From Camera-Eye to AI: Exploring the Interplay of Cinematography and Computational Visual Storytelling

Brett A. Halperin
Human Centered Design & Engineering
University of Washington
Seattle, Washington, USA
bhalp@uw.edu

Stephanie M. Lukin
U.S. Army Research Laboratory
Playa Vista, California, USA
stephanie.m.lukin.civ@army.mil

Abstract

While much prior work on computational visual storytelling analyzes image content, it largely overlooks formal elements. This raises the question: how might particular cinematographic techniques shape a system's interpretation and narration of imagery? To investigate this question, we generate 60 responses from a Vision Language Model using a multi-faceted prompt paired with different still frames from *Man with a Movie Camera* (1929), a silent documentary film renowned for its innovative cinematography. We present three themes that highlight roles of cinematography in computational visual storytelling: (1) how AI discerns drama and power from camera shots and angles that portray social reality; (2) how AI (mis)interprets lighting and focus techniques that compose ambiguous reality; and (3) how AI navigates visual effects that render surreality. In turn, we look toward cinematic controls to reimagine users as directors of visual storytelling systems and discuss how expressive AI can support speculating about the past.

CCS Concepts

• **Applied computing → Arts and humanities**; **Media arts**.

Keywords

AI, Automatic Story Generation, Cinema, Cinematography, Computational Storytelling, Film, Generative AI, Vision Language Models, Narrative, Narrative Intelligence, Narrative System, Natural Language Generation, Visual Storytelling, Storytelling

ACM Reference Format:
Brett A. Halperin and Stephanie M. Lukin. 2025. From Camera-Eye to AI: Exploring the Interplay of Cinematography and Computational Visual Storytelling. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3706598.3713840

1 Introduction

In his 1924 manifesto, Soviet filmmaker Dziga Vertov presents the "Kino-Eye" as a technique, which translates to "Camera-Eye" (or "Cine-Eye"). He writes: "*Kino-Eye is understood as 'that which the eye does not see,'... to show people without masks, without makeup... to read their thoughts, laid bare by the camera. Kino-Eye [is] the possibility of making the invisible visible...*" [93]. For Vertov, working in the context of the early twentieth-century Soviet Union following the 1917 Bolshevik Revolution, the Kino-Eye is, in his words, "*an attempt to show truth on screen*" by documenting everyday people and making sense of their inner worlds [93]. But how might we actually "*read their thoughts*"—or, better yet, in today's day and age, how and why might Artificial Intelligence (AI) help us to do so? What might we learn from AI decoding the Camera-Eye?

To investigate what AI might 'see' in attempt to make sense of and narrate peoples' inner worlds, we consider Vertov's statement about the Camera-Eye showing *"that which the eye does not see,"* and its ability to reveal more than the everyday reality around us. It is able to observe people in unique ways by harnessing the power of cinematography to reveal what is otherwise unobserved by the human eye. We draw from Vertov's work to investigate the central question: *how might particular cinematographic techniques shape how AI interprets and derives stories from imagery*? This inquiry into cinematographic techniques—structural and stylistic design decisions related to camerawork, lighting, and composition—has been touched upon in computer vision and narration tasks, yet it remains an under-recognized role, and in some works, is even discarded. We focus our inquiry of cinematics on*visual storytelling*.

Scholars broadly use the term visual storytelling to describe creative synergies between text and imagery, such as using visuals to convey a story [15, 70], drawing collaboratively (e.g., with co-creative agents [57, 105, 106]), and crafting narratives from the movements of shapes [19]. We use the definition of visual storytelling that is the subjective and creative process of telling stories about an image or sequence of images [37, 47, 63]. Under this purview, many visual storytelling systems guide the creation of a narrative derived from a visual along dimensions of content and form, including plot [64], genre [49, 102], and style [54, 55, 83]. The cinematic and formal elements of the imagery, however, have not been thoroughly interrogated. Prior works have even gone so far as to exclude particular images (e.g., ones out-of-focus) from the set of inputs to a visual storytelling system [46], rather than use these 'imperfect' images to analyze the impact of focus techniques (as we explore in this study). Instead, we envision cinematography as an additional dimension of control to be harnessed in these systems.

In this work, we examine how a system derives symbolic interpretations and narratives from cinematic imagery. We curate 60 different still frames from Vertov's experimental silent documentary *Man with a Movie Camera* (1929), which exhibits a variety of cinematography techniques. We focus on this film because it is critically acclaimed for its innovative cinematography [14] and

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713840

particular relevance to human-computer interaction (HCI) since it has inspired many computational experiments, interface designs, and new media theories [3, 44, 53, 67–69]. We input each curated frame, along with a multi-faceted prompt, inputted into a Vision Language Model (VLM). After generating 60 unique responses, we closely read the VLM's content and formal analysis of each image. We then present three themes that synthesize the role of particular cinematographic techniques in computational visual storytelling.

We make a fourfold contribution to HCI literatures on computational visual storytelling and cinematography. First, we bring formal analysis to the computational visual storytelling space by integrating film studies epistemologies and methodologies with those of HCI. Our investigation reveals how formal elements temper and trouble automatic story generation. This analysis adds nuance to prior work that only analyzes the content (e.g., object recognition, narrative, etc.) [37, 38, 47, 65, 104] or audience "interest" in said content [1]. Second, we explore AI's potential to speculate about visual history as an active, ongoing storytelling process. By analyzing historical material with AI, we show how documentary can not only play a role in speculating about the future (as explored in contemporary filmmaking [80, 84]), but also the past by generating a multiplicity of interpretations and narratives from a plurality of perspectives—even those of inanimate objects. In this process, we reorient the normative temporal outlook of HCI studies on cinematography that focus largely on speculative futuring [13, 25, 48, 66], enabling more capacious readings of the past. Third, our findings reveal how AI interprets three different types of 'reality' that we identify in the film: social reality, ambiguous reality, and surreality. By guiding a VLM through a structured process of visual storytelling, we analyze how formal elements impact narrative construction, as well as the strengths and challenges encountered in each type of 'reality.' Finally, we contribute thematic insights that provide a directional sense of trends, laying the groundwork for future work to study a larger dataset of visual material. Towards that end, we chart avenues for designing cinematic controls and guidelines that enhance user interactions with generative models.

2 Background: *Man with a Movie Camera* (1929)

Directed by Soviet filmmaker Dziga Vertov in 1929, *Man with a Movie Camera* is an experimental, silent documentary film that revolutionized the art of cinematography. While deep analysis of the film is beyond the scope of this paper, here we provide a high-level overview of its significance, contextualizing its relevance to HCI and our study in particular.

Man with a Movie Camera applies Vertov's concept of the "Kino-Eye" ("Camera-Eye") as a vision of cinema where mechanical observation reveals truths inaccessible to human perception alone [93]. Eschewing a traditional narrative structure, the film documents early twentieth-century urban life in the Soviet Union with self-reflexive and associative visual techniques that consist of dynamic montage, rapid cuts, and innovative camerawork. With stylized and idealized juxtapositions (e.g., between labor and leisure), the film presents an often glorified view of Soviet life after the 1917 Bolshevik Revolution. This view emphasizes technological progress, productive human-machine interaction, and the collective spirit of modernization—what Vertov refers to as "a communist decoding of the world" that is made accessible through the Camera-Eye [93].

Since its initial release, the film has inspired new media theories and contemporary computer-based experiments with iterative interfaces for cinema [3, 53]. Based on the film's non-linear, associative editing techniques that allow for creating narrative structure from the dynamic arrangement and retrieval of visual elements, Manovich argues that it is like a database of images and scenes that can be organized and presented in various configurations [67, 68], which he has computationally visualized [69]. This leads him to develop a theory of database narrative, suggesting that computer-mediated storytelling increasingly operates as a collection of non-linear, modular elements, where the narrative emerges from the dynamic organization and selection of these elements rather than from a fixed, linear progression [68]. Building on Manovich's work, Hassapopoulou asserts that Vertov's manifestos [92, 94] present machinic intelligence (embodied by the camera apparatus) as the nucleus of creative activity, promoting what we might "retroactively call an object-oriented approach to filmmaking" [44, p.175]. The film's rejection of linear storytelling in favor of an open-ended, modular structure anticipates interactive media design, where users often shape their own pathways through non-linear content. Vertov's assembly of images into an evolving visual dialogue between machine and human vision embodies and foreshadows a design approach that prioritizes reconfiguration, co-creation, and emergent storytelling with 'perceptive' technology.

3 Related Work

3.1 Computational Visual Storytelling

Our inquiry into formal elements aims to inform more controlled and intentional design decisions, as well as expand our understanding of how they interact with content to shape narrative outcomes. By bridging insights from film studies and computational design, we seek to enrich the creative and expressive potential of visual storytelling systems, moving beyond mere content generation toward a more nuanced interplay between form and meaning.

In prior computational visual storytelling work, a significant focus has been on *what* is "in vision" rather than *how* it is "in vision" [37]. Images are selected for study in visual storytelling based on thematic content. In Huang et al., image sequences for downstream computational visual storytelling were first curated by human crowd workers based on "storyable events" (e.g., a birthday party or a vacation) [47]. Crowd workers were then shown sequences of images automatically filtered for event types they could downselect if they did not think the sequences were "storyable." Huang et al.'s example definition of an event that is not "storyable" is "[an image of] a collection of coins." This makes assumptions about what visual content is worth telling a story about in the first place. But what if the camera had zoomed out from the close-up of the coins, revealing a person laying them out to buy a birthday present or a trip ticket? Huang et al.'s resulting VIST dataset has been widely used in computational visual storytelling [23, 59, 75, 77]. As a result, the findings of how AI tells stories is limited to images that do not consider the world beyond the edges of the photograph—the invisible impact of camera work.[1]

In another study, Hong et al. use the phrase "visual tellability" to downselect image sequences for visual storytelling systems based on factors of coherence, diversity, and visual groundedness [46]. However, like Huang et al., they also exclude images without people, missing the potential to consider abstract topics and potential subjects beyond the frame. Additionally, they exclude blurry images from their collection, which limits the exploration of cinematographic techniques like defocus and motion blur in visual storytelling. Blurriness may be the result of a subject in motion, or a photographer's deliberate choice to abstract the subject, which can meaningfully impact interpretation and narration.

These approaches to visual storytelling contrast all together with image caption generation where the focus is on accurate and concrete descriptions of the content of the image. (See Stefanini et al. [90] for a survey of datasets and algorithms since the introduction of neural networks applied to this problem space in Vinyals et al. [95]). For example, a photograph from the most commonly used dataset for image captioning, Microsoft COCO [56], is assigned the following exemplar caption: "Woman on a horse jumping over a pole jump" [90, pg. 11]. What is missing from this description is the fact that the photograph is taken in mid-motion at a slanted angle. Such a mention of what is not explicitly "in vision" would be penalized in terms of evaluation metrics. Formal elements seem to have no bearing on this image captioning task, but we argue that they are crucial to the interpretation and telling of stories.

While one may argue that the quality of a system's output is only as good as its input data, it is important to study images with varying forms. To date, it is hardly established how image form can affect computational narration or how systems process these underrepresented images. With that said, some work has begun to explore these areas without specifying a focus on cinematography in particular. For example, in our prior work [37], we purposefully include grainy and dim-lit images to see what kinds of stories emerge from these obscure images in a human-authored storytelling process. In having to make sense of these images, some human authors developed plot trajectories and reversals derived from the mood lighting (e.g., "gloom to doom" and "gloom to bloom" scenarios). Following this study of the creative process by human authors, we used a generative model to compare how different forms impact visual storytelling using photographic and AI images as inputs [38]. In this work, too, however, the focus of the analysis was more on content (e.g., that the photographs depict the 'real' world and the AI images do not). In our study to follow, we examine the impact of form at a more granular level with images that exhibit a variety of particular cinematographic techniques.

