
SMPTE
MI
MOTION IMAGING JOURNAL
MPEG MEDIA TRANSPORT
ENCODING EFFICIENCY
BROADCAST SIGNALING
VSF NMOS Specification
Fast Metadata Framewo

TECHNICAL PAPER

KEYWORDS GENERATIVE AI // GENAI // VIDEO GENERATION // VIDEO GENERATION MODELS // PROMPT ENGINEERING

As LLM´s já transformoram a forma como criamos textos e imagens — e agora estão revolucionando o universo do vídeo. O presente artigo revela como os "Modelos de Geração de Vídeo" (VGMs) estão reduzindo custos de produção (coisa de até 40%), acelerando fluxos de trabalho e abrindo novas possibilidades criativas para o audiovisual. Você, caro leitor, descobrirá como a engenharia de ***prompts*** e conceitos de cinematografia se tornam ferramentas-chave para obter resultados profissionais, além de conhecer aplicações práticas em pré-produção, edição, efeitos visuais e monetização de acervos. Uma leitura essencial para quem quer se preparar para o futuro da mídia, vem que é aprendizado na certa!

Tom Jones Moreira
(tvdigitalbr@gmail.com)

Video Production with Generative AI

By Brent Rabowsky

Abstract

There are many kinds of Generative AI (GenAI) models, from audio to text generation models. However, video generation models (VGMs) are foundational for video production and are gaining importance as their expressiveness increases. VGMs differ from other GenAI models in that VGM users must have domain-specific knowledge to leverage VGMs to express their artistic vision optimally. Specifically, it can be demonstrated that VGM users must know the basics of the visual language of cinematography to translate their creative vision into quality output correctly. Prompt engineering is used to craft VGM output into something useful and aesthetically viable. Using prompt engineering, it is possible to ask a VGM to instantiate the basic elements of cinematography, such as camera placement, camera movement, shot composition, shot size, focus/depth of field, and lighting. Besides VGMs, other GenAI models are useful for pre- and post-production tasks.

Although Generative AI (GenAI) initially captured attention for its ability to generate text, its reach has been rapidly extended to audio, image, and video generation. Recent advancements in video generation models (VGMs) have proved their value for many use cases, including previsualization, second unit/b-roll, and short-form content. As these VGMs have evolved, it has become clear that prompt engineering is the most important technique for their effective use. Prompt engineering allows users to apply the visual language of cinematography to craft the VGM output into something useful and aesthetically viable. Well-crafted prompts enable users to incorporate cinematographic elements such as camera placement, camera movement (for example, tracking shots), shot composition (how elements in a scene are arranged), shot size (for example, close up), focus/depth of field, and lighting. With effective prompt engineering, GenAI can be employed across a wide variety of pre-production, production, and post-production video workflows. The efficiency gains for human creatives enabled by GenAI are crucial to the long-term economic viability of media companies. As the CEO of Sony Pictures Entertainment, Tony Vinciquerra,

stated in 2024, "The biggest problem with making films today is the expense. We will be primarily looking at ways to produce films for theaters and television more efficiently, using AI."[1] In one case, using GenAI virtual production techniques reduced the production budget of a sci-fi film by more than 40%, enabling the production and release of a film that otherwise would have been too expensive and never made.[2] It also is important to keep in mind that although the focus herein is on using GenAI to generate and process video frames, GenAI also has a much wider application in media workflows via the generation and processing of audio, images, and text.[3]

Video Generation Applications and Limitations

VGMs generate short videos, typically of two minutes duration or less, when guided by input prompts consisting of images, text, and/or other videos. Although VGMs are still in an early evolutionary phase, they have already been applied in many kinds of published content. Released short-form content mainly based on VGM output includes ads, music videos, and other short films. Long-form content has also been produced with VGM assistance for specific tasks such as previsualization. Media companies are also exploring other possible uses of VGMs. For example, many media companies have archives of older, dormant properties. Using GenAI to produce short-form content featuring these properties could be a cost-efficient way to test their current appeal. GenAI can also transform or summarize long-form content into more easily consumable short-form content, another potential way to reuse and monetize media archives. Other possibilities include using VGMs in the creation of educational and training videos.

