Exploring Machine Vision through Generative AI Video

Yu-Shien Yang

OCAD University, Toronto

ABSTRACT

This paper examines recent advancements in generative text-to-video AI models, with a focus on Meta's *Make-A-Video* and Google's *Video Diffusion Models (VDM)*. Both approaches tackle key challenges in generating coherent video content from text prompts, such as limited paired text-video datasets and maintaining temporal consistency. The paper also explores cutting-edge platforms like OpenAI's *Sora* and MiniMax's *Hailuo AI*, which have made significant strides in producing realistic, high-resolution videos. Despite these advances, persistent challenges remain, including semantic accuracy, nuanced temporal understanding, and ethical concerns.

Keywords

Text-to-Video Models, Video Diffusion Models, machine vision, video synthesis

1. Motivate

Text-to-video generation has emerged as a frontier in AI, promising to translate textual ideas into moving images. Building on the success of text-to-image (T2I) models like DALL-E and Stable Diffusion, researchers have been striving to create text-to-video (T2V) models that produce coherent video clips from prompts (Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, Yu-Gang Jiang, 2024). The motivation is clear: video is a rich medium capturing motion and temporal dynamics that single images cannot. Achieving this would be a milestone for generative AI, enabling new forms of creative expression, rapid content creation, and simulation of the physical world in motion (OpenAI, 2024). For instance, Meta AI's *Make-A-Video* and Google's *Video Diffusion Models* (VDM) both tackle the challenge of generating temporally coherent, high-fidelity videos via deep learning. These works aim to overcome the scarcity of paired text–video data and the technical difficulty of modeling realistic motion, while leveraging advances in image diffusion models.

At the same time, the rise of machine-generated video invites broader reflection on machine vision and AI aesthetics. As Daniel Chávez Heras argues, "machine vision puts pressure on existing epistemic modalities, as it can be at once analytical and generative (Heras, 2024)". In other words, AI systems not only analyze visual data but increasingly create it, blurring the line between interpreting the world and imagining it. The notion of computers "seeing" and now "dreaming" videos for us raises fundamental questions: What does it mean for a cinematic image to be synthesized by an algorithm? How does AI-generated motion reconfigure our relationship as spectators to what is seen? The vast majority of digital video online already flows through algorithmic pipelines unseen by humans, and now generative models can produce hallucinated moving images where motion is algorithmically generated instead of captured. This change in visual media forms the backdrop and motivation for examining the latest T2V research. By comparing two papers – "*Make-A-Video: Text-to-Video Generation without Text-Video Data*" and "*Video Diffusion Models*," we can better understand the state-of-the-art methods and the goals driving their development. In addition, drawing on insights from *"Cinema and Machine Vision"* helps place these technical advances within a critical framework of AI aesthetics and spectatorship.

2. IDEATE

2.1 Core Approaches

The two papers have different purposes yet related approaches to text-to-video generation. Make-A-Video introduces a method to create videos from text without requiring paired text-video training data. Instead, it leverages pre-trained text-to-image models and unlabeled video. The key idea is to learn what the world looks like from images and how it moves from videos. Specifically, the authors start with a strong T2I diffusion model and extend it into the temporal domain by adding new spatial-temporal layers and attention modules that learn motion dynamics from video-only data. This transfer learning drastically reduces the need to train a video model from scratch. The model is further refined through a pipeline of frame interpolation and super-resolution steps that increase frame rate and output resolution. By decomposing the problem. *Make-A-Video* effectively marries the strengths of image generation (diverse content and appearance) with video understanding (temporal coherence). In essence, it accelerates T2V training by instantaneously transferring the knowledge from a previously trained T2I network to a new T2V one (Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman, 2022).

VDM, on the other hand, takes a different approach by extending the diffusion probabilistic models that revolutionized image generation directly into video. They show that minimal changes to the standard image diffusion architecture can yield high-quality video. In their approach, a 3D U-Net (a neural network with 3D convolutions) is trained to denoise video data, generating a fixed number of frames per sample. Crucially, the model is trained on both image and video datasets simultaneously, treating a single image as a "video" of one frame, which stabilizes training and improves quality.

To generate longer videos than the fixed frame count it was trained on, *VDM* introduces a conditional sampling technique. Essentially, the model can be fed its own output as partial conditioning to extend a sequence in time (analogous to how language models generate text iteratively). This allows for temporal extrapolation and spatial-temporal super-resolution without needing an enormous one-shot model. The simplicity of this design was intentional: the authors demonstrate that even little modification other than straightforward architectural changes to accommodate video data is sufficient to produce coherent videos (Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet, 2022). This is a testament to the generality of the

diffusion model framework (Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, Karsten Kreis, 2023).