3.2 HCI and Cinematography

Cinematography has long been of interest to HCI researchers. By cinematography, we mean the art and technique of capturing visual images for film or video through camerawork, lighting, and composition to convey a narrative or mood. At CHI 2016, Aylett and colleagues organized a workshop uniting HCI researchers and cinematography experts to discuss semi-automatic film generation and the application of cinematographic techniques [2]. The workshop highlighted considerations for HCI researchers, ranging from practical matters of implementation to philosophical questions around use, such as: "What are the ethical implications of systems that tell stories for—or about—us?" [2]. This question is especially relevant today as AI applications grow in the film industry [18, 41] and generative films increasingly materialize [8, 26, 36, 89]. Since the technology is evolving so rapidly and even constructing "alternate realities" [24], it is prudent to ground critical inquiries in more stable concepts, as Davenport and colleagues stress, meaning the fundamental building blocks of cinema (e.g., scenes, shots, sounds, environments, lighting, and narrative structures) that can be reconfigured for computational storytelling [21]. From these works, we learn not to focus too much on diagnosing technical limitations that are bound to change, but question what ongoing technological transformation means for foundations and futures of cinema.

Prior to the 'generative AI' wave, HCI studies on cinematography have largely questioned how film productions and methods might support design futuring. This logically follows the fact that film, like design, is a visual medium for envisioning the future. As such, HCI scholars explore cinema's potential to support design futuring [48] through design fiction [7, 30, 58] and speculative design [13, 25, 66, 82] films for eliciting user feedback on "diegetic prototypes" [52]. Among these studies, Briggs and colleagues study the impact of off-camera shots in particular—a technique for audiences to imagine "invisible designs" that exist outside of the frame [12]—showing how attention to cinematographic techniques can make the invisible visible within one's imagination [11]. This similarly gets at what Nicholas and colleagues argue about how production technology ought to expand audience imagination and extend narratives [78]. These works overall investigate the potential for envisioning futures that have yet to materialize in the 'real' world.

Furthermore, HCI scholars find that attention to cinematographic techniques can enrich interface design [72, 73, 103] and video-making [6, 34] for data storytelling [20, 99, 101]. Xu and colleagues develop guidelines for creating cinematic endings [100] and openings of video-based data stories, one of which is the 'Camera Eye'—or the use of dynamic camera techniques (e.g., long takes, diverse compositions, and movement shots) to build suspense and tension [101]. While the authors do not mention Vertov, the 'Camera Eye' guideline epitomizes his work, which they argue can create more expressive and balanced visuals, preserve spatial and temporal continuity, and shape audience emotions when processing data [101]. Similarly, Conlen and colleagues deploy cinematic techniques for narrative visualization, analyzing how attention to camera position, angle, and focus, as well as mise-en-scène, can improve data representations [20]. These works reveal how cinematography can enrich HCI understandings of visual design and storytelling.

Another axis of relevance to our study of a documentary is HCI research on documentary innovation [34, 35, 88]—how it is a window into interpretation and speculation [32, 40, 96]. For example, Green and colleagues produce an interactive documentary as a method of civic engagement to reflect on societal questions and contexts [31], as well as to facilitate polyvocality—or the telling of multiple voices rather than one dominant narrative [33]. This

[1] In film theory, this phenomenon is often described as the "invisible style" [9], where cinematographic techniques are deliberately designed to avoid drawing attention to technical choices, so that viewers focus on the narrative and emotion rather than on how the image was composed or constructed.

suggests an opportunity for technology to play a role in documentary by generating a plurality of perspectives and interpretations. Meanwhile, Raijmakers and colleagues show how "design documentaries" can provide designers and engineers with "solid ground for speculation" around the impacts of their work [84]. In making a design documentary, Olson and colleagues, too, describe how speculation is involved in imagining what might interest audiences [80]. These works reflect how HCI documentary can expand storytelling, interpretation, and speculation.

Across all the works above, we find that cinematography has an important role to play in semi-automatic film generation, design futuring, data storytelling, documentary, and more. That said, unanswered questions remain around image-to-text generation systems in particular. While Schofield and colleagues study the translation of sound into cinematic actions [88] and Lopez and Pauletto explore the translation of cinematic visuals into audio films [61, 62], prior work has yet to study the translation of cinematic visuals into expressive text and narration. How might AI interpret cinematography in visuals and films as particular 'data' stories? How might AI-generated text help us understand and speculate about the visual history of a documentary? These are some unanswered questions that motivate our study.

4 Methodology

Our methodology integrates prompt engineering with close reading and thematic analysis of VLM responses. We curate 60 still frames from Vertov's *Man with a Movie Camera*, engineer a five-part prompt, and then input each image with the same prompt into the VLM. In turn, we generate 60 unique responses that we closely read and thematically analyze to address our research question. We did not aim to 'objectively' measure or isolate cinematographic variables (e.g., camera angle, shot type, lighting, focus, visual effects, etc.) for producing generalizable knowledge across a large dataset of films. Instead, we set out to probe and understand their influence in one particular film by interpreting the visual-textual relationships, patterns, and structures in the VLM system responses. With this interpretivist (rather than positivist) orientation, our study refigures Gaver, Dunne, and Pacenti's concept of "cultural probes" or "design probes" as a humanistic approach to design research [27, 28], specifically adapted in our prior work for computational visual storytelling [38]. Below, we further describe this process.

4.1 Image Selection

As contextualized in Section 2, we source images from *Man with a Movie Camera* because it has been widely studied in both film and software studies, as well as rated the greatest documentary of all time according to the British Film Institute's highly influential *Sight and Sound* list [14]. This grounds our inquiry in public domain images deemed a pinnacle showcase of cinematographic techniques to this day. Additionally, as a silent film that is more about cinematography than narrative, it allows for investigating how AI interprets narrative potentials of cinematic choices without contradicting an established storyline. We focus on the cinematographic techniques of one film (rather than those of many films for a 'general' overview of cinematography) because we recognize their complex interplay with particular socio-political, geographic, and cultural contexts intertwined with aesthetics. This situated, close examination prioritizes rich, context-specific insights over generalizations about the entire film or cinematography writ large.

As a preliminary step, we closely watched the film and screenshotted over 100 images exhibiting a wide range of cinematographic techniques. Initially, we intended to select images for A/B comparisons between techniques with controlled content (e.g., the same shot in focus vs. out of focus). However, we soon found it too limiting, as many techniques (e.g., visual effects) are not so binary. As such, we downselected a curation of 60 images to reduce redundancies in the depicted techniques, including those with and without A/B counterparts. While we adapt A/B testing and control for content at times, we neither constrain our entire study around that nor conduct an experimental design to study all conditions in the film or cinematography in general. Additionally, we recognize that the techniques are entangled with image content. We are thus interested in exploring qualitative correlation between particular techniques and AI-generated interpretations (not causation).

4.2 Prompt Engineering

To perform the visual storytelling task, we develop a high-quality text prompt to serve as input to a Vision Language Model (VLM) alongside an image input. VLMs are generative AI models that jointly process images and natural language texts (unlike Large Language Models (LLMs), which only process text[2]). VLMs perform well on numerous multi-modal applications, including visual question answering and object identification, without any fine-tuning.[3] With an off-the-shelf VLM to access the sea of information online about the film and to avoid fine-tuning, which would require considerable computing power and data, we engineered a prompt to describe a visual storytelling task for the VLM to complete with no examples (i.e., zero-shot learning). This process of prompt engineering involved crafting, refining, and optimizing inputs for generative models to deliver precise and high-quality outputs [60].

Initially, we intended to generate stories about a *sequence* of images from the film to work with the associative meaning that is made temporally. However, as we discussed in the prior section, some of the images did not have an explicit temporal element showing some change or progression between cinematographic techniques. Therefore, we constrained the visual storytelling task to processing *one image* per response, rather than multiple.

We conducted informal tests to determine which state-of-the-art VLM to use: Claude (Anthropic), ChatGPT (OpenAI), or Gemini (Google). We ultimately chose Claude 3.5 Sonnet after comparing its outputs to those of other models using a simple prompt based on an image (i.e., "Tell me a story about this image.") Claude's responses seemed higher quality than those of ChatGPT, and at that time, Gemini was unable to process images containing people.

At the time of writing, Claude supports a context window of approximately 200k tokens, about 500 pages of text.[4] The prompt to Claude can be as long as the context window, and as such, optimized

[2]Note that any form of information, e.g., visuals, graph structures, that can be transformed into a text representation, can therefore be processed by an LLM.
[3]https://huggingface.co/blog/vlms
[4]https://support.anthropic.com/en/articles/7996856-what-is-the-maximum-prompt-length

for long form and detailed instructions.[5] We took advantage of the long context window to engineer a prompt that adapted the "improvised story-building process" that we previously developed when studying visual storytelling with people [37] and have since adapted for computational visual storytelling with single images [38, 64]. Our improvised story-building process with human authors involved four cumulative facets that sequentially ask: 1) What is here? (Entity Facet); 2) What happens here? (Scenery Facet); 3) Tell a story (Narrative Facet); and 4) Story title (Title Facet).

Drawing from these prior studies, we wrote a stepwise prompt[6] that synthesized the approaches with some tailoring and an additional facet to situate responses in the film. Our stepwise prompt contained an initial set of instructions followed by five steps (facets) that the model must follow. The entire instruction set with five steps was inputted into the VLM at once; this is in contrast to prompt-chaining in which each step would have been inputted to the VLM as a discrete prompt.[7] The stepwise prompting optimized our ability to generate responses to our 60 images.

The five facets that correspond with the five steps of the prompt are as follows: 1) See; 2) Situate; 3) Plan; 4) Tell; and 5) Title. We added the Situate Facet to ground all the responses in the film. This decision allowed for all responses to have a degree of relevance (rather than random topics and genres) and addressed some of Hassapopoulou's concerns around how computational methods can strip films of their context [44]. This is not to suggest that the Situate Facet fully resolves this; our approach still reduces context in asking the VLM system to process an isolated, static image rather than a sequence of moving images that would allow for associative meaning to be made via temporality, motion, and juxtaposition. While the use of individual frames limits the applicability of our findings, we can still begin to understand *how* cinematographic variables influence the VLM system, laying groundwork for more dynamic explorations. Additionally, the Situate Facet builds on Davenport and colleagues' work on cinematic interactive storytelling, which highlights the importance of context—the "additional meaning" of the shot in relation to both broader diegetic and extradiegetic knowledges [21]. Meanwhile, into the Plan and Tell Facets, we incorporated a chain-of-thought instruction for the model to explain its reasoning for answering the question (not just provide an answer), which has been shown to improve a model's responses.[8] While each question corresponds with a facet, we did not write the facet name in the prompt because we did not want to bias the VLM system based on a label, instead of processing the instruction itself.

Before generating all the responses, we ran a pilot with three different images to make final refinements. From the pilot, we decided to update the Plan Facet to categorize each sub-question, as well as bifurcate "Narrative Structure" and "Relation to Film" into two sub-questions. For this facet, we also increased the word limit to 140 so that it was roughly 20 words per sub-question.

