Proceedings of ICADSS 2026 Symposium: Contemporary Visual Arts, AI Creativity, and Interactive Media: Theory and Practice
DOI: 10.54254/2753-7064/2026.BJ38939

Can World Models Generate Films? AI World Model and the Boundaries of Interactive Cinema

Yanao Liu

Beijing Jiaotong University Affiliated High School, Beijing, China
yanao_liu@jidjdfz.com

Abstract. This paper asks whether AI world models can generate films, and more specifically whether they can produce a form of interactive cinema in which viewers intervene in unfolding images. Drawing on André Bazin's account of cinema's founding desire for "a total and complete representation of reality," it situates contemporary world models such as LingBot-World, Happy Oyster, Runway's GWM-1, and Odyssey within a longer history of attempts to reproduce the world through moving images. The paper compares world models with virtual production, using the author's own AI-built lighting simulator as a case study of real-time light and shadow generation and its computational limits. It then reframes AI photography as a probabilistic mode of generation that repositions the photographer as a player exploring a space of possible outcomes rather than an author of a single fixed image. Finally, it turns to the material conditions of world models within the Anthropocene, following Kate Crawford's account of AI as an extractive industry, to argue that computing power, training data, and copyright are inseparable from questions of energy, labor, and planetary cost. The paper concludes that world models can already generate footage assembled into a finished film, but genuinely interactive cinema remains bounded by these material and infrastructural constraints.

Keywords: world models, interactive cinema, virtual production, AI photography, Anthropocene

1. Introduction

While watching the MediaStorm channel, known in Chinese as *yingshi jufeng*, on Bilibili, I saw that its latest episode was titled "Tossing a Coin in AI: Is the Probability 50%?" MediaStorm is a well-known Chinese technology and filmmaking channel and production studio. The episode presents an experiment by the channel that uses AI-generated coin tosses to examine whether a world model can reproduce the probabilities of the physical world.

In this MediaStorm episode, the reason why the probability of a coin landing heads might not be 50 percent is that the training material for world models comes from footage recorded in the real world. In this footage, the heads side of a coin can display a clearer portrait, denomination, or more attractive design, so most recorded coin tosses show the coin landing heads. This makes the world model's random draw mechanism more complicated when it generates a contingent event in the real

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.BJ38939

world, such as a coin toss. Because heads and tails are not equally represented in the training data, the generated results are not equally distributed either.

The coin-toss experiment raises two distinct questions: how coins behave in the physical world, and how biases in training data shape generated outcomes. The challenge for world models is therefore to reconstruct, through recorded images, a world whose behavior resembles physical reality. This ambition to reconstruct reality connects world models to a much older cinematic aspiration. In *What Is Cinema?*, André Bazin approaches the development of cinema through the desire for "a total and complete representation of reality" [1]. In "The Myth of Total Cinema," André Bazin argues that the aspiration to reproduce the world in movement, sound, color, and depth preceded the technical means of realizing it. Cinema emerges from an interaction between this imagination and the material possibilities of recording. This argument offers a starting point for examining contemporary world models. When a computational system can generate an environment and respond to actions within it, how does the ambition to reproduce a world change? What might this mean for the production and experience of cinema?

Can world models generate a film? The question concerns several different operations. A model might generate footage that a filmmaker selects and organizes into a finished work. The term interactive cinema therefore identifies a central problem for this study: what happens when a viewer can intervene in the unfolding images? Interactivity during production must also be distinguished from interactivity during viewing. A filmmaker's navigation through a generated world may still result in a fixed film for its audience.

The paper proceeds in four sections. The first examines definitions of world models and distinguishes their demonstrated capabilities from claims about filmmaking. The second develops a cinematic account of world models through virtual production and the light and shadow simulator. The third considers generation as a mode of cinematography, focusing on probability, creative control, and contingency. The fourth examines computational constraints, training data, and the dilemmas of the Anthropocene. Together, these sections return to the opening question by specifying what a world model generates, how that output becomes cinema, and whose work and resources make the process possible.