2.2 Goals and Motivations

Both works share the high-level goal of pushing generative modeling into the video domain, but with different emphases. *Make-A-Video* explicitly seeks to sidestep the scarcity of text-labeled video data. Collecting huge video datasets with descriptions (analogous to LAION-5B for images) is infeasible; thus, the goal is to reuse abundant image-text data and unsupervised video to achieve T2V. The authors highlight how it would be wasteful to train T2V models from scratch when there already exist models that can generate images. The motivation is efficiency and immediacy, that is, unlock text-to-video generation now, by standing on the shoulders of image models.

In contrast, *VDM* are motivated by validating diffusion as a paradigm for video. Diffusion models had already proven their worth for images and audio. The authors aimed to show they can handle the greater complexity of video while achieving high quality metrics on standard benchmarks. They explicitly sought to demonstrate progress on multiple fronts: unconditional video generation (random videos), video prediction (continuing a given sequence), and text-conditioned generation. This broad goal reflects an interest in diffusion as a general solution for generative video, leveraging its stability and mode coverage.

2.3 Results

Both papers report strong results. *Make-A-Video* delivered striking proof-of-concept clips that captured complex motions and scenarios described only by text. The paper showcases clips like "*a dog wearing a superhero outfit with red cape flying through the sky*," where the output shows the dog realistically flying with its cape fluttering. The motions are smooth and logical, such as a camera pan across a static scene or an animal running. The authors claim that in all evaluated aspects, spatial and temporal resolution, fidelity to the text prompt, and visual. This claim is supported by both user studies and quantitative metrics. For instance, *Make-A-Video* outputs were reportedly preferred by users over previous methods and showed better alignment with text descriptions (Meta, 2023).

Technically, a significant achievement is that the model can learn motion patterns from unlabeled videos and apply them appropriately, suggesting the emergence of a learned "world dynamics prior." For example, given a prompt about books on a table with sunlight, the model kept the books static but made the lighting change with subtle camera movement, mimicking real-world behavior. This indicates a degree of common sense learned from video data.

VDM also reports compelling outcomes. On standard benchmarks for short video generation (evaluated on datasets like UCF-101 or Sky Time-lapse), their diffusion approach set new records. For instance, for unconditional video generation, the model achieved the lowest Fréchet Video Distance (FVD) scores to date, indicating the samples were very close to real video distributions. On a video prediction task, it likewise achieved impressive performance. These quantitative gains suggest that diffusion models can surpass GAN-based or autoregressive models that were common in earlier works.

Figure 1
Meta AI's incredible result of a video clip. Prompt: A dog wearing a Superhero outfit with red cape flying through the sky (Meta, 2023)

Additionally, *VDM* presented early results on a large text-conditioned video generation task, demonstrating that diffusion models can be conditioned on text embeddings to produce relevant video content. While those text-to-video results were described as "promising" rather than flawless (due to likely low resolution or short duration), it was a milestone showing the feasibility of scaling up diffusion for multimodal generation. Qualitatively, the results appeared as somewhat blurry or simplistic videos that nevertheless followed the prompt. Notably, the authors' conditional sampling method enabled the generation of longer sequences than prior diffusion-based attempts, pointing toward breaking the length barrier through iterative generation. Overall, *VDM*'s results reinforced the diffusion model's potential, achieving strong Inception Score (IS) and FVD metrics and establishing a baseline for future improvements.

2.4 Limitations

Despite their advances, each paper acknowledges important limitations.

Make-A-Video's reliance on image-based learning means it cannot capture video-specific nuances that have no image analogue. The authors note that without true text-video pairs, the model may fail on prompts where the action or temporal aspect is implied only through motion. For example, understanding "a person waving left-to-right" versus "right-to-left" requires associating text with directionality, something their model cannot learn from static images alone. Thus, temporal ambiguity in language remains a challenge.

Another limitation is the length and complexity of generated videos. *Make-A-Video* was demonstrated on short clips with a single scene. Generating longer narratives with multiple events or scene changes is left for future work. There are also telltale artifacts in some outputs. Common issues include slight flickering or the subject's appearance drifting over time. While the motion is generally fluid,

maintaining perfect identity consistency is hard without explicit temporal constraints.

VDM also faced their own constraints. Training a 3D U-Net on video is extremely memory-intensive, which likely forced the authors to use relatively low resolutions and frame counts. This limits the visual fidelity and duration of what the model can produce. Indeed, the text-conditioned results in *VDM* were preliminary; by the authors' admission, they chose not to release the model or any public demo, partly due to concerns about ethical and safety implications, but possibly also because the model was not yet at a stage for wide use.

It is worth noting that in the paper, the authors caution that like any generative model, it can produce biased or harmful content and "reflect the biases of its training datasets". From a technical perspective, diffusion-based video generation at scale was – and is – computationally heavy and somewhat slow at inference (requiring many denoising steps per frame). The *VDM* paper's focus on methodology meant that aspects like higher resolution synthesis or fine detail were left for subsequent improvements.