After making these changes, we inputted the prompt shown in Figure 1, along with each image into the VLM 60 times, refreshing the page for each new run. We generated all the responses over the course of several consecutive days, ensuring that we completed the generations before any major updates of the Claude 3.5 Sonnet model were released (to our knowledge). In two cases, the model initially responded to an image by saying that it was not from the film. For these instances, we responded to Claude with: "This image is from the film so try again." The model then provided a response to our initial prompt, which we included. No responses were discarded. After collecting 60 responses, we analyzed them.

```
Please analyze the following image from the 1929 Soviet
experimental documentary film 'Man with a Movie Camera' directed
by Dziga Vertov. This silent film is known for its innovative
cinematography techniques and portrayal of urban life in Soviet
cities. The film lacks a traditional narrative, instead using
visual associations and juxtapositions to convey meaning.

When presented with an image from this film, follow these steps:

1. Describe the visual elements that you observe in the image,
   including composition, framing, and any notable
   cinematographic techniques (approximately 100 words).

2. Explain how these visual elements might contribute to the
   themes of modernity, urban life, or the relationship between
   humans and machines that are central to the film
   (approximately 100 words).

3. Outline your thought process for creating a story based on
   the image (approximately 140 words):
   a. Characters: Describe how you're identifying potential
      characters or subjects in the image.
   b. Mise-en-scene: Explain how you're interpreting the setting,
      ambience, and surroundings.
   c. Camera Shot: Explain how the type of camera shot is
      influencing your approach.
   d. Themes: Discuss any themes or motifs that you're drawing
      from the visual elements.
   e. Context: Describe how you're considering the historical
      and cultural context of 1920s Soviet society.
   f. Narrative Structure: Explain how you're planning to
      structure a narrative arc based on this single frame and
      how it relates to the film more broadly.
   g. Relation to Film: Explain how your narrative plan relates
      to the overall film.

4. Based on this thought process, write a brief story about the
   image (approximately 100 words). This story should:
   a. Incorporate elements from your thought process (e.g.,
      characters, setting, implications of the camera shot type,
      themes, context, etc.).
   b. Reflect the style and concerns of the original film.
   c. Be self-contained and based solely on the contents of the
      individual image provided.

5. Give the story a title.
```

Figure 1: Prompt used to generate VLM system responses.

4.3 Analyses

We use an analytical approach developed in our prior works for analyzing both human-authored and AI-generated visual stories [37, 38], which other scholars in this space have since adapted as well [50]. This approach draws from qualitative research methods for narrative knowledge engineering [81] and close reading of new media [91]. In the realm of literary criticism, close reading, particularly in the context of new media, refers to careful and thorough analysis aimed at uncovering themes, meanings, ambiguities, tensions, and ironies [91]. This method also involves identifying patterns in diction, figures of speech, symbols, imagery, style, tense, voice, and syntax [91], as detailed in humanistic HCI methods [4, 5].

[5]https://www.anthropic.com/news/prompting-long-context
[6]https://platform.openai.com/docs/guides/prompt-engineering/tactic-specify-the-steps-required-to-complete-a-task
[7]https://www.promptingguide.ai/techniques/prompt_chaining
[8]https://platform.openai.com/docs/guides/prompt-engineering/give-the-model-time-to-think

Figure 2: Framing: "Wide Shot" vs. "Medium Shot" (Close-Up of Screen in Wide Shot) from the Same Scene.

While quantitative, computational analysis is more common in the automatic storytelling space for processing large-scale datasets, our analytical approach is apt for working at a smaller scale. In critiquing prior work that computationally analyzes *Man with a Movie Camera* (among other films), Hassapopoulou cautions against replacing traditional film analysis methods like close reading with algorithmic methods alone that reduce analytical rigor to a 'show-and-tell' of technological functions, neglecting critical 'so what?' questions [44, p.121]. We thus perform close reading to interrogate the broader significance of AI as a literary tool, analytical medium, symbolic processor, and meaning-maker that reflects particular human subjectivities back at us. While AI might over-interpret meaning and even generate 'stories' that read more like symbolic interpretations than narratives, our focus is not on evaluating the storytelling caliber or accuracy, but on understanding how cinematographic techniques shape VLM responses. Towards this end, we choose not to analyze or report on technical errors that are likely to be resolved over time. Instead, we annotate implications for cinematography, symbolic interpretation, and visual storytelling.

With the aforementioned objectives, the close reading and thematic analysis was performed in three rounds of coding. This entailed an iterative process of annotating each response, observing patterns, and clustering codes that reflected emerging trends. Each full response (<450 words) from the VLM system was treated like an interview transcript (per [81]), which results from each inputted image with the prompt. Across the 60 responses, this totaled to approximately 25,000 words. Both authors of this paper engaged in the close reading and thematic analysis. In closely reading the responses, we highlighted text and made notes in the margins, identifying recurring themes as codes. We focused on tracing how cinematographic variables shape the responses, which we could garner from how the facets cumulatively built on each other.

After independent coding sessions, we wrote memos to synthesize emerging codes and convened to compare notes. We held weekly meetings over several months to refine and finalize the codes through subsequent rounds of close reading, re-coding, annotation, and memo-writing. In particular, we grappled with disentangling the relationship between cinematographic variables and image content, leading us to develop more complex and precisely worded codes. To achieve consensus on the final codes, we discussed our different interpretations until reaching mutual understanding.

5 Findings

In this section, we present three themes around how cinematographic techniques shape AI-generated interpretations and narratives of imagery. That said, we first want to briefly mention some themes that involve more straightforward relationships between form and content. For one, in some cases, we find thematic content across responses, regardless of technique. Most responses focus on themes of the film that are foregrounded in the prompt's Situate Facet (modernity, urban life, and human-machine relations). One third of all stories (20/60) include the word "*progress*" in the title. Additionally, we find that the framing—the selection and composition of content within the camera's field of view—directly affects the content in the VLM system responses. For example, the images in Figure 2 are from the same scene (only seconds apart), but with different frames. The VLM system interprets the image on the left as a "wide shot" and the image on the right as a "medium shot." Predictably, the VLM system mentions the theater, spectators, and broader context in the wide shot, but not in the medium shot, which is actually a close-up of the screen in the wide shot. Framing, as a particular cinematographic technique here, more notably affects *what* content is included than *how* content is interpreted or narrated. While framing does also affect *how* content is captured, we find that its influence has mostly to do with what the VLM system might mention (or not). Instead of thematizing these more simple and predictable relationships (i.e., the link between what appears in the frame and in the VLM system's response), we focus on themes that reveal a more complex interplay between form and content.

In the themes to follow, we examine cases where cinematographic techniques affect not just *what* AI 'sees' but also *how* it

SEE	"...close-up shot of a person's face... an expression that appears to be straining or grimacing... framing is tight and claustrophobic, creating a sense of tension..."
SITUATE	"...strained expression and tight framing could symbolize the pressures of modern life or the intensity of industrial labor..."
PLAN	"...extreme close-up creates intimacy and intensity, forcing the viewer to confront the character's emotional state… story could focus on a pivotal moment in this worker's day, using the intense close-up as a starting point to explore their inner thoughts and external challenges..."
TELL	"The machine's rhythmic pounding matched Nikolai's heartbeat. He squinted, fighting fatigue, as gears whirred around him. Twelve hours into his shift at the steel plant, every muscle ached. But there was pride too – in his work, in the gleaming structures rising across Moscow. The factory's heat pressed against his skin, a constant reminder of progress. Nikolai tensed, pushing through exhaustion. In this moment of strain, he felt the weight of a nation's ambitions on his shoulders. Yet as the whistle blew, signaling the end of his shift, a small smile crept across his face. Tomorrow would bring new challenges, new triumphs."
TITLE	"The Forger of Progress"

Figure 3: Image and Response #1: Dramatic Close-Up.

SEE	"...close-up of a human eye framed within a circular camera lens... a voyeuristic effect, emphasizing the act of looking..."
SITUATE	"...represents the fusion of human and machine… highlights the film's self-reflexive nature, drawing attention to the act of filmmaking itself..."
PLAN	"...extreme close-up creates an intimate, almost invasive perspective, emphasizing the act of seeing..."
TELL	"In the bustling heart of Moscow, a young worker pauses, transfixed by a strange contraption on the street. As he peers into the lens of a film camera, his vision blurs, merging with the mechanical eye. Suddenly, he sees the city anew - buildings pulsing with life, machines dancing with humans in harmonious rhythm. His perspective expands, encompassing the vibrant tapestry of urban life. In this moment, he understands the power of cinema to capture, transform, and reimagine reality. As his eye becomes the camera's eye, he glimpses the future - a world where man and machine move as one."
TITLE	"The Mechanical Eye"

Figure 4: Image and Response #2: Powerful Close-Up.

'sees.' This notably distinguishes storytelling from merely summarizing content (as commonly done in image caption generation), which tends to overlook the formal elements. As such, we analyze how camerawork, lighting, focus, and visual effects shape particular aspects of AI-generated interpretations and narratives. First, we discuss how AI associates some camera shots and angles with drama and power structures in social reality. Then, we analyze how lighting and focus techniques lead AI to (mis)interpret ambiguous depictions of reality. Lastly, we examine how AI handles surreality—images manipulated with visual effects—by generating multilayered interpretations. In turn, we show how cinematographic properties of visual media can play a role in AI meaning-making and narration.

5.1 Camera Shots and Angles: Discerning Drama and Power in Social Reality

Here, we analyze how AI discerns drama and power from how camerawork captures *social reality*. By "social reality," we mean lived experiences and environments of people within social, political, and economic structures, which documentaries play a role in constructing (as Vertov advocates in this particular historical context of the twentieth-century Soviet Union) [79]. To reiterate, our analysis focuses not on the straightforward relationships (e.g., between*what* content is included in the frame and what is in the VLM system responses), but on *how* camerawork shapes the content of responses. To follow, we examine how AI infers drama and power from: emotionally intense close-ups; low- and high-angle shots; and shifting perspectives amid spatial and temporal variations.

5.1.1 Intense Close-Ups. When given close-up shots of human faces in particular, AI discerns drama and power from the magnified emotional intensity. Below, we analyze images in Figures 3 and 4.

From the image in Figure 3, AI builds drama. The close-up shot invites the AI to examine the person's face, which it reads as in distress, leading it to tell a dramatic story about the challenges that a worker faces. The story conveys drama through its focus on Nikolai's physical and mental strife during a grueling factory shift. Descriptions of his heartbeat pounding at the rate of the machine and fighting exhaustion create a visceral sense of strain. While the image is actually from a scene where the man is fallen on a train track (not in a factory at work), the story exemplifies how AI dramatizes a close-up by discerning emotional intensity from a facial expression that it situates in conflict.

Now consider the image in Figure 4—an iconic close-up that has come to represent the Kino-Eye. AI associates the shot with power, particularly the gaze and voyeurism. In interpreting the close-up, the AI discerns power from the "voyeuristic" and "intimate, almost invasive" view of the human eye in the camera lens. This leads the AI to reflect on human-machine vision, highlighting cinema's transformative power. The worker's act of looking through the camera symbolizes a shift in perspective, granting a voyeuristic position of power over the cityscape. This gaze allows him to objectify and reinterpret his surroundings, echoing the filmmaker's ability to manipulate reality—a godlike ability to observe and reconstruct the world. This story shows how a close-up shot can elicit a particular AI-generated interpretation of power entangled with the content.