2. What is a world model

World models were initially applied in the field of autonomous driving, where they were used to simulate vehicles' decision-making and responses under complex road conditions [2]. However, the uncertainty of their content generation, together with the high degree of realism in their scene simulations, has unexpectedly revealed their potential for cinematic storytelling—this combination of unpredictability and realism seems almost inherently suited to moving-image creation.

Although the development of world models around the world is still at an early stage, the work of several Chinese and international teams has already begun to show initial results [3, 4]. In China, for example, LingBot-World can generate a 16-fps interactive world simulator with virtually no latency. It also allows users to switch camera perspectives in real time through WASD, offering a more immersive sense of presence and greater freedom of exploration [5]. Another example is Happy Oyster. The developers officially define it as: "Happy Oyster is an open-ended world model for real-time world creation and interaction" [6]. Users can "direct" scenes through text, voice, or images, or explore infinitely expanding worlds, and the system supports continuous audio-video generation [6]. With its powerful computing capacity, it can already generate up to three minutes of 720p video content [7]. Although it remains limited in duration, this is already enough to offer a glimpse of the

DOI: 10.54254/2753-7064/2026.BJ38939

enormous potential behind world models, as well as the technical mechanisms that remain to be perfected.

Runway's GWM-1 is another noteworthy case. "It generates frame by frame, runs in real time, and can be controlled interactively with actions—camera pose, robot commands, audio" [8]. This mechanism of "frame-by-frame generation + action-driven control" further confirms that world models are moving from "passively generating a segment of imagery" toward "actively responding to user commands and evolving in real time" [8].

Odyssey's definition of a world model approaches the concept from another angle. It does not emphasize the form of interaction, but instead directly compares the fundamental difference between language models and world models in their ability to "understand the physical world." Odyssey states that language models essentially only process text tokens and "has no concept of gravity, friction, or three-dimensional space." "It builds an internal simulation of how the physical world operates. It learns from video, sensor data, and interaction logs to predict what happens next when something moves, collides, or changes state. Drop a glass on a tile floor, and a world model can predict the trajectory of the shards. A language model cannot" [2-4].

3. My definition of the world model

3.1. World models and virtual production

A typical example of virtual production is Hollywood's use of game engines such as Unreal Engine 5 to create cinematic visual effects. Swords and Willment [9] define virtual production as a mode of film and television production that uses computer-generated content to visualize and control digital environments in real time on set. Its central feature is that the virtual environment is captured directly "in camera," rather than composited during post-production, as in traditional VFX. This system depends on an entire technological ecosystem, including real-time game engines such as Epic's Unreal Engine and Unity, virtual- and augmented-reality systems, motion capture, camera tracking, dynamic lighting, green- and blue-screen technologies, and LED screens.

A world model is different. It contains an internal simulation of how the physical world operates. The character's next action and the environment's next state are calculated in real time by the model on the basis of the current state and incoming input, rather than being staged in advance by human creators. For this reason, the earlier world models emerge and mature, the sooner virtual characters may be able to escape their existence as "pre-produced puppets" and acquire a genuine temporal life of their own—one capable of developing autonomously.

3.2. The lighting simulator as a world model

Looking back, my initial interest in exploring world models arose during an online summer course, where I first learned how to use AI models to build my own sandbox simulator. As the design and refinement of my lighting simulator entered its final stages, I began to realize that the simulator itself can, to a certain extent, be understood as a world model. It operates according to its own set of rules, and the lighting in each session is calculated in real time rather than pre-rendered as a single image. Moreover, because it was generated by AI based on my own conceptual approach, it can also be regarded as a world model in its own right.

However, the limitations of using AI were also apparent. My initial plan was to start with an empty classroom. After building its lighting environment, I would gradually expand the project to include an office, a corridor, and even a lecture hall. At first, I thought that once the classroom's

DOI: 10.54254/2753-7064/2026.BJ38939

lighting was working well, building the subsequent scenes would simply require finding images that most closely resembled real environments and copying the classroom's lighting settings and rendering setup. I expected the scenes to fit together seamlessly. In practice, however, the drawbacks of rendering in real time on the web began to emerge when I expanded the project to include an office. The browser had to recalculate the lighting for every frame. As more models were added and the lighting had to be rendered again for each new scene, the simulator's frame rate began to fall. It had originally run at more than 30 frames per second even in full screen mode, but later managed only somewhere in the teens, or even just a few frames per second, in a small window.