In summary, *Make-A-Video* pushed the envelope by harnessing existing resources but must overcome semantic gaps and limited video length, whereas *VDM* proved a concept at some cost to detail and with acknowledged societal risks.

Before moving to EXPLORE section, it's worth noting how these innovations tie into theoretical implications. The aesthetic quality of AI-generated videos and the experience of spectatorship in viewing them are uncharted territory. Chávez Heras suggests that as AI begins to produce moving images, we witness a "radical re-imagination of the visual" where the machine is both creator and viewer (Heras, 2024). The outputs of *Make-A-Video* or *VDM* can be seen as a form of "datamatic time" – a term Chávez Heras uses to describe how generative techniques dissolve conventional cinematic temporality. For instance, a generative model can conjure an endless slow-motion sunset or a perfectly looped sequence that has no real-world counterpart in duration. This challenges our notions of filmic time and causality. Moreover, the slight surrealness or "rational illusions"in these videos (a term borrowed from Paul Virilio to describe machine-generated visuals) might become a new aesthetic: one where viewers are aware that what they see is a product of computation, not reality, potentially altering their mode of spectatorship. We might become "willing subjects" to the AI's time and vision, as in classical film theory the spectator yields to the film's programmed sequence. Here, the sequence is generated on the fly by an algorithm – a new kind of contract between viewer and image. These considerations illustrate why, beyond technical achievement, researchers and critics are carefully scrutinizing what generative video means for the future of cinema and visual culture.

3. EXPLORE

The paper now will explore two current video generation models that have emerged from the latest wave of this research: OpenAI's *Sora* and MiniMax's *Hailuo AI*. Both systems became available in late 2024 and are at the cutting edge of applying ideas from prior academic models into real interactive tools. This paper investigates their capabilities, example outputs, and how their approaches align or diverge from the earlier research.

3.1 Text-to-Video (T2V)

OpenAI Sora

Sora is a text-to-video model developed by OpenAI and released to users in late 2024. Technically, Sora can be seen as a successor to concepts from both *Make-A-Video* and *VDM*. It is described as a "diffusion transformer" model, meaning it uses a transformer-based neural network within a diffusion generative process. According to OpenAI's report, Sora represents video in a compressed latent space of "spacetime patches (OpenAI, Video generation models as world simulators, 2024)".

This involves first encoding raw video into lower-dimensional latent codes, then chopping that latent into patches that become tokens for a transformer. By doing so, Sora achieves remarkable flexibility, it can generate videos of variable resolution, aspect ratio, and length without needing separate models. This is a major leap over prior academic models that were fixed to 4 seconds at 256px. Sora can also generate single images as a special case (a 1-frame "video"), underscoring its generalist design. In practice, Sora's user experience is integrated with ChatGPT. A user provides a detailed prompt, possibly along with an initial or reference image/video, and Sora will produce a short video clip that tries to match the description. Early access users and red-team testers were able to generate clips demonstrating a wide range of scenes – from cinematic cityscapes to fantastical creatures (Leffer, 2024).

To test the capabilities of Sora T2V, I used the same prompt as Meta's *Make-A-Video*: "*A dog wearing a superhero outfit with red cape flying through the sky.*" The result showed a dog dressed in a superhero costume flying through a sky of blue and white clouds. The dog wore a tight-fitting blue suit with a golden emblem on its chest, yellow wrist guards, and a red cape flowing behind - basically a canine version of a superman. Its facial expression was serious and heroic, adding a mix of humor and drama to the scene.

Stylistically, the video generated by Sora had a hyper-realistic animated look, with richer details and more vivid colors, somewhat reminiscent of characters from Pixar or Illumination films. Compared to MetaAI's version, Sora's depiction of the superhero dog was noticeably more anthropomorphic, in other words, the character looked more human than dog-like, similar to the anthropomorphic animal designs seen in the animation *BoJack Horseman* (Bob-Waksberg, 2014). In terms of movement and facial expressions, Sora's output had greater dramatic tension and more dynamic, expressive motion than *Make-A-Video*.

However, the video's color saturation was overly intense, which could pose significant challenges in commercial post-production workflows. To address this, I modified the prompt to: "*A dog wearing a superhero outfit with red cape flying through the sky, low contrast and saturation, LOG color space*" in an attempt to generate footage resembling LOG or RAW color space.

Figure 2
Sora-generated video, using the same prompt as Make-A-Video. https://sora.com/g/gen_01jqjrkmewewqrx5bvezfhrkgp

Despite this adjustment, it became clear that Sora still has room for improvement in terms of recognizing professional file formats and color management workflows. This remains one of the major challenges in AI-generated videos. In professional filmmaking, LOG or RAW formats are standard, as they allow for consistent color grading across different shots, something not achievable with Rec. 709 color space, which is rarely used during professional filmmaking.