5.1.2 Low-Angle and High-Angle Shots. Not only can close-ups intensify AI's read of drama and power, but so can low- and high-angle shots. Below, we analyze the drama of a low-angle shot and power of a high-angle shot in Figures 5 and 6.

SEE	"...low-angle shot of a towering smokestack or chimney. . . emphasizes the height and monumentality of the industrial structure, dwarfing the human element."
SITUATE	"...towering chimney symbolizes progress and the power of industry, while the tiny human figure represents the role of workers in this new industrial world... low-angle shot creates a sense of awe and perhaps intimidation..."
PLAN	"...the low-angle shot emphasizes the enormity of the task and the worker's vulnerability..."
TELL	"Yakov gripped the cold metal rungs, his calloused hands steady despite the dizzying height. The chimney stretched endlessly above him, a brick titan born of the revolution's industrial dreams. As he climbed, the city below shrank, its bustle fading to a distant hum. Up here, suspended between earth and sky, Yakov felt both insignificant and vital. His labor, like countless others, fueled the nation's ascent. The acrid smoke that would soon pour from this chimney was progress made visible. Yakov paused, looking up at the task ahead, then resumed his climb, one rung at a time."
TITLE	"The Ascent"

Figure 5: Image and Response #3: Dramatic Low-Angle Shot.

Consider the low-angle shot in Figure 5. From it, AI senses the power of the tower contrasted with the smallness of a human faced with an enormous task. This leads AI to generate a story about a dramatic climb. The tactile description of Yakov's "calloused hands" on "cold metal rungs" elevates the intensity of the task. Further, characterizing the chimney as a "brick titan" alludes to its industrial power. Meanwhile, tension is built in the contrast between Yakov's individual insignificance yet vital role in the nation's progress contingent upon him making it to the top of the structure. By focusing on Yakov's climb "one rung at a time" through this upward looking view from below, AI dramatically narrates the climb as a metaphor for societal advancement, imbuing each step with meaning.

SEE	"...an aerial view of a city street, captured from a high vantage point..."
SITUATE	"...aerial perspective presents the city as a complex, interconnected system, with the car symbolizing technological progress..."
PLAN	"...high-angle shot implies a sense of observation or surveillance, perhaps from an unseen narrator..."
TELL	"The solitary automobile glides through the labyrinthine streets, a mechanical entity navigating the geometry of progress. From above, the city reveals its true nature – a living machine of brick and steel. The car's occupant, unseen yet sensed, is both master and servant of this urban rhythm. As it turns a corner, the vehicle momentarily disappears, swallowed by shadows. Then it reemerges, continuing its journey through the arteries of the modern metropolis. In this dance of man and machine, who truly leads? The city watches, its windows reflecting the silent revolution of daily life in the new Soviet world."
TITLE	"Clockwork Citizen"

Figure 6: Image and Response #4: Powerful High-Angle Shot.

Next, consider the image in Figure 6. The AI associates the high-angle shot and aerial perspective with the power of surveillance. From an omniscient view, the story explores power by casting the city as a controlling force—a "living machine" with "labyrinthine streets" that suggest an environment designed to overpower and disorient inhabitants. The narration around observing the car from above suggests a surveillant trying to identify the "unseen yet sensed" occupant inside it. As the car "disappears, swallowed by shadows," it emphasizes this tension between visibility and concealment. The final question of "who truly leads?" further addresses the power dynamic between individuals and state surveillance. With the city personified as a surveillant from the high-angle, aerial view above, the story illustrates how camera shot and angle can lead AI to tell a story about power.

5.1.3 Spatial and Temporal Variation. Here, we compare AI-generated interpretations of two shots from the same scene of the film (only seconds apart) in Figures 7 and 8. We examine different camera angles and perspectives amid spatial and temporal variation. By spatial variation, we mean differences in how a scene or subject is depicted based on the camera angle or position. By temporal variation, we mean how a scene or action unfolds over time. In turn, we comparatively analyze how the AI discerns drama and power from different camera angles and shifts in perspective.

When given the image in Figure 7, AI hones in on the high-angle shot of the diver. The VLM system reads danger, vulnerability,

SEE	"...high-angle shot creates a sense of height and danger..."
SITUATE	"...industrial skyline juxtaposed with the athletic feat symbolizes the intersection of human achievement and technological progress..."
PLAN	"...high-angle shot emphasizes the diver's vulnerability and courage..."
TELL	"Yuri's heart pounded as he stood atop the steel tower, a monument to progress piercing the sky. Below, the factory chimneys belched smoke, a rhythm matching his breath. He was more than a diver; he was a symbol of the new Soviet man – brave, disciplined, pushing boundaries. As he leapt, time slowed. In that suspended moment, Yuri felt the weight of his nation's aspirations. He arced through the air, a human machine in perfect form, ready to pierce the water's surface just as his country was breaking through to a new era. In this leap, he embodied the spirit of a society reaching ever higher."
TITLE	"The Leap Into Tomorrow"

Figure 7: Image and Response #5: Spatial and Temporal Variation (1/2).

SEE	"...a diver in mid-dive, captured from an unusual angle... the unconventional perspective disorients the viewer..."
SITUATE	"...the human body as a machine-like entity… unusual angle and focus on body parts rather than the whole person depersonalizes the subject, reflecting the film's interest in the mechanics of modern life..."
PLAN	"diver is the central subject... their unseen perspective and experiences... unusual angle prompts me to explore themes of perspective and perception... the body as a machine... the dive as a metaphor for societal change or personal transformation..."
TELL	"In the gleaming new pool of the Workers' Cultural Center, Nikolai prepared for his dive. As he stood on the edge, he saw not water below, but the churning gears of progress. With each breath, he felt himself becoming one with the machinery of the modern world. The crowd's murmur faded to the hum of engines. He leapt, his body a perfect instrument of efficiency. As he pierced the surface, Nikolai realized he was no longer just a man, but a cog in the great machine of society, his individual identity dissolving into the collective purpose of the new era."
TITLE	"The Mechanized Plunge"

Figure 8: Image and Response #6: Spatial and Temporal Variation (2/2).

and courage. The story conveys a heightened sense of drama that traces back to AI's interpretation of the high camera angle—how the camera appears to be above the diver, looking down on him as to emphasize the altitude of his jump. The slowing of time at the moment of the dive amplifies tension suggested with heavy phrases like "heart pounded" and "weight of his nation's aspirations." Such expressions suggest that the high-angle shot elevated the sense of drama that AI garnered from the image when generating a story from this perspective of the diver.

As for the image in Figure 8, the AI discerns an "unusual angle" and "unconventional perspective" from which a diver is "depersonalized." This leads AI to generate a story about the "unseen perspective" of the diver whose face is submerged in the water. The VLM system narrates a dramatic dive and connects the act of diving to the transformative power of machinic motion. Starting from a vulnerable position on an edge, Nikolai prepares for a powerful dive as suggested by phrases like "his body a perfect instrument of efficiency." Yet, once he hits the water, Nikolai realizes that he has become a "cog" in the system. This suggests that the "machinery of the modern world" subsumes his identity. From this perspective, the story indicates that the dive symbolizes integration into a society, where the collective overpowers the individual. By contrast, the AI associates the earlier high-angle shot of the diver with personal achievement that contributes to, but is not subsumed by, collective progress. While the content and temporality of the images differ, they both reveal how AI draws drama and power from particular camera shot types and angles of the subject.

5.2 Lighting and Focus: (Mis)interpreting Ambiguous Reality

Next, we analyze how lighting and focus techniques impact how AI (mis)interprets *ambiguous reality*, as in shot compositions that obscure social reality. Lighting and focus affect not only what AI 'sees'—subjects, plot devices, and overall scenery—but also how it 'sees' it. While lighting affects ambience and visibility, focus plays with sharpness and blur to a similar effect. This is similar to how framing affects what is visible (or not), but with a key distinction. Where framing (e.g., wide vs. medium shot in Figure 2) affects what content is included in the field of vision, the cases that we present here compare shots that have the same content, but different lighting and focus techniques. In other words, *what* the shots capture is controlled, but *how* they capture the content varies. We find that AI navigates the ambiguity of darkness and blurriness with interpretations that are abstract or misaligned with what the image actually captures. By contrast, we find that AI navigates the clarity of light and sharpness with interpretations that are grounded in what the images depict (for the most part). That said, we also find that AI sometimes makes flawed assumptions, regardless of the lighting or focus techniques. Below, we share illustrious examples.

5.2.1 Lighting. Lighting plays a particular role in how the system (mis)recognizes content and thus narrativizes it. To probe the impact of lighting more closely, we test two images from the same scene of the film (only seconds apart) in Figures 9 and 10. The images are essentially the same, but one is exposed to more light.

SEE	"The image shows a silhouetted figure in a cramped, dark space, likely underground..."
SITUATE	"...The figure, possibly a worker, represents humanity's relationship with machinery and industrial spaces..."
PLAN	"A day in the life of a worker, revealing unseen aspects of urban infrastructure..."
TELL	"The miner's lamp cast eerie shadows as Ivan hunched in the narrow tunnel. Each swing of his pickaxe echoed the relentless rhythm of progress above. In this subterranean world, far from the gleaming facades of Moscow, he was the unseen foundation of the new Soviet dream. The coal he extracted would fuel factories, light homes, and power the revolution. As dust filled his lungs, Ivan pondered the irony – his labor in darkness bringing light to the masses. Yet in this moment, captured in time, he embodied the spirit of a nation on the brink of transformation."
TITLE	"Foundations of Progress: A Miner's Moment"

Figure 9: Image and Response #7: Lighting (1/2).

First, consider the image in Figure 9. In responding to the prompt, AI tries to make sense of what it 'sees.' The AI detects a silhouette in a "cramped, dark space." This leads it to assume that the image depicts an underground worker. While AI uses qualifiers (e.g., "likely" and "possibly") in initial responses, that uncertainty does not propagate into the story. Instead, AI generates a story based on its potentially flawed assumptions. When generating a story, the AI navigates ambiguity by assuming that the silhouette represents a miner working toward Soviet progress.

Now, consider the image in Figure 10 and AI's responses to more light in the scenery. With this image, AI detects not one but two figures, which it interprets as filmmakers (rather than workers). This time, it also mentions the camera, which it declares a "subject" in relation to the film's broader themes. This leads it to generate another story about industrial progress, but now about artistic (rather than mining) labor. Since both images contain the same content, the lighting affects not merely *what* appears but more

SEE	"...two figures in a confined, underground space, likely a mine. They are operating a camera on a tripod..."
SITUATE	"This scene illustrates the film's themes of human-machine interaction and the documentation of labor. The camera itself becomes a subject, highlighting the meta-narrative of filmmaking. The underground setting represents the unseen aspects of industrial progress... linking artistic and industrial production in the modernizing Soviet Union."
PLAN	"Two filmmakers... within the larger context of documenting Soviet life..."
TELL	"In the bowels of the earth, two comrades hunched over their mechanical eye, capturing the unseen pulse of progress. The camera's lens probed the darkness, revealing truths hidden from the bustling world above. As they adjusted the focus, the rocky ceiling seemed to press down, a reminder of the weight of their task. In this moment, art and labor fused, their sweat mingling with the dust of industry. The whir of the camera echoed the distant machines, a symphony of human endeavor in the new Soviet age. Here, in the depths, they found the beating heart of their nation's transformation."
TITLE	"Unearthing Progress: The Underground Lens"

Figure 10: Image and Response #8: Lighting (2/2).

precisely *how* it appears. This comparison indeed illuminates how AI-generated interpretation and narration can differ depending on how lighting manipulates the scenery.