Another central problem in developing the lighting simulator lies in the real-time refinement of light and shadow. For example, when the camera angle is rotated, the computer must adjust the shadow and light through the window into each corner. Similarly, when the angle of a light source is moved, the light entering the classroom should be blocked by walls, desks, chairs, and other objects in a physically plausible manner, thereby producing the corresponding shadows. At the same time, constructing each scene places a computational burden on both the AI system and the web-based platform. As the number of scenes increases, the simulation of light sources consumes a significant amount of GPU computation, which can lead to lag and reduced performance. This constitutes one of the central bottlenecks—and practical limits—of constructing a lighting simulator through AI.

My lighting simulator provides a practical point of comparison, showing how the construction of a convincing world depends on the relationship between visual design and computational capacity. World models thus open a new possibility for the film industry while continuing cinema's longstanding ambition to recreate reality through moving images.

4. Another form of photography: AI photography as generation

AI photography can be understood as another form of photography because it changes what it means to take a photograph. In conventional photography, the shutter captures a moment that already exists in the world. In AI photography, the image is generated by sampling from a probability distribution conditioned by a prompt. The prompt plays a role analogous, though not identical, to composition, framing, focal length, shutter speed, and aperture. It narrows the range of possible outcomes without determining a single image in advance.

The term "generation" has become increasingly familiar in the age of AI. It should not be understood as the retrieval of a finished answer from a database, nor as creation in the conventional sense. Instead, generation refers to the production of one possible continuation from a probability distribution conditioned by the given context. In a language model, the basic unit of generation may be a token; in an image model, it may be an image; in a video model, it may be a sequence of frames. In a world model, however, generation extends further: the system generates a succession of spatial and temporal states that respond to user input. A world model does not merely generate an isolated image. It can transform a script in the user's mind into a scene, a space, and a sequence of possible events that unfold over time.

4.1. Photography as a probability game

AI photography can be understood as a probability game, although "game" here functions as a metaphor for the process of repeated sampling. Each prompt produces a range of possible outcomes, and the same prompt may produce different results at different moments. The central operation of AI photography is therefore to generate repeatedly until one obtains an image that corresponds sufficiently to the photographer's intention.

DOI: 10.54254/2753-7064/2026.BJ38939

In photography mediated by a physical camera, composition, framing, focal length, shutter speed, and aperture directly shape the image through the camera's optical and material conditions. A prompt does not function in exactly the same way. It does not directly control every element of the image; instead, it conditions the probability distribution from which the image is generated. A single word, a modifier, or a slight change in phrasing can alter the composition, lighting, atmosphere, or spatial arrangement of the result.

The photographer's agency consequently lies less in determining the image than in establishing the conditions under which images can emerge. Through repeated prompting, comparison, and selection, the photographer gradually approaches an image that corresponds to the one imagined. The final result is therefore neither fully controlled nor entirely random. In traditional photography, the shutter captures a moment that already exists in the external world. Within a world model, however, the shutter becomes the precondition of generating the image.

This difference becomes clearer when we consider the uncertainty involved in photographic and film production. During the preparation and shooting of a conventional photograph or film, weather, lighting, location, and environmental conditions may change continuously. Even when these factors are carefully planned, they cannot always be reproduced on demand.

A world model operates differently. If the user wants a particular result, the model can generate the scene repeatedly until one version becomes satisfactory. This does not mean that a world model is fully controllable. But there is still something that cannot be controlled: the user cannot determine what the model will produce exactly. What becomes more manageable is the process as a whole, because the user can repeat the generation many times and select the result that best matches the intended image. The world model therefore exchanges computational time and repeated sampling for a partial reduction of contingency. Conventional photography, by contrast, must largely endure the contingency of the physical world.