Interestingly, in the second output, the dog was portrayed as a more realistic animal rather than an anthropomorphized figure. Despite the prompt only changing the color space and not the character description. This variation in interpretation reveals a potential inconsistency in how the model processes textual prompts, which could be an important area for further research.

Figure 3
Sora-generated video, using the prompt to change the color space. https://sora.com/g/gen_01jqn3bv1ff52b29yvq9y47qt7

Sora also supports video extension, given an existing short video, it can continue it or change its style. This implements the idea of autoregressive extension from *VDM* in a user-facing way. For instance, one could upload a 5-second clip of a landscape and ask Sora to "continue this scene for another 5 seconds with similar style," and it will produce a seamless continuation. The overall experience has been likened to using DALL-E but for motion. OpenAI also enforces content safeguards in Sora, filtering prompts about violence, sexual material, identifiable people, etc., in line with their safety policies. This means user requests that violate those rules will be refused or altered.

Hailuo AI

Hailuo AI is another new text-to-video generator, introduced by the Chinese startup MiniMax. Launched to the public in September 2024, Hailuo quickly gained popularity, especially among creators in online communities, because of its ability to produce vivid and fluid videos up to around 6 seconds in length. The name "*Hailuo*" (海螺, meaning conch shell in Chinese) suggests a tool that can conjure rich audio-visual experiences from a small prompt.

What sets *Hailuo* apart, according to reports, is the quality of motion and realism in its outputs. Early adopters noted that *Hailuo* produced human movements that were "much more fluid and lifelike" than those from some U.S. models (Franzen, 2024). The videos tend to have coherent backgrounds and subjects that move naturally with fewer distortions. An example shared by the company includes a generated video of a cat hopping off a table, where the motion of the cat and the camera's slight shake felt convincingly real. Another example from a demo reel shows a person running on a beach at sunset, with the waves moving and the person's silhouette consistent, a scene that could pass for actual footage at a casual glance (Franzen, 2024). These anecdotal examples line up with the claim that *Hailuo* outputs "ultra realistic" videos. The system likely employs a diffusion or transformer-based architecture similar to others (though details have not been fully published, being a commercial product). It might incorporate large Vision/Language models for interpreting prompts and a video diffusion backbone for generation, possibly enhanced by motion-specific training.

During my exploration, I also used the same prompt as I did with *Make-A-Video* and *Sora*: "*A dog wearing a superhero outfit with red cape flying through the sky*." In terms of generation speed, *Hailuo* was noticeably slower. Sora could generate a 10-second 720p video in about two minutes, whereas Hailuo took 26 minutes to produce a 6-second 720p clip.

Moreover, with the same prompt, the dog's flying motion in *Hailuo*'s output appeared quite stiff, as if only keyframes were used to set values along the X-axis while the background was uniformly scaled, resulting in a motion that breaks the logic of real-world physics and cinematic camera movement. The dog also lacked any natural flying gestures or expressive facial changes. Visually, the colors were even more garish and flashy compared to *Sora*'s output.

Figure 4
Hailou- generated video, using the same prompt as Make-A-Video and Sora. https://hailuoai.video/share/ai-video/xKOvz3jGrvwm

To test *Hailuo*'s understanding of professional video formats and color management, I used the same adjusted prompt: "*A dog wearing a superhero outfit with red cape flying through the sky, low contrast and saturation, LOG color space*." The results showed similar issues to Sora's, indicating that Hailuo also struggles with interpreting professional video formats and color space requirements.

Figure 5
Hailou-generated video, using the prompt to change the color space. https://hailuoai.video/share/ai-video/dzEYgw3Djody

However, one strength of *Hailuo* was its consistency in the character's movements and visual style across videos using the same prompt. However, if the goal is to generate dynamic, expressive movements and facial expressions like those seen in *Sora*'s output, the prompts may need to be more detailed and specific.

To test this hypothesis, I modified the prompt to: "*A dog wearing a detailed superhero outfit with a red cape, powerfully soaring through the sky in a dynamic flying pose, with intense, dramatic facial expression and determined eyes, against an epic backdrop of clouds and cinematic lighting*." Despite the refined prompt, the resulting animation still lacked ideal motion, the dog's limbs showed almost no movement apart from the cape fluttering in the wind. However, the enhanced environmental and lighting descriptions did help strengthen the video's dramatic tone.

This suggests that in T2V generation, the structure and specificity of the prompt still play a crucial role in determining the quality of the output.