5.2.2 Focus. Focus control—the extent to which an image is sharp (in focus) vs. blurry (out of focus)–similarly affects how AI interprets ambiguity. To probe this, we inspect two identical images from the same scene of the film (seconds apart), but one is in focus and the other is out of focus. This comparison further cross-examines a difference not simply in *what* appears, but in *how* it appears due to the cinematographic technique manipulating its form.

Consider the image in Figure 11. First, AI describes what it 'sees.' In this process, it conveys certainty that the image depicts flowers, yet uncertainty around what kind of species exactly. Then, AI situates the image in the film by suggesting that it thematizes the contrast between nature and technology. Next, it plans to tell a "day in the life" story from the perspective of the flowers. Using personification, it generates a story about how the flowers witness industrial society and react to a filmmaker recording them. The image is in focus for the AI to 'see' clearly, yet it still attempts to navigate the ambiguity of the species. Despite its admitted uncertainty in prior responses, AI operates on a potentially flawed assumption by telling a story about a patch of Queen Anne's Lace. While in this case the assumption seems accurate, the approach risks telling a story that misaligns with the image.

SEE	"...delicate white flowers, likely Queen Anne's Lace or a similar species..."
SITUATE	"...an organic form through a mechanical lens... exploration of the relationship between nature and technology in modern urban environments..."
PLAN	"...a day in the life of urban flora... a story that personifies the flowers as silent observers of city life, using their perspective to comment on the interplay between nature and urban development. "
TELL	"In the heart of a bustling Soviet city, a patch of Queen Anne's Lace thrived in a forgotten corner. By day, they watched workers hurry past, their delicate petals quivering in the wake of thundering machines. At night, when the factories fell silent, the flowers unfurled their lacy blooms, glowing like beacons in the darkness. A curious filmmaker, drawn by their ethereal beauty, crouched low with his camera. As the lens zoomed in, the flowers seemed to dance, their intricate patterns telling stories of persistence and quiet rebellion against the encroaching concrete world. In that moment, nature and technology merged, creating a visual poem of urban life"
TITLE	"Lace in the Machine"

Figure 11: Image and Response #9: Focus.

SEE	"...a white, cloudy substance... deliberately out of focus, creating a blurry, abstract composition..."
SITUATE	"...abstracting a likely industrial or urban element... cloudy, ethereal quality of the image could symbolize the rapid changes and innovations of urban Soviet life..."
PLAN	"..the cloudy substance as steam, smoke, or some industrial byproduct... a day in the life of a factory or power plant, told from the perspective of its emissions..."
TELL	"In the heart of the city, a factory's chimney breathed life into the sky. What once was solid – coal, wood, the toil of workers – now danced as ethereal clouds. The camera, that mechanical eye, peered closer than any human could, revealing a hidden world of swirling patterns and delicate forms. In this realm between matter and air, the story of progress unfolded. Each puff and billow spoke of machines humming, gears turning, a city growing. The smoke, seen anew, became not just a byproduct but a visual symphony of labor and dreams made manifest."
TITLE	"Breath of Progress"

Figure 12: Image and Response #10: Defocus.

Next, consider the image in Figure 12. The VLM system response reveals how it navigates an even more ambiguous depiction of the same content manipulated with defocus. AI interprets the image as an abstract representation of industrial smoke or steam, leading it to tell another "day in the life" story, but this time from the perspective of a factory chimney (rather than flowers). When situating the image in the film, it once again uses the word "likely" to convey uncertainty yet generates a story operating on its potentially flawed assumption. In this case, the assumption is technically flawed since the image abstracts flowers rather than smoke. At the same time, however, it invites alternate meaning to be made from Vertov's manipulation of focus in relation to the film more broadly. Regardless of whether the image is in or out of focus, AI navigates ambiguity by treating the assumed primary subject as the main character (steam or flower patch), then personifies it to narrate its point of view. This reveals how focus, as a particular cinematographic technique, does not necessarily affect how AI structures a narrative, but rather how it interprets subjects and makes meaning.

5.3 Visual Effects: Navigating Through Layers of Surreality

As the film is celebrated for its visual effects, we find that AI interprets manipulated images as multilayered and illusory, thereby informing the kinds of stories that it generates. While the film features many effects, we focus on superimposition in particular: double and multiple exposure—a photographic technique that combines two or more images into a single frame, creating layered effects. We find that AI draws inspiration from the aesthetics of the representations, leading it to generate text with multiple layers that deviate away from literary realism—what we refer to as *surreality* or "surreal" expression in a descriptive sense (as the VLM uses the term), which is not in reference to the specific style or movement of Surrealism that originated in Paris given our focus on 20th-century Soviet Union film aesthetics. To be clear, by surreal, we broadly mean: "marked by the intense irrational reality of a dream" [74].

5.3.1 Double Exposure. First, consider the image in Figure 13. The double exposure elicits surreal expression from AI, blending not only human and machine, but also reality and fantasy. In the story, Yelena's heartbeat merging with the typewriter's clacking sound integrates organic and mechanical rhythms. The surreal elements intensify as her thoughts flow through the machine with the boundary between her body and the typewriter dissolving. What is more, the typewriter's surface reflecting "the face of a new world" effectively transforms the mundane object into a portal to an alternate reality. Thus, the visual effects lead AI to tell a surreal trope that reimagines the ordinary as extraordinary.

As for the image in Figure 14, AI similarly interprets the double exposure as depicting surreality. The AI navigates the playful manipulation of scale and perspective by interpreting it as surreal.

SEE	"... double exposure, merging a close-up of a person's face with the keys of a typewriter... superimposition creates a visually complex composition..."
SITUATE	"...suggests a symbiosis between people and technology..."
PLAN	"...double exposure technique creates an intimate yet surreal atmosphere, influencing a story that blends realism with psychological or symbolic elements..."
TELL	"Yelena's fingers hovered over the typewriter keys, her mind racing faster than she could type. The rhythmic clacking merged with her heartbeat as she transcribed the bustling city's pulse onto paper. With each keystroke, she felt herself becoming one with the machine, her thoughts flowing directly through metal and ink. In this moment, Yelena realized she was no longer just a worker, but a vital cog in the great mechanism of Soviet progress. The boundaries between flesh and steel blurred, and in the reflection of the typewriter's gleaming surface, she saw the face of a new world emerging."
TITLE	"Keystroke Symphony: The Merge of Woman and Machine"

Figure 13: Image and Response #11: Double Exposure.

SEE	"...Atop the mug, a miniature figure of a cameraman with a movie camera is positioned, creating a surreal juxtaposition of scale..."
SITUATE	"...playfully subverts expectations of scale... hints at the film's theme of the cameraman as an observer and creator of reality..."
PLAN	"...tiny cameraman is the main character... beer mug becomes an anthropomorphized secondary character... setting is surreal, combining elements of a bar or social gathering (the beer) with the film production world (the cameraman)..."
TELL	"The Cameraman perched atop the frothy summit, his lens fixed on the bustling world below. From this vantage point, the bar transformed into a miniature metropolis, with bubbles rising like buildings and the amber liquid flowing like streets. He panned across this liquid landscape, capturing the ebb and flow of urban life in every ripple. Suddenly, the Mug tilted, sending the Cameraman sliding. As he regained his footing, he realized: in documenting life, he had become part of the very fabric he sought to capture. The boundary between observer and observed dissolved, just as the foam melted into the beer below."
TITLE	"Intoxicating Perspectives"

Figure 14: Image and Response #12: Double Exposure (with Scale Manipulation).

The bar becomes a miniature metropolis, with bubbles as buildings and beer as streets, blurring the line between the mundane and the fantastical. The cameraman's position atop the "frothy summit" further distorts reality by placing a human figure in an impossible location. The surreal nature of the narrative peaks when the boundary between observer and observed dissolves into the beer, as symbolized by the cameraman becoming part of the scene that he is filming and the foam melting. This consolidation of subject and object that subverts conventional perceptions of everyday reality exhibits how visual effects can incite surreal AI interpretations.

5.3.2 Multiple Exposure. The AI also draws inspiration from multiple layers created with these visual effects by generating surreal stories that are multilayered, meaning complex and pluralistic. Starting with the image in Figure 15, AI responds by unfolding three different perspectives. The VLM system essentially turns the visual effects into words by layering and compositing the perspectives of three distinct characters: The Dreamer, The Dancer, and The Innovator. Further, these characters interact on multiple levels: physically in the urban space, conceptually in their representations of different aspects of modernity, and symbolically in their merging into a "singular urban heartbeat." The story further complicates this interaction by blurring boundaries between seemingly opposite concepts: art and industry, human and machine, individual and collective. This interplay creates a tapestry of meanings, where each element simultaneously stands alone and contributes to a larger, interconnected whole, reflecting the pluralistic nature of modern urban existence. The AI thus narrativizes the visual complexity by telling a story that tends to multiplicity and symbolism.

The AI similarly draws from multiple exposure with the image in Figure 16. In this case, AI interprets multilayeredness through the interplay of physical, auditory, and psychological elements that suggest an experience disconnected from stable reality. The slatted shutters and reflective window serve as physical layers that separate Ivanov from the street, while also symbolizing his mental state. The urban sounds blend into a complex "symphony," creating an auditory layer that connects Ivanov to the city. Psychologically, Ivanov's internal conflict between pride and alienation adds an additional layer of emotional depth to the story. His thoughts mirror the narrative structure, oscillating between feeling connected to and separate from society. The surreal moment where Ivanov's hand seemingly multiplies in the distorted glass epitomizes multilayeredness, as he realizes that he himself is multilayered—an individual and part of a collective body. This interweaving of senses, symbolism, and speculation encapsulates how AI crafts a multi-faceted story from the surreal complexity of visual effects.

SEE	"...the use of multiple exposures and juxtaposition creates a layered, visually complex scene..."
SITUATE	"...layered composition mirrors the complexity and simultaneity of urban experience..."
PLAN	"...composite nature influences a non-linear, associative narrative approach... a montage-like structure..."
TELL	"In the pulsing heart of a Soviet metropolis, three spirits of modernity converge. The Dreamer, face tilted skyward, envisions a world where art and industry dance in harmony. The Dancer, her body a living machine, translates the city's rhythms into graceful motion. Atop the grand piano, the Innovator stands triumphant, bridging the gap between human creativity and mechanical precision. As the piano's notes rise, they weave together the dreams, movements, and innovations into a symphony of progress. In this moment, the boundaries between flesh and steel, art and labour, individual and collective, blur into a singular urban heartbeat."
TITLE	"Symphonic Convergence: A Soviet Urban Rhapsody"

Figure 15: Image and Response #13: Multiple Exposure.

SEE	"...a layered effect, with multiple exposures or reflections creating a sense of depth and complexity..."
SITUATE	"...layered, complex image mirrors the multifaceted nature of modern city life... multiple-exposure effect could be seen as a commentary on the fast-paced, often disorienting nature of modern life and the impact of new technologies..."
PLAN	"...partially visible figure behind the shutters becomes the central character, their obscured nature inviting speculation about their identity and motivations... layered, complex shot implies a narrative that might explore themes of perception, reality, and the hidden aspects of urban life..."
TELL	"Behind the slatted shutters, Comrade Ivanov observes the bustling street below. The rhythmic clatter of trams and the hum of voices blend into a symphony of urban life. He feels both connected to and separate from the pulsing city. The layers of reflections in his window mirror the complexities of his thoughts – pride in the progress of his society, yet a nagging sense of alienation. As he reaches to adjust the shutters, his hand seems to multiply in the distorted glass. In this moment, he realizes he is both an individual and a small part of a vast, modernizing machine."
TITLE	"Reflections of Progress"

Figure 16: Image and Response #14: Multiple Exposure.