VGMs can be applied across all stages of production: pre-production, production, and post-production. In pre-production, VGMs are used to generate previsualization assets such as animatics. This can be done in conjunction with image generation models to generate storyboards and concept art. At the production stage, VGMs can be used to generate "raw footage." Much current interest pertains to generating B-roll and second unit type footage. Concerning post-production, in 2024, a leading editing software provider incorporated VGMs for the first time to generate additional frames in an editor's timeline.[4] There are also many other possible uses in post-production for VGMs, which will be discussed separately below.

As of early 2025, VGMs have many limitations. Resolution typically is limited to 1080p or less, though 4K resolution is on the feature roadmap for major VGM providers. Also, VGMs typically produce only short clips with lengths of two minutes or less. This is due to multiple issues, including hardware limitations related to available GPU memory and difficulties pertaining to the lack of visual continuity or consistency of objects. Leading VGMs have now achieved an acceptable level of object consistency within a single VGM output video. However, different VGM outputs using the same or similar prompts can be quite different due to the stochastic nature of VGMs. Perhaps most importantly, the degree of expressive control of VGM output can be limited for the same reason, as well as limitations in the training data. VGMs are still learning the nuances of camera movement, for example. As a result of VGMs' current limitations, to date, they have had limited application to "high-end" or premium long-form content and have been leveraged most extensively for ads and other short-form content.

Some of the expressive control limitations of VGMs are illustrated by considering the efforts that went into the widely seen VGM-based short film *Air Head* (2024). The unusable to usable VGM output ratio was 300:1, with many output clips requiring extensive editing with post-production techniques.[5] Balanced against these limitations are the efficiency gains enabled by VGMs. For example, it took only two weeks for a team of just three people to produce *Air Head*. VGMs