Figure 6
Hailou-generated video, using the refined prompt to enhance the character and camera movements. https://hailuoai.video/share/ai-video/yKGr1kv3y3EK

3.2 Image-to-Video (I2V)

Although the *Make-A-Video* and *VDM* papers did not delve deeply into the Image-to-Video function, this feature could actually be more practical for commercial applications. It allows clients to specify their product and model, and generate a video from just a single photo, significantly reducing the time, cost, and communication efforts required for production.

Given its potential, this paper will also test and compare the performance of OpenAI's *Sora* and *Hailuo* in this area. To begin, I extracted stills from videos I had previously produced for commercial brands to use as test samples.

Figure 7
A still photo from videos I had previously produced for commercial brands

When no prompt was provided, *Sora* generated a video in which the characters became distorted. The Asian woman on the right turned into a woman with darker skin, and the man on the left appeared with different clothing. In contrast, *Hailuo* did not exhibit such issues, the characters remained consistent (though their movements were minimal and somewhat unnatural), and the camera slowly zoomed in.

From these results, we can see that when dealing with more complex compositions and character designs, AI-generated videos still tend to show inconsistencies, such as changes in appearance or clothing. While *Hailuo* demonstrated fewer of these issues, its character movements were noticeably stiff, indicating that current AI systems still lack a full understanding of the logic of the physical world.

Figure 8
Sora's image-to-video test result.
https://sora.com/g/gen_01jqqnbvhwe8sbbfce2q74ffkx

This suggests that in addition to relying on prompts, creators planning to use AI tools for video generation should aim to keep visual compositions as simple as possible to ensure smoother post-production. Moreover, AI tools can serve as a helpful pre-visualization resource, allowing creators to simulate and plan scenes in advance, streamlining the overall filming and editing process.

Figure 9
Hailou's image-to-video test result.
https://hailuoai.video/share/ai-video/dzEoK7J7pL0k

3.3 The Capacity to Comprehend Cultural and Stylistic Nuances

One of the challenges in using AI for creative or commercial purposes is its ability to understand culture, values, and style. Since most brand campaigns are theme-driven each season, AI tools need to accurately capture the client's desired visual direction, and ideally, allow for corrections when the output misses the mark.

To evaluate how well AI can interpret cultural and stylistic cues, the following section will use the "Millennium aesthetic" that has recently gained popularity in Japan as a case study to test the cultural and stylistic comprehension of OpenAI's *Sora* and *Hailuo*.

Using the following prompt, I tested *Sora* and *Hailuo* individually to evaluate their comprehension of cultural context." *A nostalgic 1990s Japanese millennium-era TV commercial featuring a young Japanese female idol as the spokesperson. She is promoting a stylish CD Walkman. The scene is bright and colorful, with retro aesthetics, VHS-style visual texture, and dynamic camera movements typical of late 90s Japanese commercials. The idol has long black hair, wears a fashionable school uniform with a modern twist, and smiles charmingly as she demonstrates the CD player in various everyday scenes, walking in the city, sitting in a classroom, relaxing at a café. Japanese text flashes on the screen with bold fonts, upbeat J-pop music plays in the background, and the ad ends with the idol giving a wink to the camera and the product slogan echoing with nostalgic synth effects. The mood is cheerful, youthful, and filled with 90s optimism."*

Figure 10
Sora's video cultural understanding capability test result.
https://sora.com/g/gen_01jqqws2n7fd4vxmyv5a6ppqnf

Figure 11
Hailou's video cultural understanding capability test result.
https://hailuoai.video/share/ai-video/7A0EwgqwypyN

3.4 Alignment with Research Approaches

Sora and *Hailuo* both embody the fruits of ideas seeded by works like *Make-A-Video* and *VDM*, yet they also illustrate different

philosophies in reaching state-of-the-art performance. *Sora*'s design very much aligns with the *VDM* paradigm, it is explicitly a diffusion model for video with a twist of transformer architecture. Just as the article of *Make-A-Video* mentioned. *Sora* was trained on "publicly available videos as well as licensed videos" plus an image dataset (Leffer, 2024). In spirit, *Sora* extends the *VDM* idea of using a single model for variable-length video: the patch-based approach allows arbitrary sizes, which is an elegant solution to the fixed-size limitation in the original *VDM* implementation. *Sora* also resonates with *Make-A-Video* in that it leverages existing image generative technology: notably, OpenAI built Sora after releasing DALL-E 3, and indeed Sora's technology is described as an "adaptation of the technology behind DALL-E 3". This implies that the transformer and decoder used in DALL-E 3's image generation were extended to handle video patches (OpenAI, Sora System Card, 2024) – conceptually like how *Make-A-Video* extended an image model's weights to video.