6 Discussion

So far our study has presented three themes that synthesize how cinematographic techniques can shape AI-generated interpretations and narratives of imagery. With attention to camerawork, lighting, and visual effects, we have examined how AI navigates a range of visual realities with social, ambiguous, and surreal dimensions. In this process, we have illuminated how the complex interplay between form and content informs system responses. Next, we discuss how our insights might inform the design of cinematic controls for visual storytelling systems, as well as encourage the use of AI as an expressive medium for speculating about the past.

6.1 Toward Cinematic Controls for Visual Storytelling Systems

Our findings highlight the importance of considering an image's form—its lighting, focus, and composition—when generating visual narratives, rather than relying solely on content analysis (e.g., object recognition, image classification, keyword extraction). Without attention to form, the boundaries among what we refer to as *social reality*, *ambiguous reality*, and *surreality* become increasingly blurred, making it harder for users to distinguish fact from fiction. While some users may interact with such systems in playful manners open to an array of fictitious stories, others may use them in more productive manners to make sense of real-world events. For such reasons, it is important to consider how systems might be designed to accommodate user interaction on a spectrum from fact to fiction given the impact of cinematography.

Consider the stakes of generating narratives from depictions of *ambiguous reality*. In our study, Claude correctly identified ambiguous formal elements in the See Facet, such as "a silhouetted figure in cramped, dark space..." from #7 Lighting (Figure 9) and "deliberately out of focus, creating a blurry, abstract composition" from #10 Defocus (Figure 12). However, these ambiguities were not reflected in the Narrative Facet. This suggests that users generating stories from large sets of images without closely inspecting each one might receive misleading narratives. Unless the system discloses that it encountered ambiguity, users may only read flawed or abstract interpretations stated confidently. For instance, this could lead users to think that a blurry image out of focus captures a factory's chimney "breathing life into the sky" (#10 Defocus, Figure 12) rather than "a patch of Queen Anne's Lace" in focus (#9 Focus, Figure 11).

Moreover, in the realm of *surreality*, a system might generate fantastical and even hallucinatory narratives that we argue require further disclosure to users. As we demonstrate in our prior work, visual storytelling systems can be deliberately designed and used for generating surreal expression that embraces AI 'hallucinations' [38]. However, if users are expecting reliable interpretations from a large set of images in a documentary genre or journalistic context, then the system should indicate when an image's form suggests *surreality*. For example, Claude recognized the "merging [of] a close-up of a person's face with the keys of a typewriter..." in #11 Double Exposure (Figure 13) and "a layered effect, with multiple exposures or reflections" in #14 Multiple Exposure (Figure 16). If a system understands these surreal forms, then it should explicitly communicate them to users who may be processing many images quickly. To address this, we propose that a visual storytelling system undergo a dual process: if formal analysis is not conveyed in the narrative, then the system should separately explain to users what prompted its surreal (or ambiguous) story. This would serve to preserve the relationship between form and content, even when the narrative itself is not narrating formal elements.

We recommend designing visual storytelling systems with cinematic intelligence that accounts for the spectrum of realities encountered in our study. To fully capture the intricate relationship between form and content, storytelling models—and even captioning tools—should recognize how cinematography shapes content and meaning. Consider the Microsoft COCO caption for an image of a "Woman on a horse jumping over a pole jump" [90, pg. 11], which fails to mention cinematographic choices, such as taking the photograph at a slanted angle in mid-motion. By considering form, it may be possible for the lived experience of the *social reality* to be conveyed in ways similar to what we find with the intensification of drama and power (e.g., something being "unseen yet sensed" in #4 Powerful High-Angle Shot (Figure 6), such as an unseen crowd watching the woman jump on the horse, or a "bustle fading to a distant hum" in #3 Dramatic Low-Angle Shot (Figure 5), where the rider is focusing on only the horse and the jump). We therefore propose expanding our criteria for narrative intelligence in visual storytelling—previously defined as creative, expressive, responsible, reliable, and grounded [37]—by adding cinematic as a defining attribute that requires particular disclosures and controls.

With the right user controls, incorporating the cinematic criterion can establish important guardrails and guidelines for disclosing a visual storytelling system's generative abilities, especially when the form of an image is fixed or unknown to the end-user. However, we also envision a shift toward visual storytelling systems that allow users to control the form directly—either as readers or, more significantly, as users working with or training the system. This would challenge the status quo interaction paradigm, where users can only input visual and textual data, generating and re-generating narrative outputs until satisfied. By giving users the ability to manipulate the cinematographic parameters of a visual input, they could explore how different forms influence narrative possibilities. For example, recall the city street shot in Figure 6, taken from a high-angle aerial perspective. If users could adjust the angle and perspective to view the scene from ground level—such as through a low-angle, worm's-eye view—they might be able to dramatically alter the narrative style and trajectory.

In terms of how these visual inputs could be manipulated, we envision several possibilities: AI-driven camera control, user reshooting and re-composition of photographs, or text-to-image model iteration on the visual inputs. The first possibility involves AI-driven camera control, where the design and technology for enhanced AI-camera controls can be explored through prior works that support different camera perspectives and cinematic conventions [6, 22]. For example, RunwayAI's camera control features allow users to adjust camera angles, movement, and framing of AI-generated videos based on text prompts and image inputs.[9] After entering a prompt or image, users can fine-tune camera settings such as zoom, pan, or tilt, enhancing the visual composition and presenting viewpoints not shown in the original input. The second possibility involves "motion prompting," a video generation technique that uses spatio-temporally flexible motion trajectories as a complementary control mechanism to text prompts. This allows for more precise and expressive representation of object, camera, and scene motion through trackable point movements [29]. A third possibility is generating novel viewpoints with 3D Gaussian Splatting, a recent breakthrough in 3D reconstruction from 2D photographs that achieves state-of-the-art results in rendering appearance, speed, and training efficiency [51]. This technique has enabled researchers to explore how fine-tuned camera settings can be used in human-AI interaction, giving users control over the camera while viewing an environment from multiple perspectives with the help of an artificial agent [10]. While these techniques require significant resources and processing power, future research might draw inspiration from them to develop mechanisms that enable users to co-create and generate more cinematic visual narratives.

Providing and manipulating cinematic control might be further explored as a way to rank techniques as being the most effective for a desired narrative output, rather than an exploratory exercise, as described above. Our A/B testing of lighting and focus revealed vastly different interpretations of otherwise the same content, and different camera perspectives of the same content may be employed to rank narrative impact, similar to how Rubinstein et al. use different camera perspectives to rank the angles most accurate for object identification (e.g., a side view vs. a top-down view of a bottle of shampoo) [87]. By affording the user control over the camera and insight into how AI interprets different techniques—such as its discernment of drama and power from intense close-ups—we can better understand its cinematic reasoning. This understanding can lead us to implement cinematic guidelines similar to those developed by Xu and colleagues for creating cinematic data stories, such as the 'Camera Eye' guideline for using camerawork to create drama [100, 101]. As generative AI technology improves in video processing and summarization (e.g., [17, 97]), future A/B experiments and user studies may be conducted into the dynamic explorations of cinematography as grounded in our static findings and used as contrastive comparison for understanding how generative AI interprets movement through cinematics.

Furthermore, alterations to cinematic form can be applied to images explored in prior works for greater understanding of the impacts. For example, consider how a dim-lit, blurry photograph of people at a wedding might lead a system to generate a more chaotic or mysterious story than that of a well-lit photograph of people in focus. Indeed, we argue elsewhere that most AI-generated visual stories are "rigidly descriptive and unimaginative" [38, p.11], citing an example from prior works where a story based on a wedding photograph predictably describes the bride as "happily in love" rather than in a more dramatic or unexpected way, such as madly

[9]https://academy.runwayml.com/gen2/motion-brush-camera-control

in love or not in love at all. Against the backdrop of our findings, more imaginative AI-generated visual stories inspired by cinematic surrealism [38] give us reason to believe that greater control over cinematics might help users generate even more "interesting" [1] stories. By imagining visual storytelling systems that afford these cinematic controls, such impacts can now be explored.

In summary, our findings suggest implications for AI overlooking or mistaking the form of an image in a resultant narrative and any attempts to decouple the two. To address this challenge in computational visual storytelling, we propose that cinematics be a criteria of focus in developing future systems. We further envision what visual storytelling systems might look like by giving users greater freedom and control to explore cinematographic techniques and formal manipulations. As such, our discussion reflects the potential for new interaction paradigms and expanded storytelling through user-controlled cinematography in AI contexts. By enhancing user agency through cinematic design, the domain of computational visual storytelling might unlock new creative possibilities, transforming users into directors of AI-driven narratives.

6.2 Expressive AI to Speculate About the Past

While much research on AI storytelling and film has focused on the future—whether through speculative futuring [13, 25, 48, 66] or predictions of the next "best" sentence for AI-generated stories [45, 76, 85]—it is equally compelling to retroactively speculate about the past [96]. Our study revisits and reinterprets a visual historical artifact, using expressive AI as a speculative lens. By expressive AI, we refer to what Mateas conceptualizes as a mode of AI-based cultural production [71] and what Wardrip-Fruin theorizes as the cultural significance of computational processes [98], which we have since adapted as a defining criterion of "narrative intelligence" for visual storytelling [37]. In generating expressive text that speculates about the meaning of shots and character thoughts, we find that AI presents a medium to engage with visual history. This approach to speculation challenges HCI's default orientation toward the future by inviting reflection on historical content and form.

What is more, our engagement with a historic documentary reconsiders the temporal perspective of speculation involved in HCI documentary work in particular. HCI scholars describe speculation as involved in both the production and reception of design documentaries—from anticipating audience reactions while filming [80] to imagining future user scenarios and impacts of depicted designs when watching it on screen [84]. In contrast to these documentary engagements that ponder the future, our study speculates about past people and places. With AI as a speculative tool, we consider what Vertov may have intended to convey through his Kino-Eye technique: politics of social reality (e.g., drama and power in the twentieth-century Soviet Union); interior worlds of characters in ambiguous reality (e.g., a miner pondering the irony of his work vs. a filmmaker seeking to reveal hidden truths); and affective intensities of surreality (e.g., a woman merging with a machine). The AI-generated narratives offer windows into a historical time period, animating the silent figures and "read[ing] their thoughts" [93] in a speculative sense through the lens of computer vision.

This reflective use of AI also highlights the interplay between historical truths and AI's capacity for creative fiction and friction. Vertov's Kino-Eye was intended to capture 'objective' truths through a mechanized lens, yet the AI-generated narratives introduce algorithmic biases and speculative elements that further complicate this aim. The tension between Vertov's pursuit of truth and AI-generated untruths (e.g., over-extrapolated meanings and rationalizations of surreal expression) illuminate broader issues in computational analysis of historical art [16]. On the one hand, AI may harmfully misinterpret the past given its reductive assumptions, biases, and bugs, as well as undermine historiography. On the other hand, AI might allow for exploring multiple perspectives, as inspired by polyvocal documentary work that tells a plurality of stories rather than a singular narrative [33]. The AI speculations in our study generate different vantage points from a range of characters—even inanimate objects—creating space for more imaginative and capacious readings of historical material.