5. World models: the dilemma of the Anthropocene

5.1. Computing power, bitrate, and resolution

Under the energy constraints of the Anthropocene, the first technical boundary world models encounter is the difficulty of simultaneously meeting demands for computing power, frame rate, and resolution. One basic fact is that, with each increase in resolution, the computing power required for frame-by-frame generation grows roughly with the square of the resolution—even maintaining a stable 720p image already entails considerable costs. For world models intended to "generate films," this means that image quality, smoothness, and duration can scarcely meet cinematic standards at the same time.

Limits on computing power, frame count, and resolution remain a difficult hurdle for the field. LingBot-World, discussed earlier, is a typical example: although it achieves interactive generation with almost no latency and lets users change viewpoints in real time using WASD controls, its output frame rate is capped at 16 fps—short of cinema's commonly used 24 fps, let alone the higher frame rates needed for slow motion. Even as its freedom of interaction approaches complete openness, its computing power remains insufficient to support a corresponding level of smoothness. LingBot-World's low frame rate is, in essence, a computational compromise forced by the competing demands of "real-time interaction" and "generation quality."

Similar compromises arise in duration and image quality. Happy Oyster supports continuous audiovisual generation, letting users "direct" scenes through text, voice, or images, yet even with substantial computing power it can generate only around three minutes of 720p content—"three

DOI: 10.54254/2753-7064/2026.BJ38939

minutes" and "720p" have almost become shared ceilings for this generation of world models. Runway GWM-1 addresses the same constraint differently, sacrificing visual complexity or generation duration for real-time, action-driven interaction instead. This trade-off is no accident: moving from 720p to 1080p does not simply raise the required computing power by half but to nearly 2.25 times the original amount, and moving to 4K brings an even steeper increase. No company can push image quality, frame rate, and duration to their limits simultaneously, because the computing power, energy, and hardware that sustain them are themselves scarce resources.

Resolution and bitrate are distinct parameters. Platform bitrate limits and compression can degrade image quality, affecting the video material used to train world models such as LingBot-World, Happy Oyster, and Odyssey. Many streaming platforms and social media services impose strict bitrate limits on videos uploaded by creators, directly affecting the quality of the finished work. To reduce costs and computational demands, some platforms even quietly lower bitrates without informing creators, substantially degrading image quality or audiovisual synchronization. This gap between bitrate and resolution may appear to be merely a matter of "user experience," but it indirectly determines the quality ceiling of world models' training data: LingBot-World, Happy Oyster, and Odyssey all fundamentally rely on video material from platforms and the internet for training. In other words, if the training material already contains distortion, blurring, or audiovisual misalignment caused by bitrate compression, the motion and detail information from which the model learns may also be affected. This concern is echoed in Odyssey's definition: it emphasizes that language models have "no concept of gravity, friction, or three-dimensional space," whereas a world model "learns from video, sensor data, and interaction logs to predict what happens next when something moves, collides, or changes state." The bitrate problem therefore ultimately points to a question of "cinematic quality": whether world models can generate convincing motion, light, and shadow is closely related to how much sufficiently clear and coherent footage of the world they have seen.

5.2. Whose world is being generated

The computational limits of world models raise questions about who supplies the resources and images that make generation possible, and who bears its costs. Godfrey Reggio's *Koyaanisqatsi* [10] offers a cinematic starting point for examining these relations. Through time-lapse photography, slow motion, and Philip Glass's repetitive score, the film presents industrial infrastructure, urban movement, and mass consumption as parts of an accelerating system. Read retrospectively through the Anthropocene, it makes visible the dependence of technological civilization on land, energy, and material resources. The apparent freedom to reshape the world rests on extensive material operations that remain largely outside individual perception.

Kate Crawford's *Atlas of AI* [11] places artificial intelligence within this material system. She describes AI as an "extractive industry" dependent on the extraction of energy, minerals, labor, and data [11]. Her discussion of mining, data centers, and water consumption challenges the apparent immateriality of the "cloud," locating computation within the Earth and its resource systems [11]. Applied to world models, this perspective reveals that generating a virtual environment depends on transforming and consuming resources in the physical world. The ability to produce images on demand therefore raises questions about how the labor and environmental costs supporting that production are distributed.