Hailuo's approach, while not published academically, can be inferred to align with these trends as well. The fluidity of motion in *Hailuo*'s outputs suggests a model strongly focused on temporal consistency, a key concern of both papers we reviewed. It likely uses a diffusion model or a GAN with heavy temporal constraints. Given the timeline, *Hailuo* may have been influenced by the *Make-A-Video* paper or similar work (Meta's model was publicized in late 2022) in terms of leveraging existing image model knowledge and adding motion through separate modules. It could also be using an architecture akin to Runway Gen-2 (which uses latent diffusion for video).

Both models, importantly, incorporate the safety and ethical considerations flagged by prior research. *Sora* refuses certain prompts (like those involving violence or specific people) and watermarks outputs with metadata to indicate AI origin. Similarly, *Hailuo*'s developers would likely have filters (especially given content regulations in China). Although Hailuo offers an option to generate content without a watermark, recent regulations from the Chinese government require all AI-generated content to include one (Sharwood, 2025). This suggests that as more negative use cases emerge, AI-generated videos may become increasingly subject to mandatory regulations. These practical measures show how turning research into products requires addressing the potential misuse highlighted in works like *VDM*.

4. TEST

Having explored *Sora* and *Hailuo*, I will now evaluate how these modern systems' behavior matches or diverges from the research claims made in *Make-A-Video* and *Video Diffusion Models*. In many ways, *Sora* and *Hailuo* are the realizations of those earlier research aspirations, so it is illuminating to see where theory meets practice and where unexpected differences arise.

4.1 Capability

One of the core claims of *Make-A-Video* was that leveraging image knowledge plus unlabeled video is a viable path to text-to-video generation without text-video pairs. *Sora*'s performance strongly affirms this claim. OpenAI did not publicly disclose the size of *Sora*'s training data, but it is known that they used a large amount of video (with AI-generated captions) in addition to images (Mauran, 2024). *Sora* demonstrates that a model can indeed learn to align text and video without explicit human-provided video captions: the "re-captioning" strategy provided the needed bridge (Richie Cotton, Matt Crabtree, 2024). This mirrors *Make-A-Video's* approach of using image captions video for motion, suggesting that *Sora*'s impressive results are a validation of that concept on a larger scale. *Hailuo*'s success likewise indicates that paired data scarcity is not a showstopper. MiniMax could build a great model presumably with a combination of existing resources, possibly leveraging public video and image-text data and clever model design . In fact, Hailuo's fluid motions fulfilling text prompts show that the model has learned text-motion mapping, likely through similar indirect means that *Make-A-Video* proposed.

Sora, being a latent diffusion transformer, achieves strikingly high fidelity – arguably beyond what the original *VDM* paper demonstrated. For example, where *VDM* might produce a low-res clip of an action with some blur, *Sora* can now produce near-photorealistic 720p or 1080p videos with consistent detail. This suggests that the scalability of diffusion models (one of their key advantages in images) carries over to video when sufficient resources are applied. *Hailuo*'s use of advanced diffusion or similar generative tech also underscores that diffusion has become the de facto standard approach for cutting-edge video AI.

Both *Sora* and *Hailuo* show evidence of strong temporal coherence, a central criterion in *VDM*'s evaluation. Videos from these models usually do not suffer from the severe flicker or disjointed frame problem that was common a few years ago; the motion is learned as a smooth function, exactly as diffusion models intended by treating video as a contiguous signal.

4.2 Temporal Understanding

However, certain research claims highlight ongoing challenges. *Make-A-Video*'s authors pointed out their model couldn't handle cases like distinguishing left vs. right in motion without paired data. Does *Sora* fix that? Maybe not.

Some users noted that *Sora* still struggles with some complex physics and logical consistency . Early users found that mirrored text (like signs) or left-right orientations can confuse it, resulting in inconsistent output on those specifics. This shows that even the state-of-the-art model inherits some limitations that were theoretically anticipated: certain nuanced spatiotemporal relations are hard to infer without explicit training (Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, Robert Geirhos, 2025). If a prompt says "person turns left," the model might not reliably know which direction left is from the camera's perspective, which is a very granular comprehension issue.

Hailuo's behavior in this regard is less documented but given its focus on visual realism over strict semantic accuracy, it might similarly not guarantee correctness for such fine-grained instructions (e.g., it can make a person wave, but if asked for a specific hand wave, it might not consistently get left vs. right hand correct). The test result in this paper also shows the limitation identified by research persists to a degree in these two products.