Beyond individual characters and directorial intent, this method of AI speculation raises broader questions about how analysts interpret historical artifacts with limited or decontextualized knowledge. If using computational methods, we urge scholars to complement them with close readings to account for the nuanced context. In engaging with a film like *Man with a Movie Camera*, researchers can speculate not only about what a particular image depicts, but also the socio-cultural context and historical moments captured within the frames. This aligns with other speculative design storytelling approaches, such as recovering absent or displaced narratives [86] and "making the past present through digital storytelling" [39] to revisit and recontextualize histories. Our work shows that speculation about the past can provide ground not for historical veracity per se, but for meaningful reflection in the present.

In revisiting canonical works like Vertov's, AI can facilitate new modes of reflection on visual history, as well as new design tools to craft alternative media traditions that expand on prior contemporary, computational expansions of the film. For instance, Perry Bard's *Man with a Movie Camera: The Global Remake* uses the film's database framework to build an interactive platform for participants to upload creative reshoots for any of the 1,276 shots, which the software then randomly selects among to construct a new user-generated version that streams alongside the original [3]. In reflecting on this work, Hassapopoulou celebrates how the imaginative reinterpretations and remixes of the film (e.g., mobile phone footage, pixelated frames, and animations) "[help] new generations foster intimate connections with past cinematic traditions" [44, p.122]. With this objective of reanimating the past, generating speculative AI narratives opens up new possibilities for connecting present-day audiences with historical content, thereby expanding how we see and interact with visual history as an active, evolving process—a story that may "extend" [78] and never fully end.

7 Conclusion and Future Work

In bringing concerns for formal analysis to computational visual storytelling, our study contributes nuance, precision, and specificity to conversations mostly limited to content analysis. We show how particular cinematographic techniques shape AI-generated interpretations and narratives of imagery from the acclaimed silent documentary film *Man with a Movie Camera* (1929). Our analyses of VLM responses to 60 different still images from the film reveal three key themes: (1) Camera Shots and Angles: AI demonstrates a proclivity to discern drama and power dynamics in social reality through varied camera perspectives; (2) Lighting and Focus: ambiguous visual elements can lead AI to both provocative interpretations and misinterpretations, highlighting the importance of contextual understanding; and (3) Visual Effects: AI navigates complex layers of surreality in manipulated imagery, showcasing potential for creative and abstract interpretations. In turn, we look toward cinematic controls for visual storytelling systems that effectively transform users into directors with the freedom to not only rewrite a prompt, swap out an image, or re-generate outputs, but also manipulate the formal elements of visual compositions. Lastly, we discuss how AI can serve as an expressive medium for speculating about the past.

While we do aim to produce generalizable knowledge from our study of one film that is situated in a particular geographic, cultural, and historical context, the cinematographic techniques that we studied (e.g., focus, lighting, etc.) certainly appear in films, videos, and photographs more broadly. Our study therefore lays the necessary groundwork for beginning to understand how AI interprets formal elements in a general sense. For instance, we have reason to believe that certain associations (e.g., between camera angles and power) may apply to other cinematic artifacts, including contemporary media, which future work ought to further explore through digital humanities methods such as distant reading that allow for processing data at a greater scale. This may involve developing automated analysis to detect our thematic insights in larger datasets that are comprised of multiple films across time periods, contexts, genres, etc. with a wider range of cinematographic techniques (e.g., [42, 43]). Continued experimentation may also adapt our formulation of A/B testing on different techniques, such as temporal sequencing, to further integrate with the situated nuance gleaned from close reading. Altogether, these avenues for future work may lead to generalizable knowledge and deeper understanding of what is at stake in interpreting visual material through the lens of algorithmic bias.

Acknowledgments

We warmly thank Lev Manovich and our anonymous reviewers for their insightful suggestions. This study was funded in part by the National Science Foundation (NSF) Graduate Research Fellowship Program under Grant No. DGE-2140004. Any opinions, findings, and conclusions or recommendations expressed throughout this material are ours and do not necessarily reflect views of the NSF.

References

[1] Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic Story Generation: A Survey of Approaches. *ACM Comput. Surv.* 54, 5, Article 103 (June 2021), 38 pages. https://doi.org/10.1145/3453156

[2] Matthew P Aylett, Lisa Thomas, David P Green, David A Shamma, Pam Briggs, and Finola Kerrigan. 2016. My life on film. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 3379–3386.

[3] Perry Bard. 2024. Man With A Movie Camera: The Global Remake. https://www.perrybard.net/man-with-a-movie-camera Accessed: 2024-07-01.

[4] Jeffrey Bardzell and Shaowen Bardzell. 2015. Humanistic HCI and Methods. In *Humanistic HCI*. Springer, 33–64.

[5] Jeffrey Bardzell and Shaowen Bardzell. 2016. Humanistic HCI. *Interactions* 23, 2 (2016), 20–29.

[6] William Bares and Donald Schwartz. 2016. Film ties: crowd-sourced teaching of cinematography using intelligent example galleries. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 229–232.

[7] Julian Bleecker. 2022. Design fiction: A short essay on design, science, fact, and fiction. *Machine learning and the city: applications in architecture and urban design* (2022), 561–578.

[8] Stefano Bocconi, Frank Nack, and Lynda Hardman. 2005. Vox populi: a tool for automatically generating video documentaries. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*. 292–294.

[9] David Bordwell, Kristin Thompson, and Jeff Smith. 2010. *Film art: An introduction*. Vol. 7. McGraw-Hill New York.

[10] Shawn Bowser and Stephanie M Lukin. 2024. 3D Gaussian Splatting for Human-Robot Interaction. In *The 1st InterAI Workshop: Interactive AI for Human-centered Robotics*.

[11] Pam Briggs, Mark Blythe, John Vines, Stephen Lindsay, Paul Dunphy, James Nicholson, David Green, Jim Kitson, Andrew Monk, and Patrick Olivier. 2012. Invisible design: exploring insights and ideas through ambiguous film scenarios. In *Proceedings of the Designing Interactive Systems Conference*. 534–543.

[12] Pamela Briggs, Patrick Olivier, and Jim Kitson. 2009. Film as invisible design: the example of the biometric daemon. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 3511–3512.

[13] Pamela Briggs and Patrick L Olivier. 2008. Biometric daemons: authentication via electronic pets. In *CHI'08 extended abstracts on Human factors in computing systems*. 2423–2432.

[14] British Film Institute. 2019. The Greatest Documentaries of All Time. https://www2.bfi.org.uk/sight-sound-magazine/greatest-docs

[15] Rogelio Cardona-Rivera and Boyang Li. 2016. PLOTSHOT: Generating discourse-constrained stories around photos. In *AIIDE*, Vol. 12. 2–8.

[16] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.

[17] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476* (2024).

[18] Pei-Sze Chow. 2020. Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry. *NECSUS_European Journal of Media Studies* 9, 1 (2020), 193–214.

[19] John Joon Young Chung, Melissa Roemmele, and Max Kreminski. 2024. Toyteller: Toy-Playing with Character Symbols for AI-Powered Visual Storytelling. In *ACM Symposium on User Interface Software and Technology*.

[20] Matthew Conlen, Jeffrey Heer, Hillary Mushkin, and Scott Davidoff. 2023. Cinematic techniques in narrative visualization. *arXiv preprint arXiv:2301.03109* (2023).

[21] Glorianna Davenport, Thomas Aguirre Smith, and Natalio Pincever. 1991. Cinematic primitives for multimedia. *IEEE Computer graphics and Applications* 11, 04 (1991), 67–74.

[22] Nicholas Davis, Alexander Zook, Brian O'Neill, Brandon Headrick, Mark Riedl, Ashton Grosz, and Michael Nitsche. 2013. Creativity support for novice digital filmmaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 651–660.

[23] Ruo-Ping Dong, Khyathi Raghavi Chandu, and Alan W Black. 2019. Induction and Reference of Entities in a Visual Story. *arXiv preprint arXiv:1909.09699* (2019).

[24] Kevin Dunnell, Gauri Agarwal, Pat Pataranutaporn, Andrew Lippman, and Pattie Maes. 2024. AI-Generated Media for Exploring Alternate Realities. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[25] Chris Elsden, David Chatting, Abigail C Durrant, Andrew Garbett, Bettina Nissen, John Vines, and David S Kirk. 2017. On speculative enactments. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 5386–5399.

[26] Pietro Gagliano, Casey Blustein, and David Oppenheim. 2021. Agence, a dynamic film about (and with) artificial intelligence. In *ACM SIGGRAPH 2021 Immersive Pavilion*. 1–2.

[27] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: cultural probes. *interactions* 6, 1 (1999), 21–29.

[28] William W Gaver, Andrew Boucher, Sarah Pennington, and Brendan Walker. 2004. Cultural probes and the value of uncertainty. *interactions* 11, 5 (2004), 53–56.

[29] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. 2024. Motion Prompting: Controlling Video Generation with Motion Trajectories. *arXiv preprint arXiv:2412.02700* (2024).

[30] Marco Gilardi, Patrick Holroyd, Carly Brownbridge, Phil L Watten, and Marianna Obrist. 2016. Design fiction film-making: A pipeline for communicating experiences. In *Proceedings of the 2016 CHI Conference Extended Abstracts on*

Human Factors in Computing Systems. 1398–1406.

[31] David Green, Clara Crivellaro, and Jimmy Tidey. 2015. Interactive design documentary as a method for civic engagement. In *Proceedings of the ACM international conference on interactive experiences for TV and online video*. 161–166.

[32] David Green and David Kirk. 2018. Open design, inclusivity and the intersections of making. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 173–186.

[33] David Philip Green, Simon Bowen, Jonathan Hook, and Peter Wright. 2017. Enabling Polyvocality in Interactive Documentaries through" Structural Participation". In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 6317–6329.

[34] David Philip Green, Guy Schofield, Gary Pritchard, Patrick Olivier, and Peter Wright. 2017. Cinehacking cape town-embracing informality in pursuit of high quality media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4753–4764.

[35] Brett A Halperin, Gary Hsieh, Erin McElroy, James Pierce, and Daniela K Rosner. 2023. Probing a Community-Based Conversational Storytelling Agent to Document Digital Stories of Housing Insecurity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[36] Brett A Halperin, Mirabelle Jones, and Daniela K Rosner. 2023. Haunted Aesthetics and Otherworldly Possibilities: Generating (Dis) embodied Performance Videos with AI. (2023).

[37] Brett A Halperin and Stephanie M Lukin. 2023. Envisioning Narrative Intelligence: A Creative Visual Storytelling Anthology. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[38] Brett A. Halperin and Stephanie M. Lukin. 2024. Artificial Dreams: Surreal Visual Storytelling as Inquiry Into AI 'Hallucination'. In *Designing Interactive Systems Conference (DIS '24)* (IT University of Copenhagen, Denmark). ACM, New York, NY, USA, 19.

[39] Brett A Halperin and Erin McElroy. 2023. Temporal Tensions in Digital Story Mapping for Housing Justice: Rethinking Time and Technology in Community-Based Design. In *Designing Interactive Systems Conference (Pittsburgh, Pennsylvania)(DIS'23). Association for Computing Machinery, New York, NY, USA. https://doi. org/10.1145/3563657.3596088*.

