instructed to utilize John Truby's premise writing format from *The Anatomy of Story* [Truby 2007] and Aristotle's three-act structure from *Poetics* [Aristotle 1996] to ensure narrative coherence. The Screenplay generates content based on factual JSON files from the perception stage and incorporates transitional inter-titles to enhance story continuity.

A core function of the Screenplay involves associating and imagining the real-world object and background attributes recorded in the JSON documents summarized by Perception. We instructed it to retain physical attributes such as size, quantity, and dimensions, but let GPT-4o semantically re-imagine the names of the objects and backgrounds to match the corresponding movie genre. As illustrated in Figure 4, an orange and the box containing it, recorded by the camera, were re-imagined into props or characters

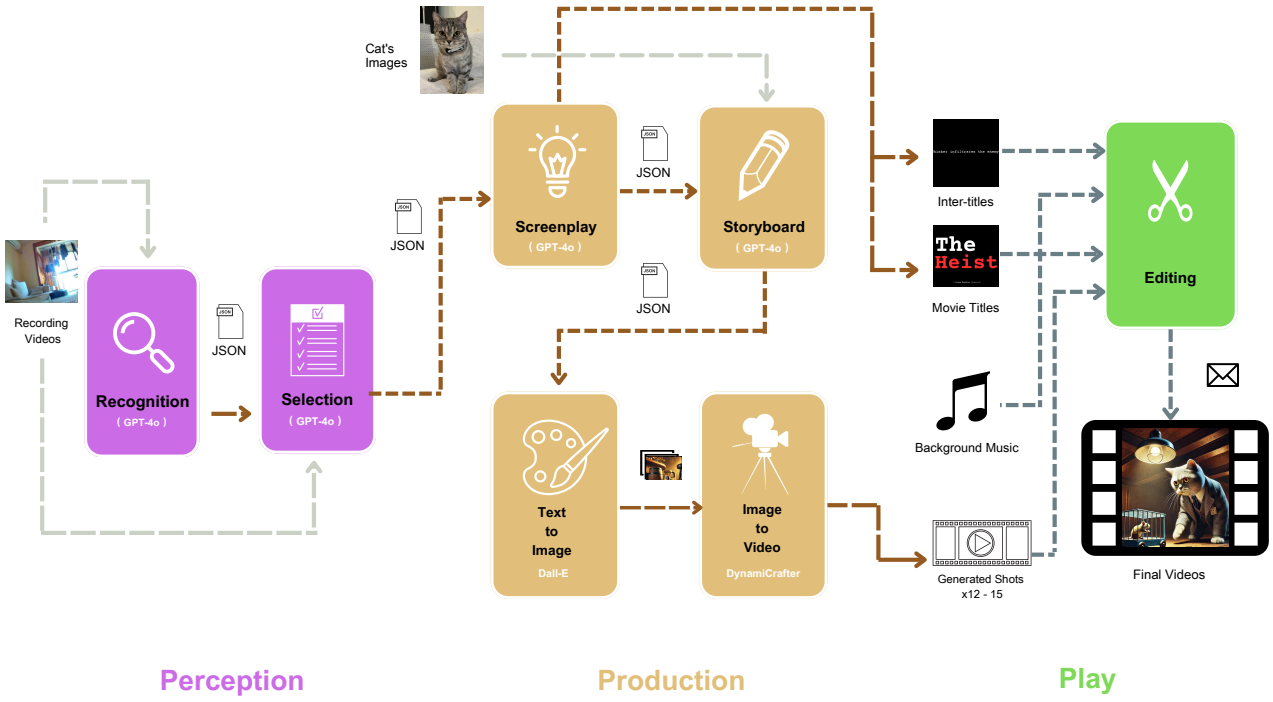

Figure 3: The technical pipeline of *Cinema Meowdiso* consists of three modules: Perception, Production, and Play. (©Xuanyang Huang and Xiaoyun Zhong)

relevant to different movie genres. The characteristics of the orange and the box were abstracted into objects with quantity and spatial functions. After the Screenplay's imaginative process, they were transformed into different types of objects that still retain their quantitative or spatial carrier functions. Background information such as atmosphere and lighting was preserved, while the environment's name was left to Screenplay to imaginatively associate with scenes from the corresponding movie genre. Real-world objects and background information are transformed into narrative elements within different narrative spaces through the semantic conversion of GPT-4o. Subsequently, the Storyboard phase will imagine more detailed attribute descriptions for these objects and background information.

Next, the Storyboard is responsible for providing detailed descriptions for image generation and handling the depiction of the protagonist cat. First, it converts the story and object information re-imagined by the Screenplay into a JSON file for text-to-image processing. Then, it summarizes the necessary information for each storyboard image frame, including details of all entities, such as background descriptions and objects, to form text-to-image prompts.

The Storyboard module manages the cat's appearance and attire, extracting detailed attributes from a photo using GPT-4o. We instruct GPT-4o to extract textual information about the cat's breed, fur color, pupil size, body size, age, and other detailed attributes from a single full-body photo of the cat. It also modifies the cat's accessories and outfits to fit the film genre. This process aims to reproduce the cat's appearance in the final video. Additionally, the Storyboard enriches objects and environments imagined by the Screenplay, transforming real-world attributes into story-specific properties. While the Screenplay focuses on genre-based object correspondence, the Storyboard adds details to these elements. This semantic mapping of real-world objects into fictional events enhances the detail in generated visuals and facilitates subsequent image generation.