Figure 12
Sora's generated video showed an error in distinguishing between left and right hands. Prompt: A Japanese woman is waving her left hand to the audience

Figure 13
Hailuo has the same issue as Sora in not fully understanding real-world physics. Prompt: A Japanese woman is waving her left hand to the audience https://hailuoai.video/share/ai-video/38vmwGJdWWWA

Another aspect is video length and multi-scene generation. *Make-A-Video* saw multi-scene storytelling as future work. *Sora* has extended length significantly, but does that mean it can do a true multi-scene story? In tests, *Sora* can maintain coherence for tens of seconds, typically it works best with a single continuous scene. However, when I tested it with a script-like, multi-scene prompt, Sora seemed to successfully merge the two scenes together. That said, when the prompts become more complex, it seems to lack an internal mechanism for handling hard scene cuts or sustaining long narrative arcs, instead, it tends to favor continuous transitions.

This is in line with the research state, there isn't yet a clear solution for the model to know when to transition scenes or handle very long-range dependencies. *Hailuo*, being limited to around 6 seconds, avoids this question entirely by focusing on short clips. Users who want longer videos with *Hailuo* have to stitch multiple outputs manually. In essence, current models validate research progress in short-form video generation but have not fully conquered long-form video generation.

Figure 14
Sora successfully transitioned from the beach scene to the forest scene, but only limited to simple cut shots. https://sora.com/g/gen_01jqykcj3efvvvwc7s2ths03aa

Figure 15
Hailuo used AV club-style effects to transition between the two scenes, but due to the limited video length, the transition felt somewhat unnatural. https://hailuoai.video/share/ai-video/XX0ok2rWl8Y9

4.3 Single Frame Quality

The quality improvements claimed by *Make-A-Video* are evident. *Sora*'s videos are high resolution and often breathtaking in quality for AI-generated media, with dynamic range in lighting, correct perspective, and even emergent 3D consistency (OpenAI researchers noted Sora "figured out how to create 3D graphics from its dataset alone", e.g., it can show a subject from multiple angles smoothly (Levy, 2024)). This emergent spatial consistency is exactly what one hopes for in a model learning "world dynamics," it indicates the model has some internal representation of objects in space, not just flat sequences of frames (Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, Aditya Ramesh, 2024). This goes beyond what *Make-A-Video* demonstrated; it is a result of larger scale training and perhaps

architectural choices like transformers that encourage such representations.

Hailuo's coherence also speaks to these models indeed learning a quasi-3D understanding; videos of people rotating or cameras panning look correct in perspective. These match the aspirational claims that generative models can learn internal physics or world models from data.

Yet, limitations remain visible: for instance, *Sora* and *Hailuo* are both known to falter on fine text (like reading words on signs) and detailed faces up close. A prompt that requires reading or writing will likely produce gibberish text in the video. This mirrors the experience with image models and was indeed an issue noted indirectly in research (as part of general limitations and biases). It's not a focus of either paper, but it's a known shortcoming that persists – the models don't truly "understand" text as content within the image frames, so they can't reliably generate legible text. This is a point where spectatorship becomes interesting: a human viewer immediately spots these glitches, highlighting the difference between a video that is generatively plausible vs. one that is semantically grounded in all details.

4.4 Creative Implications

From a creative perspective, the models exhibit behaviors anticipated by theory: an "emergent grasp of cinematic grammar" was noted in *Sora*. Steven Levy observed *Sora* sometimes did unprompted shot changes, as if it learned when to cut or change camera angle for dramatic effect. This is a fascinating alignment with ideas from film studies and AI aesthetics: the AI is not just generating random movements, but it appears to be developing a sense for how to present a scene, reminiscent of the "patterns of cinematography" that one might teach or analyze in human-made films. This emergent property was not explicitly claimed in the papers, but it is a logical extension of training on large amounts of video. It resonates with Chávez Heras's notion of "key aesthetic and epistemic assumptions embodied in the practice of machine-seeing films," Here, the machine might be internalizing patterns of cinematography and replaying them. Thus, in testing *Sora* and *Hailuo*, I find that they not only meet many technical expectations set by prior research (regarding coherence, quality, leveraging existing data) but also start to surface new questions: Are these models inadvertently learning film-making conventions? Are they creative or just imitative? My observations suggest a bit of both, they are imitating patterns in data, but in doing so, they sometimes produce novel combinations (e.g., a fantasy creature in a documentary style shot) that feel creative.

In summary, the behavior of *Sora* and *Hailuo* largely confirms the research claims from 2022: high-quality short videos are achievable with diffusion/transfer learning methods, and key limitations (data biases, some semantic gaps, length constraints) are still active areas to work on. Where they diverge, the differences are quantitative (much higher resolution now, longer duration) and operational (safety mitigations, user interface), rather than fundamental conceptual breaks. We are seeing a steady evolution rather than a sharp break, these products are the next step along the path charted by *Make-A-Video* and *VDM*.

5. REFLECT

5.1 Critical Synthesis

The evolution of generative video models, from pioneering research to today's deployed systems, paints a picture of rapid progress coupled with expanding implications. In reflecting on my findings, a key theme emerges: generative video is moving from a technical demo towards a practical medium, and this transition brings both exhilaration and caution.