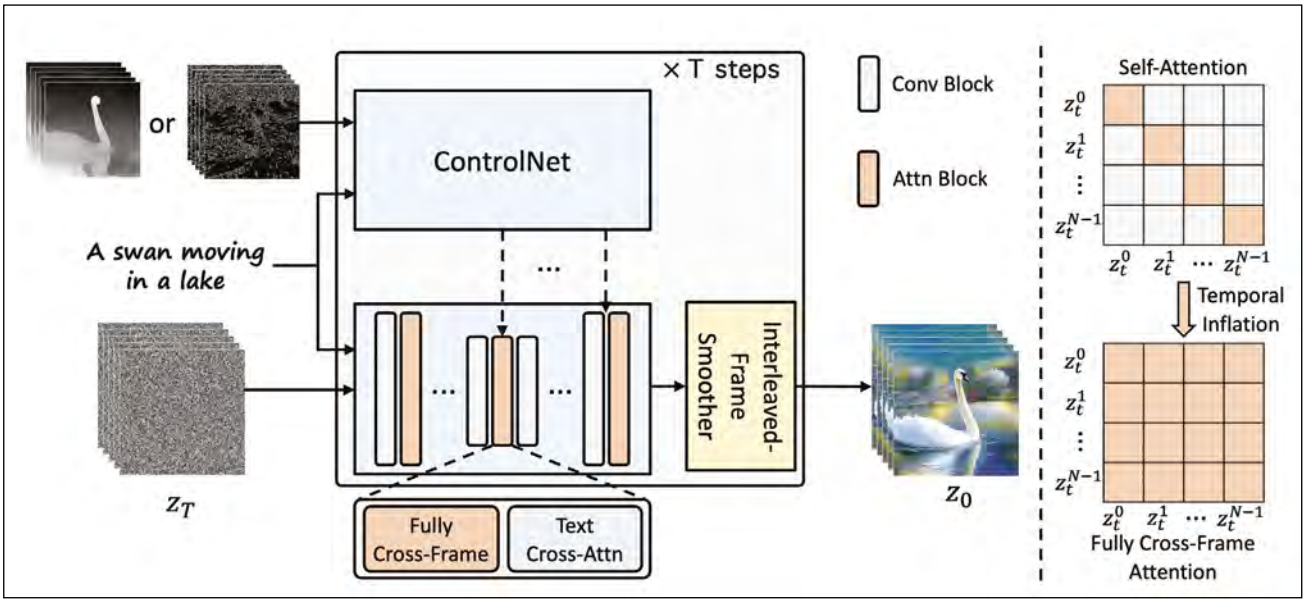

FIGURE 1. ControlVideo VGM architecture centered around an image generation model, ControlNet.

may also open new creative possibilities that are not otherwise feasible. An example is the music video, *The Hardest Part* (2024), where the lead creative aimed to produce a film that looked like a "strange feed into memories from another dimension."[6] A similar sentiment was expressed by the creative director for the first VGM-produced ad (for toy retailer Toys R Us), who stated that using a VGM "allows you to explore your imagination... in motion."[7]

How Video Generation Models Work

VGMs are a kind of Foundation Model (FM), a category of neural net models that can deliver responses to a wide variety of tasks. VGMs are an extension into the time dimension of image generation models, which centrally leverage a diffusion model component. Diffusion models are developed by slowly adding random noise to data and training the model to reverse this "diffusion" process, thereby shaping the noise into an aesthetically pleasing output.

Extending the models into the time dimension significantly increases the overall model complexity, as is apparent by examining the architecture of open-source VGMs such as Stable Video Diffusion and ControlVideo. Like any typical VGM, at the core of ControlVideo is an image generation model, in this case, ControlNet, as seen in the ControlVideo architecture diagram of **Fig 1**.[8] In addition to the image generation model, other components are necessary to extend the output into the time dimension. As the video output is generated, it is processed by a cross-frame interaction module to ensure consistency of objects between frames, then an interleaved frame smoother module to avoid flicker, and also a sampler module to produce videos in a clip-by-clip manner to reduce the burden on GPU memory and computing.

In addition to the complexity of VGM architecture, the process of training VGMs is also complex and challenging. The most intensive phase of the training process is pretraining, which involves training a VGM on a very large, unstructured, and unlabeled dataset so the VGM can handle a wide range of video generation tasks. It also is possible to do "continuous pretraining" as more video files become available to keep a VGM "up to date" and deepen the VGM's knowledge of visual language and concepts. Another important form of training is fine-tuning, which allows users to adapt existing trained VGMs for a particular task, often using a relatively small labeled training dataset. For example, in the case of a VGM that will be used to produce video in an anime style, it would be possible to fine-tune the VGM to adhere to the style of a particular anime series by fine-tuning the VGM on episodes of that series. The overall training regimen for a VGM may combine these techniques. For example, Stable Video Diffusion employs a multi-step training process, including pretraining and fine-tuning, to ensure optimal model quality while training at scale.[9]

As mentioned, any FM requires a large amount of data during training for best results. In the case of VGMs, tens of petabytes of video files could be used during the training process; a larger media archive might have millions of hours of footage amassed over multiple decades. Preparing these assets to be suitable as training data for a VGM would require substantial time and money. A pipeline to process the assets before training would be necessary. This pipeline would include many processing steps such as (1) identifying third-party content that must be redacted; (2) adding metadata to inform training; and (3) transcoding all of the disparate formats to a common format suitable for training input. To process a petabyte-sized media archive, this pipeline might need to run for several weeks to process the entire archive. Despite the heavy lift in preparing assets for VGM training, it may become a significant opportunity for media companies to monetize their media archives if various issues, such as pricing models, can be resolved.[10] In one publicized deal involving training a VGM on a studio's archive, the studio emphasized that the VGM would be customized to the studio's proprietary portfolio and used to augment workflows.[11]

The Visual Language of Cinematography

To achieve professional-quality results with VGMs, it is necessary to apply the visual language of cinematography. This distinguishes VGMs from other types of FMs in that other types do not require users to have specialized expertise; instead, users can draw upon common background knowledge. For example, using a text generation model to create a document summary can be as simple as writing a prompt to the effect of "summarize this text." However, advanced prompt engineering techniques will help optimize the output of any FM.