Once the cat's appearance and scene information have been processed, the data is integrated into a JSON file and sent to the text-to-image module for processing. This JSON file includes the image information for each storyboard frame and is handed over directly to DALL·E, integrated within GPT-4o, for image generation. Apart from adding shot information such as shot type and focal length, no additional prompts or keywords were included. For each video, 12-15 static images are ultimately generated for subsequent image-to-video processing.

The final step in production involves converting images into videos using the DynamiCrafter Image-to-Video model [Xing et al. 2023]. The camera movement and animation intensity were minimized, with the video frame rate set to 4 frames per second and each video segment lasting approximately 6 to 8 seconds. No new prompts were added during video generation to reproduce the

Figure 4: The Screenplay module imagines image sequences of oranges as different types of movie props in various film genres. (©Xuanyang Huang and Xiaoyun Zhong)

created content. These video sequences are stored locally and sequenced according to the storyboard for final post-production processing in the Play module.

3.3 Play

The Play module is the final stage where the synthesized film is composed and played. In this phase, an automated system finalizes the film composition. Video clips are sequenced, and inter-titles from the Screenplay are inserted to maintain plot continuity. Each film is about 2 minutes and 30 seconds long.

Artistic treatments give the videos a unified style reminiscent of early silent films, including adding noise for film grain, selecting specific fonts for titles and inter-titles, and incorporating custom opening and closing sequences. Each film genre is pre-assigned five background music tracks of a unified style. The completed films are sent to the pet owner with video links sent via email.

4 DISCUSSION AND FUTURE WORKS

In *Cinema Meowdiso*, the cat's role transforms from a pet to a co-creator involved in the creation of the work. It serves as both the cinematographer and the protagonist in the film. We use LLMs as the medium for content expansion, augmenting and imaginatively extending the records made by the pet. The cat's daily movements and the objects it observes act as random elements and occurrences, generating greater narrative diversity and adding an element of unpredictability to the generated content.

The system integrates the cat's image into the generated films. LLMs are utilized for semantic interpretation and data transformation, with GPT-4o employed to imaginatively translate the owner's familiar scenes into various cinematic settings and objects. This process transforms familiar cats and scenes into diverse types of stories, thereby creating an intriguing correspondence between the fictional and real worlds. Additionally, *Cinema Meowdiso* introduces boundless narrative possibilities. AI-generated procedural content continually innovates, and although it operates within

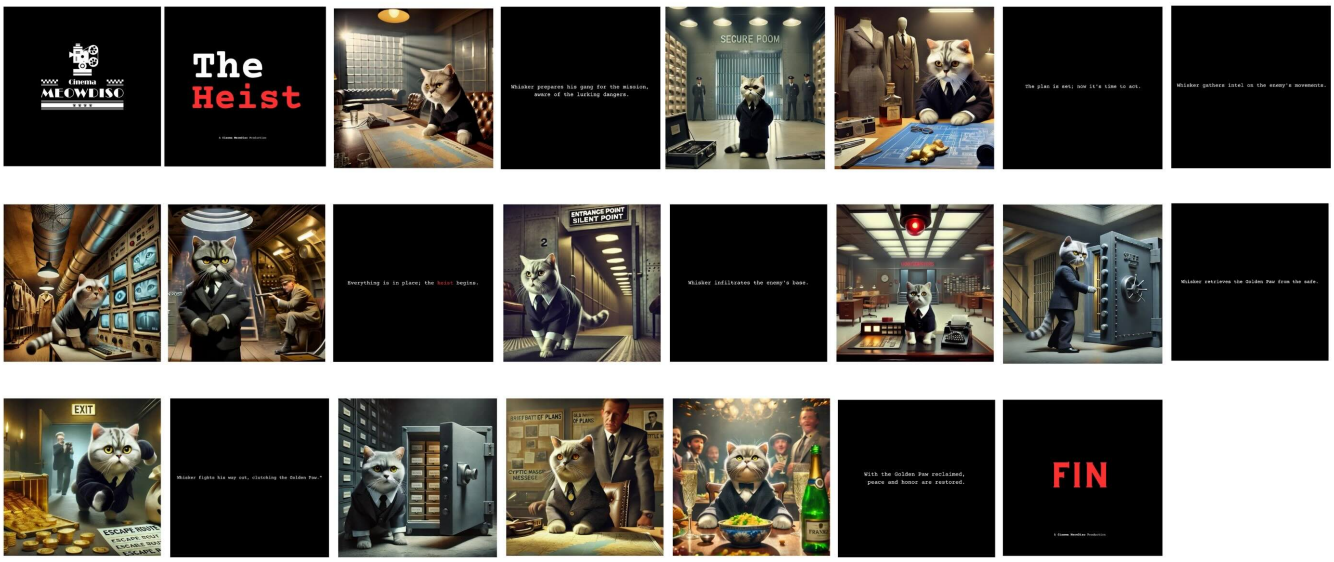

Figure 5: The final movie screenshots. (©Xuanyang Huang and Xiaoyun Zhong)

genre frameworks, LLMs can create multiple narrative dimensions. This capability enhances the storytelling potential of cinema and expands the future horizons of cinematic media.