The resources sustaining generation also include existing images and the labor embodied in them. When videos and photographs become training material, the work of filmmakers, photographers, and other creators enters the infrastructure through which new images are generated. Therefore,

DOI: 10.54254/2753-7064/2026.BJ38939

copyright concerns emerge here as questions about the conditions under which those works are obtained and used, and how their creators' contributions are recognized.

Whether training materials were lawfully obtained and used must be distinguished from whether a particular generated output infringes copyright. These questions cannot be resolved by assuming that infringement is inevitable. The scale of training datasets nevertheless creates difficulties for tracing sources and establishing responsibility. Privacy and representational bias introduce further concerns: the records through which a model learns may privilege particular experiences while excluding others. The opening coin-toss example illustrates how the distribution of recorded images can shape generated outcomes. A world model's apparent realism does not, by itself, establish that it represents the world comprehensively or without bias.

6. Conclusion

Returning to the initial question: can world models generate films? If "film" refers to a finished moving-image work that can be selected, edited, and ultimately fixed, the answer is already close to "yes." World models can begin with a single static image and generate scenes, spaces, and sequences of events unfolding over time. Creators can then select and organize this material into a fixed film for an audience. If, however, "film" further refers to the "interactive cinema" named in the title, in which viewers can intervene in and alter the unfolding images during viewing, the answer requires considerably more caution: this boundary is currently determined primarily by material conditions.

The relationship between world models and cinema thus reenacts what Bazin calls "the myth of total cinema": the desire to reproduce a complete world always precedes and exceeds the technical means of realizing it. World models take this desire to new heights, aspiring to generate a world that can be entered and altered; their learning, however, remains confined to "images of the world." The coin that always lands heads up in the training material reminds us that world models faithfully reproduce the visual records humans have left of the world, together with their various biases. For this reason, the question "Can world models generate films?" ultimately extends beyond technical feasibility to a question that demands further attention: whose images, whose labor, and whose resources does such a film use, and who pays its planetary costs?

References

[1] Bazin, A. (1967). The myth of total cinema. In H. Gray (Trans.), *What is cinema?* (Vol. 1, pp. 17–22). University of California Press. (Original work published 1958)
[2] Cameron, O. (2026, February 17). *Why we must build world models*. Odyssey. https://odyssey.ml/why-we-must-build-world-models
[3] Odyssey. (n.d.). *Odyssey*. Retrieved September 10, 2026, from https://odyssey.ml/
[4] Cameron, O. (2025, May 28). *Introducing Odyssey-1: A playable world model*. Odyssey. https://odyssey.ml/introducing-odyssey-1
[5] Robbyant Team. (n.d.). *LingBot-World: Advancing open-source world models* [Code repository]. GitHub. Retrieved September 10, 2026, from https://github.com/Robbyant/lingbot-world
[6] Happy Oyster. (n.d.). *Happy Oyster: Real-time world model for interactive creation*. Retrieved September 10, 2026, from https://www.happyoyster.com/
[7] Alibaba Cloud Community. (2026, April 20). *Alibaba launches HappyOyster, a world model product for real-time immersive creation and interaction*. Alibaba Cloud. https://www.alibabacloud.com/blog/alibaba-launches-happyoyster-a-world-model-product-for-real-time-immersive-creation-and-interaction_603048
[8] Runway. (2025, December 11). *Introducing GWM-1*. https://runway.com/research/introducing-runway-gwm-1
[9] Swords, J., & Willment, N. (2024). The emergence of virtual production – A research agenda. *Convergence: The International Journal of Research into New Media Technologies, 30*(5), 1557–1574. https://doi.org/10.1177/13548565241253903

DOI: 10.54254/2753-7064/2026.BJ38939

[10] Reggio, G. (Director). (1982). *Koyaanisqatsi* [Film]. Institute for Regional Education.
[11] Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.