While cinematography dates back to the late nineteenth century with the invention of the first motion picture cameras, the most notable early explication of the related concepts and terminology was first published in 1949 in John Alton's pioneering book, "Painting with Light."[12] The book codified observations about what works and what does not, with interesting insights illustrating the many complex decisions a cinematographer must make. For example, Alton stated that a shot looks best when the foreground is darkest, the midground is correctly exposed, and the background is the lightest. Although Alton's book is known for his insights regarding lighting techniques, cinematography involves much more than just lighting.

Practitioners may have varying perspectives on the essential elements of cinematography. However, any list of the core elements likely would include the following:

- Camera placement
- Camera movement
- Shot composition
- Shot size
- Focus/depth of field
- Lighting

These elements of cinematography must be considered in relation to the many rules that have evolved over 100 years of trial and error. One example is Alton's rule about the progression from dark to light to create a feeling of depth. Another example concerns the lateral direction of characters' motion, which can significantly impact how viewers perceive the action. Left-to-right motion may be viewed more positively compared to right-to-left motion. While breaking such rules can sometimes produce good results, failure to understand the rules and apply them as appropriate more often than not

A futuristic car drives from frame left to frame right; 4K; cinematic; dolly out, camera pans left to right.

FIGURE 2. VGM output, where the prompt asked for a moving futuristic car tracked by combined camera motions.

will lead to substandard results.[13] Applying these rules to craft VGM output is possible through prompt engineering techniques, discussed next.

Prompt Engineering Techniques for Video Generation

By applying the cinematography principles discussed, one can craft effective prompts for VGMs. An initial choice that must be made for VGM prompt engineering is to select the prompt input type or mode: (1) another video; (2) an image; and (3) text. Depending on the VGM, these input types can be used in combination or on their own. Videos as an input type typically are used in editing use cases, as discussed below, such as adding or removing objects from video frames, changing the overall aesthetic of a clip, or adding a specific effect.

Images can provide the model with a starting point, for example, to ensure the VGM adheres to a particular character or object design. For example, a piece of concept art depicting a character can be supplied to a VGM as part of a combined prompt, including both the concept art image and a text prompt that describes how to "animate" the character. Another use of image inputs is to create simple ads. Amazon Ads allows advertisers to make a video ad with one click simply by supplying an input product image. For example, an image of a coffee cup on a beach could be animated into a video showing steam rising from the cup, with waves rolling onto the beach in the background.[14] Image and video inputs can be considered inverse operations: image input sets an aesthetic that the VGM animates, while video input sets an animation (motion) for which the VGM can set an aesthetic.

As a first example of a text-only prompt for a VGM, it is instructive to begin with a relatively simple prompt. In **Fig. 2** above, a simple one sentence prompt is followed by output from the Amazon Nova Reel model.

Despite its simplicity, this prompt incorporates combined camera motions and the left-to-right motion rule described earlier, resulting in an aesthetically interesting output. However, writing a longer prompt with greater complexity can achieve even more artistic control over the output. As of early 2025, model providers typically limit text prompts to lengths from hundreds to a few thousand characters. For a prompt length limit of 2,000 characters, the user's prompt may range from two to four paragraphs, enabling greater expressiveness at the cost of higher prompt complexity. For example, in the above prompt, the background could be specified in detail, along with more information about the car itself, rather than leaving these decisions to the VGM.

Given the profusion of cinematographic principles and the complexities of their interactions, how does one start writing a more complex VGM prompt? One suggestion is to divide the prompt into sections for scene, subject, and camera movement.[15] An example of this template plus an actual prompt is the following:

[camera movement]: [establishing scene]. [additional details].