In our future work, we aim to enhance AI-generated narratives by incorporating detailed character backgrounds and exploring non-linear storytelling. While LLMs currently produce structurally complete stories based on specified genres, it often defaults to traditional hero narratives. We will address this by diversifying narrative structures and deepening character development. Additionally, we will investigate the long-term impact of integrating AI and pets into narrative creation, including ethical implications of involving pets in creative processes. Rigorous attention will be paid to ethical considerations and data privacy concerns to ensure responsible technology deployment.

5 CONCLUSION

In *Cinema Meowdiso*, we envision a collaborative creation model involving pets, AI, and human guidance. This co-creation process expands the possibilities for creative cooperation between pets and AI, introduces new narrative experiences for automated storytelling, and grounds fictional narratives in real-world data.

The title of *Cinema Meowdiso* pays homage to *Cinema Paradiso* (1988) [Tornatore 1988], whose main theme is the extension of human emotions and memories through film as a medium. In our project, film extends pet experiences, creating narratives beyond their real life. This tradition of emotional storytelling is enhanced by AI and creativity, perpetuating film as a cultural medium.

Acknowledgments

We sincerely thank Huang Wei, Kedouh, Yeming Li, and Ziyi Zhang for their support. We also appreciate the technical support from the CMA thrust at HKUST(GZ). Additionally, we would like to thank our star cat, Whisker (Cengceng), for her performance.

References

Aristotle. 1996. *Poetics* (reissue ed.). Penguin Classics, London. Translated by Malcolm Heath.

Sofian Audry. 2021. *Art in the age of machine learning.* The MIT Press, Cambridge, MA.

Lucien Castaing-Taylor and Véréna Paravel. 2020. Leviathan. https://www2.bfi.org.uk/news-opinion/sight-sound-magazine/reviews-recommendations/film-week-leviathan. Accessed: June 27, 2024.

James Coupe. 2014. Jalousie Room. http://jamescoupe.com/?p=2037. Accessed: June 27, 2024.

Andrea DenHoed. 2018. The Turn-of-the-Century Pigeons That Photographed Earth from Above. https://www.newyorker.com/culture/photo-booth/the-turn-of-the-century-pigeons-that-photographed-earth-from-above. Accessed: June 27, 2024.

John Downer. 2017. Spy in the Wild. BBC Natural History Unit. Documentary TV series.

Dejan Grba. 2017. Avoid Setup: Insights and Implications of Generative Cinema. *Leonardo* 50, 4 (2017), 384–393. https://doi.org/10.1162/LEON_a_01456

Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599.

Lev Manovich. 2002. *The Language of New Media.* MIT Press, Cambridge, MA.

Jon McCormack, Camilo Cruz Gambardella, Nina Rajcic, Stephen James Krol, Maria Teresa Llano, and Meng Yang. 2023. Is Writing Prompts Really Making Art?. In *Artificial Intelligence in Music, Sound, Art and Design*, Colin Johnson, Nereida Rodríguez-Fernández, and Sérgio M. Rebelo (Eds.). Springer Nature Switzerland, Cham, 196–211.

Arthur I. Miller. 2019. *32 Ross Goodwin and the First AI-Scripted Movie.* MIT Press, Cambridge, MA, 225–230.

Hua Xuan Qin, Shan Jin, Ze Gao, Mingming Fan, and Pan Hui. 2024. CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation with LLM-Powered Chatbot Avatars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 1051, 19 pages. https://doi.org/10.1145/3613904.3642105

Ian Steenstra, Prasanth Murali, Rebecca B. Perkins, Natalie Joseph, Michael K Paasche-Orlow, and Timothy Bickmore. 2024. Engaging and Entertaining Adolescents in Health Education Using LLM-Generated Fantasy Narrative Games and Virtual Agents. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24).* Association for Computing Machinery, New York, NY, USA, Article 126, 8 pages. https://doi.org/10.1145/3613905.3650983

Jana Sterbak. 2024. Waiting for High Water. https://fonderiedarling.org/en/waiting-for-high-water. Accessed: June 27, 2024.

Giuseppe Tornatore. 1988. Cinema Paradiso. Film. Original release date: 1988-11-17.

John Truby. 2007. *The Anatomy of Story: 22 Steps to Becoming a Master Storyteller.* Faber & Faber, New York, NY, USA.

Jinbo Xing et al. 2023. Dynamicrafter: Animating Open-Domain Images with Video Diffusion Priors. *arXiv preprint arXiv:2310.12190* (2023).

Yubo Zhao and Xiying Bao. 2023. Narratron: Collaborative Writing and Shadow-playing of Children Stories with Large Language Models. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 119, 6 pages. https://doi.org/10.1145/3586182.3625120

Joanna Zylinska. 2017. *Nonhuman photography*. The MIT Press, Cambridge, MA.