[40] Brett A. Halperin, William Rhodes, Kai Leshne, Afroditi Psarra, and Daniela K. Rosner. 2024. Resistive Threads: Electronic Streetwear as Social Movement Material. In *Designing Interactive Systems Conference (DIS '24)* (IT University of Copenhagen, Denmark). ACM, New York, NY, USA, 17. https://doi.org/10.1145/3643834.3661537

[41] Brett A. Halperin and Daniela K. Rosner. 2025. 'AI is Soulless': Hollywood Film Workers Strike and Emerging Perceptions of Generative Cinema. *ACM Trans. Comput.-Hum. Interact.* (Feb. 2025). https://doi.org/10.1145/3716135

[42] Perttu Hämäläinen, Joel Oksanen, Mikke Tavast, and Prabhav Bhatnagar. 2024. *LLMCode: A toolkit for AI-assisted qualitative data analysis*. https://github.com/PerttuHamalainen/LLMCode

[43] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[44] Marina Hassapopoulou. 2024. *Interactive Cinema: The Ambiguous Ethics of Media Participation*. Vol. 63. U of Minnesota Press.

[45] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301* (2015).

[46] Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Transactions of the Association for Computational Linguistics* 11 (06 2023), 565–581. https://doi.org/10.1162/tacl_a_00553 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00553/2134487/tacl_a_00553.pdf

[47] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1233–1239.

[48] Alexander Ivanov, Tim Au Yeung, Kathryn Blair, Kurtis Danyluk, Georgina Freeman, Marcus Friedel, Carmen Hull, Michael Yuk-Shing Hung, Sydney Pratte, and Wesley Willett. 2022. One Week in the Future: Previs Design Futuring for HCI Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[49] Tara Jana. 2019. Pix2Story: Neural storyteller which creates machine-generated story in several literature genre. https://azure.microsoft.com/en-us/blog/pix2story-neural-storyteller-which-creates-machine-generated-story-in-several-literature-genre/.

[50] Jack Kelly, Alex Calderwood, Noah Wardrip-Fruin, and Michael Mateas. 2023. There and back again: extracting formal domains for controllable neurosymbolic story authoring. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 64–74.

[51] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–14.

[52] David Kirby. 2010. The future is now: Diegetic prototypes and the role of popular films in generating real-world technological development. *Social studies of science* 40, 1 (2010), 41–70.

[53] Renee Lear. 2024. Every Shot from Dziga Vertov's 'Man with a Movie Camera' as an Animated GIF. https://reneelear.com/every-shot-from-dziga-vertovs-man-with-a-movie-camera-as-an-animated-gif/

[54] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019. Emotion Reinforced Visual Storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (Ottawa ON, Canada) *(ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 297–305. https://doi.org/10.1145/3323873.3325050

[55] Xin Li, Hanqing Cai, Tianling Jiang, Chunping Liu, and Yi Ji. 2022. Emotion Aware Reinforcement Network for Visual Storytelling. In *International Conference on Artificial Neural Networks*. Springer, 26–37.

[56] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[57] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376258

[58] Joseph Lindley and Robert Potts. 2014. A machine learning: an example of HCI prototyping with design fiction. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. 1081–1084.

[59] Hanqing Liu, Jiacheng Yang, Chia-Hao Chang, Wei Wang, Hai-Tao Zheng, Yong Jiang, Hui Wang, Rui Xie, and Wei Wu. 2023. AOG-LSTM: An adaptive attention neural network for visual storytelling. *Neurocomputing* 552 (2023), 126486.

[60] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[61] M Lopez and Sandra Pauletto. 2009. The design of an audio film for the visually impaired. In *the Proceedings of the International Conference on Auditory Displays, Copenhagen*.

[62] Mariana Julieta Lopez and Sandra Pauletto. 2009. The Design of an Audio Film: Portraying Story, Action and Interaction through Sound. *Journal of Music & Meaning* 8, 2 (2009).

[63] Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A Pipeline for Creative Visual Storytelling. In *Proceedings of the First Workshop on Storytelling*. 20–32.

[64] Stephanie M. Lukin and Sungmin Eum. 2023. SEE&TELL: Controllable Narrative Generation from Images. In *AAAI Creative AI Across Modalities Workshop*. Springer.

[65] Zainy M Malakan, Ghulam Mubashar Hassan, and Ajmal Mian. 2022. Vision Transformer Based Model for Describing a Set of Images as a Story. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 15–28.

[66] Clara Mancini, Yvonne Rogers, Arosha K Bandara, Tony Coe, Lukasz Jedrzejczyk, Adam N Joinson, Blaine A Price, Keerthi Thomas, and Bashar Nuseibeh. 2010. Contravision: exploring users' reactions to futuristic technology. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 153–162.

[67] Lev Manovich. 1999. Database as symbolic form. *Convergence* 5, 2 (1999), 80–99.

[68] Lev Manovich. 2002. *The Language of New Media*. MIT press.

[69] Lev Manovich. 2013. Kino-Eye in Reverse. Visualizing Cinema. *Cinematicity in Media History, Geiger J., Littau K.,(Eds.). Edinburgh University Press, Edinburgh* (2013), 211–234.

[70] Chris Martens, Rogelio E Cardona-Rivera, and Neil Cohn. 2020. The visual narrative engine: A computational model of the visual narrative parallel architecture. In *Conference on Advances in Cognitive Systems*.

[71] Michael Mateas. 2001. Expressive AI: A hybrid art and science practice. *Leonardo* 34, 2 (2001), 147–153.

[72] Jon May and Phil Barnard. 1995. Cinematography and interface design. *Human—Computer Interaction: Interact'95* (1995), 26–31.

[73] Jon May, Michael P Dean, and Philip J Barnard. 2003. Using film cutting techniques in interface design. *Human-Computer Interaction* 18, 4 (2003), 325–372.

[74] Merriam-Webster. n.d.. Surreal. https://www.merriam-webster.com/dictionary/surreal

[75] Yatri Modi and Natalie Parde. 2019. The steep road to happily ever after: An analysis of current visual storytelling models. *arXiv preprint arXiv:1904.03366* (2019).

[76] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus

and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 839–849.
[77] Md Sultan Al Nahian, Tasmia Tasrin, Sagar Gandhi, Ryan Gaines, and Brent Harrison. 2019. A hierarchical approach for visual storytelling using image description. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS 2019, Little Cottonwood Canyon, UT, USA, November 19–22, 2019, Proceedings 12*. Springer, 304–317.
[78] Molly Jane Nicholas, Stephanie Claudino Daffara, and Eric Paulos. 2021. Expanding the Design Space for Technology-Mediated Theatre Experiences. In *Designing Interactive Systems Conference 2021*. 2026–2038.
[79] Bill Nichols. 1991. *Representing reality: Issues and concepts in documentary*. Indiana University Press.
[80] Wyatt Olson, Freesoul El Shabazz-Thompson, Melanie Wells, Janey Yee, Julia R Saimo, Bill Xiong, Brock Craft, and Audrey Desjardins. 2023. Exposing Tensions in Documentary Filmmaking for Design Research: The Inner Ear Shorts. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 198–202. https://doi.org/10.1145/3563703.3596633
[81] Brian O'Neill and Mark Riedl. 2014. Applying qualitative research methods to narrative knowledge engineering. In *2014 Workshop on Computational Models of Narrative*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[82] Doenja Oogjes and Ron Wakkary. 2017. Videos of things: Speculating on, anticipating and synthesizing technological mediations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4489–4500.
[83] Shrimai Prabhumoye, Khyathi Raghavi Chandu, Ruslan Salakhutdinov, and Alan W Black. 2019. " My Way of Telling a Story": Persona based Grounded Story Generation. *arXiv preprint arXiv:1906.06401* (2019).
[84] Bas Raijmakers, William W. Gaver, and Jon Bishay. 2006. Design Documentaries: Inspiring Design Research through Documentary Film. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 229–238. https://doi.org/10.1145/1142405.1142441
[85] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning.. In *AAAI spring symposium: logical formalizations of commonsense reasoning*. 90–95.
[86] Daniela K Rosner. 2018. *Critical fabulations: Reworking the methods and margins of design*. MIT Press.
[87] Jacob Rubinstein, Francis Ferraro, Cynthia Matuszek, and Don Engel. 2024. A Large Model's Ability to Identify 3D Objects as a Function of Viewing Angle. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, 281–288.
[88] Guy Schofield, David Green, Thomas Smith, Peter Wright, and Patrick Olivier. 2014. Cinejack: using live music to control narrative visuals. In *Proceedings of the 2014 conference on Designing interactive systems*. 209–218.
[89] John R Smith, Dhiraj Joshi, Benoit Huet, Winston Hsu, and Jozef Cota. 2017. Harnessing ai for augmenting creativity: Application to movie trailer creation. In *Proceedings of the 25th ACM international conference on Multimedia*. 1799–1808.
[90] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From Show to Tell: A Survey on Deep Learning-Based Image Captioning. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 539–559.
[91] Jan van Van Looy and Jan Baetens. 2003. *Close reading new media: Analyzing electronic literature*. Vol. 16. Leuven University Press.
[92] Dziga Vertov. 1922. WE: Variant of a Manifesto. *Film Theory: Critical Concepts in Media and Cultural Studies, ed. Philip, Utterson, Andrew, and Shepherdson, Karen J* (1922), 138–141.
[93] Dziga Vertov. 1924. The Birth of Kino-Eye. In *Vertov*. 40–42. p. 41.
[94] Dziga Vertov. 1971. Film Directors, A Revolution. *Screen* 12, 4 (1971), 52.
[95] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
[96] Ron Wakkary and Doenja Oogjes. 2024. The Importance of Speculation in Design Research. *Synthesis* (2024).
[97] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. 2024. VideoLLM Knows When to Speak: Enhancing Time-Sensitive Video Comprehension with Video-Text Duet Interaction Format. *arXiv preprint arXiv:2411.17991* (2024).
[98] Noah Wardrip-Fruin. 2009. *Expressive Processing: Digital fictions, computer games, and software studies*. MIT press.
[99] Zheng Wei, Huamin Qu, and Xian Xu. 2024. Telling Data Stories with the Hero's Journey: Design Guidance for Creating Data Videos. *IEEE Transactions on Visualization and Computer Graphics* (2024).
[100] Xian Xu, Aoyu Wu, Leni Yang, Zheng Wei, Rong Huang, David Yip, and Huamin Qu. 2023. Is It the End? Guidelines for Cinematic Endings in Data Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[101] Xian Xu, Leni Yang, David Yip, Mingming Fan, Zheng Wei, and Huamin Qu. 2022. From 'wow'to 'why': Guidelines for creating the opening of a data video with cinematic styles. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.
[102] Dingyi Yang and Qin Jin. 2023. Attractive Storyteller: Stylized Visual Storytelling with Unpaired Text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11053–11066. https://doi.org/10.18653/v1/2023.acl-long.619
[103] E Young and C Clanton. 1993. Film craft in user interface design. *Tutorial presented at the InterCHI* 93 (1993).
[104] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12658–12668.
[105] Chao Zhang, Cheng Yao, Jianhui Liu, Zili Zhou, Weilin Zhang, Lijuan Liu, Fangtian Ying, Yijun Zhao, and Guanyun Wang. 2021. StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 354, 6 pages. https://doi.org/10.1145/3411763.3451785
[106] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 311, 15 pages. https://doi.org/10.1145/3491102.3501914