Establishing wide-angle shot with deep focus: A white toy poodle with light apricot colored ears is standing in the middle ground of a meadow full of flowers. The sky is bright blue with a few wispy white clouds. The poodle begins running from left to right in the frame. As the poodle runs, the camera pans right to follow the poodle while also zooming in towards the poodle, using a realistic documentary style with the camera maintaining deep focus.

The above prompt follows the template while incorporating multiple elements and rules of cinematography. However, there is no certainty that the VGM will follow all the directions in a prompt. All GenAI models operate as stochastic processes, with output reflecting an element of randomness. For example, when the above prompt was used with the Gen3-Alpha model from Runway AI, the model generally followed the prompt. However, the model insisted on applying a depth of field that kept only the poodle in focus, blurring the foreground and background, as shown in **Fig. 3.** This occurred despite the instruction to maintain deep focus (no blurring anywhere in the frame).

Due to the stochastic nature of VGMs, there is no guarantee that the output will be usable or that the same prompt will produce similar output when reused. **Figure 4** shows output from the same model using the same prompt as pre-

FIGURE 3. VGM output where most instructions were followed (e.g., poodle in middle ground runs left to right).

viously discussed. This second output follows the prompt much less closely than the first output. While the poodle does start in the middle ground as requested, the left-to-right motion is less perceptible. Also, as in the prior example, the instruction to maintain deep focus is disregarded. Sometimes, users can overcome such issues by rewriting the prompt using different keywords or phrasing. In any event, at least in the short term, working with VGMs requires extensive experimentation and sifting through a large volume of output to find clips that best reflect the user's artistic vision.

Prompt engineering for VGMs also has to take into account the time dimension. With most types of models other than VGMs, the user does not need to be concerned with the time dimension. There is only a single unchanging prompt to produce the entire result. Several techniques have been applied to make VGM output more dynamic while retaining continuity and reducing the need to edit multiple clips together. "Prompt travel" is a technique that specifies how a prompt varies by frame number or another specific time indicator. "Chaining" is another technique that, by contrast, specifies multiple prompts at one time without specific time indicators for changes, leaving the VGM to decide on the best transition points.

Another VGM prompt engineering technique is output previews. VGM output generation, or inference, can be costly and slow. As discussed earlier, this is a stochastic process; there is no guarantee that the output will be usable given a VGM's inherent randomness. Requesting an output preview gives a user a peek into what the model might produce for a given prompt without the cost and time investment of inference for a full clip. For example, an output preview feature might enable users to see the first few frames of multiple mini-inferred clips. Based on the results, the user can then judge whether it is better to proceed to full inference or modify the prompt first.

Post-Production with VGMs and other GenAI Models

Besides generating "footage" for production tasks and assisting with previsualization tasks in pre-production, such as aiding the creation of concept art, storyboards, and animatics, VGMs and other GenAI models have many different kinds of applications. Regarding post-production, media supply chain, and archive workloads, a nonexclusive list of tasks where GenAI models can assist includes the following:

- Editing
- Visual effects
- Video search for media archives
- Video summarization for short-form content creation
- "Vubbing" for localization

In post-production, editing is a central task that can be expedited with GenAI. For example, an editing VGM that accepts a video as input can add objects to a scene and animate them. VGMs used for editing typically have two primary modes: inpainting and outpainting. Using inpainting, a user can remove and replace objects and people in scenes or even animate a previously static object. It is also possible to remove, blur, or replace the background. Outpainting, by contrast, allows the user to extend the imagery of a video frame by generating new, coherent content for an area that originally was out of frame. Besides VGMs, other kinds of FMs can also assist with editing tasks such as color grading, adjusting depth of field, or converting a clip to super-slow motion.

One of the most important applications of VGMs and other FMs is in the domain of visual effects (VFX). A traditional VFX workflow may start with a hand-drawn concept sketch of an object or character, followed by 3D modeling and texturing and then matte painting to place the object or character in an environment. This work may take a month or longer, with limited ability to quickly edit and iterate. By contrast, a GenAI-based workflow would use one or more image generation FMs to create a sketch, apply textures, and place the object or character in an environment. It is easy to iterate in minutes at each stage of this workflow simply by changing the prompts. In addition to enhancing the creative process, this GenAI-based workflow can reduce the overall workflow time from over a month to less than an hour. Concept art can also serve as input images for a VGM, enabling video previsualization of scenes.

GenAI also enables the reuse of many kinds of media archives. For these use cases, specialized models other than VGMs may be applied. For example, GenAI-based upscaling may be used to reach new audiences for content originally produced in lower resolution formats. From a monetary value standpoint, localization with GenAI enables faster and more cost-efficient access for content to the multi-billion-dollar international film and TV distribution market. Specialized GenAI models can assist with "vubbing" (visual dubbing), which involves using GenAI to lip sync actors' faces to new lines. These may be dubbed foreign language lines translated from the original or entirely new lines in the original language requested for creative reasons. Depending on the use case, these lines can be delivered either by human voice actors or generated by audio FMs.

Other forms of archive monetization and reuse are enabled by applying GenAI to generate metadata for video searches. Video understanding models (VUMs) can be applied to create a text description of what is happening in a video frame, which then can be distilled into searchable metadata

Figure 4. Same prompt as **Fig. 3**, but key instructions disregarded in the model output (e.g. limited left to right motion).

based on the VUM output. This metadata can be used to organize video clips in bins and automate searching for highly relevant clips during post-production. Unlike previous simpler AI solutions that create metadata based on input such as AI-based audio dialog transcription and celebrity and object detection, a VUM can understand the context of each video frame. For example, a VUM called VideoLLM can handle multiple video understanding tasks, including action detection, segmentation, anticipation, captioning, and highlight detection.[16] Besides enhancing the creative process and streamlining workflows, a GenAI-based video search solution can be further applied to reuse archival content when combined with video summarization. GenAI-based video summarization can efficiently summarize long-form content to either create short-form content that is more easily appreciated by audiences and monetized, or to create highlight reels, trailers and the like for marketing purposes.

Conclusion

VGMs are the most promising GenAI models for enhancing the creativity and efficiency of video production. However, unlike other kinds of GenAI models, which do not require users to have significant domain expertise, optimal usage of VGMs requires users to have some understanding of the visual language of cinematography. Applying cinematographic concepts in writing prompts for VGMs makes it possible to craft aesthetically pleasing VGM output suitable for a wide range of pre-production, production, and post-production video workflows. In addition to using VGMs to generate and transform video frames, video production can also be enhanced by other kinds of GenAI models that can assist with tasks related to processing and monetizing media archives via localization and other forms of reuse. Looking to the future, a key trend will be "multimodal-to-multimodal" models where the input may be audio, images, text, and/or video, and the output a combination of the input types. For example, such a model could generate a video with synchronized audio (e.g., ambient sound, speech). This raises the question: Will VGMs be completely subsumed within multimodal-to-multimodal models? The answer likely is no—a media company may wish to record or generate audio separately from video for many reasons. Also, a review of the evolution of existing model types suggests that there will be a place for specialized, lighter weight models that excel at specific tasks as alternatives to larger, generalized models that may be more expensive. Accordingly, in the near term, VGMs appear to have a bright future as an aid in video production.

References

1. T. Maglio, "Sony Will Use AI to Cut Film Costs, Says CEO Tony Vinciquerra," *IndieWire*, 30 May 2024. [Online]. Available: https://www.indiewire.com/news/breaking-news/sony-pictures-will-cut-film-costs-using-ai-1235010605/. [Accessed: 3 Aug. 2024]
2. N. Magoon, et al., "Tech in Content Production: Will AI Kill the Video Star?," *Bain & Company*, 28 Sept 2023. [Online]. Available: https://www.bain.com/insights/tech-in-content-production-will-AI-kill-the-video-star/. [Accessed: 3 Aug. 2024]
3. B. Rabowsky, "Applications of Generative AI to Media," *SMPTE Mot. Imag. J., 132 (8)*: 53-57, Sep. 2023. DOI: https://www.doi.org/10.5594/JMI.2023.3297238
4. A. Still, "Bringing generative AI to video editing workflows in Adobe Premiere Pro," *Adobe Blog*, 15 Apr. 2024. [Online]. Available: https://blog.adobe.com/en/publish/2024/04/15/bringing-gen-ai-to-video-editing-workflows-adobe-premiere-pro. [Accessed: 4 Aug. 2024]
5. M. Seymour, "Actually Using SORA," *fxguide*, 14 Apr. 2024. [Online]. Available: https://www.fxguide.com/fxfeatured/actually-using-sora/. [Accessed: 4 Aug. 2024]
6. M. Seymour, "1st SORA music video," *fxguide*, 8 May 2024. [Online]. Available: https://www.fxguide.com/fxfeatured/1st-sora-music-video-how-sora-is-evolving-guessing-possible-pricing/. [Accessed: 4 Aug. 2024]
7. E. David, "Toys R Us unveils first commercial made with OpenAI's Sora," *VentureBeat*, 25 Jun. 2024. [Online]. Available: https://venturebeat.com/ai/toys-r-us-unveils-first-commercial-made-with-openais-sora/. [Accessed: 11 Aug. 2024]
8. Y. Zhang et al., "ControlVideo: Training-free Controllable Text-to-Video Generation," arXiv:2305.13077, May 2023 [Online]. Available: https://arxiv.org/abs/2305.13077. [Accessed: 14 Jul. 2024]
9. A. Blattmann *et al.*, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," arXiv:2311.15127, Nov. 2023 [Online]. Available: https://*arxiv.org/abs/2311.15127*. [Accessed: 10 Aug. 2024]
10. A. Schomer, "How Much Should AI Giants Pay Hollywood?" *Variety*, 21 Aug. 2024. [Online]. Available: https://variety.com/vip/ai-licensing-training-studios-hollywood-content-film-tv-1236112960/. [Accessed: 22 Aug. 2024]
11. A. Weprin, "Lionsgate Inks Deal With AI Firm to Mine Its Massive Film and TV Library," *The Hollywood Reporter*, 18 Sep. 2024. [Online]. Available: https://www.hollywoodreporter.com/business/business-news/lionsgate-deal-ai-firm-runway-1236005554/. [Accessed: 24 Sep. 2024]
12. J. Alton, *Painting with Light*, Macmillan: New York, NY, 1949.
13. B. Brown, *Cinematography: Theory and Practice*, Focal Press: New York, NY, 2002.
14. Amazon.com, Inc., "Amazon Ads launches a new AI Video generator—here's your first look at the beta," [Online]. Available: https://www.aboutamazon.com/news/innovation-at-amazon/amazon-ads-generative-ai-video-generator-advertisers. [Accessed: 24 Sep. 2024]
15. Runway AI, Inc., "Gen-3 Alpha Prompting Guide," [Online]. Available: https://help.runwayml.com/hc/en-us/articles/30586818553107-Gen-3-Alpha-Prompting-Guide. [Accessed: 12 Aug. 2024]
16. G. Chen *et al.*, "VideoLLM: Modeling Video Sequence with Large Language Models," *arXiv:2305.13292*, May 2023 [Online]. Available: https://arxiv.org/abs/2305.13292. [Accessed: 17 Aug. 2024]

About the Author

Brent Rabowsky has worked in AI at Amazon/AWS for the past ten years, managing a team of AI specialists. He has spoken at SMPTE conferences and written about AI for several journals and publications. He has also served as the technical editor for multiple AI books for O'Reilly Media and Packt Publishing.