On one hand, models like *Sora* and *Hailuo* demonstrate unprecedented capabilities, they realize many of the aspirations that researchers had only a couple of years ago. We now have AI that can simulate the visual world in motion with surprising fidelity. This opens doors for creative professionals and everyday users alike. A filmmaker can create storyboard concepts by simply typing, a game designer can generate background scenes, an educator can create illustrative videos for a lesson without a camera. The democratization of video creation is a profound implication: much as text-to-image tools have enabled anyone to become an artist, text-to-video might enable anyone to become a filmmaker or animator at a basic level.

However, along with these opportunities come concerns that echo those voiced by scholars and commentators. The worry about misinformation and deepfakes is very real, if anyone can generate a clip of an event that never happened, society's already fragile trust in media could erode further (Metz, 2024). My exploration confirms that companies are aware of this. OpenAI's metadata tags and content filters are direct responses to these risks. Yet, as the tech spreads (with open-source efforts likely to accelerate), the onus will increasingly be on media literacy and detection techniques to keep up

There's also an economic and ethical angle: if AI video generation becomes mainstream, what happens to the jobs and crafts associated with video production? Some industry figures have paused major studio investments due to concern for AI's impact (Edwards, 2024). While others note that it will be a long time – if ever – before text-to-video threatens actual filmmaking, we should not underestimate how quickly "good enough" AI content can disrupt certain areas (e.g., advertising, stock footage company, basic entertainment, even news reenactments).

From a theoretical standpoint, this juncture is fascinating. We see a convergence of cinema and machine vision in a new way: AI models have become both producers of moving images and subjects that embody a way of seeing. A central question arises: what ought to be seen in the age of AI? These models tend to generate what they've seen in their training data – a reflection of our world as filtered through the internet's corpus of videos. This raises a concern about aesthetic homogenization: if everyone uses similar models trained on the same data, will we see a surge of videos that all look and feel somewhat the same, guided by the AI's learned aesthetic?

The spectatorship of AI-generated video may involve a loop where humans are watching machine-produced interpretations of human culture. Will we pick up on the subtle "machine vibe" in these videos? Already, some AI images have a telltale look (hyper-real lighting, slightly surreal coherence). In video, those tells might be camera movements that are too smooth or an absence of true narrative causality – as one early viewer quipped, it can look like "a bunch of images…stitched together" with odd inconsistencies

(Manovich, L., & Arielli, E., 2024). As the tech improves, these tells will diminish, but the philosophical gap remains: these videos are manifestations of statistical vision rather than lived experience or intentional cinematography. They are what some scholars might call "sightless vision" – images without a human observer at the point of creation (Heras, 2024).

This also points toward a future shaped by post-human vision. However, this isn't an entirely new domain. In fact, machine vision has been influencing our perception of reality ever since the invention of the camera, which gradually defining the coordinates and accuracy of the real world. Looking ahead, human vision, machine vision, and algorithmic vision will collectively shape the language and aesthetics of cinema.

5.2 Future Directions

Looking forward, research and development will likely focus on several fronts. One is improving semantic understanding and controllability: future models may incorporate stronger language understanding to better fulfill complex prompts and avoid errors like left-right confusion.

Another frontier is long-form content – enabling AI to generate a sequence of scenes or an entire short film with a coherent plot. This might require new architecture or hierarchical. Also, we can expect work on multimodal integration: adding audio synchronized with the video, since a truly compelling video often needs sound.

Additionally, it's worth noting that although AI-generated videos have made unprecedented progress since the release of *Make-A-Video* and *VDM*, whether in terms of quality, cultural understanding, or generation speed, there are still key challenges. As mentioned earlier, in professional video production, footage is typically captured in LOG or RAW formats to ensure a smooth post-production process. This is crucial for maintaining consistent color grading and visual style across shots and is a defining factor in achieving a "high-quality" cinematic look. Therefore, if AI-generated videos can overcome current limitations with file formats and color management, they could be truly viable for commercial use - whether in advertising or filmmaking.

5.3 Limitations and Broader Impacts

It's important to also acknowledge the enduring limitations and the contexts where these generative videos may fall short or pose problems. Bias in training data can lead to skewed outputs – perhaps the model is more likely to visualize certain professions as one gender or create content from a Western-centric perspective if that dominated the data (Heras, 2024). If generative videos become part of media production, these biases could propagate unless actively corrected.

Another limitation is the lack of true agency or intentionality in the AI. It has no understanding of why a video is interesting or what story is being told; it just tries to make it look right. This means it could inadvertently generate superficially plausible but meaningless sequences. For instance, it might create something that looks like a news report but is gibberish in content – a hollow shell. Spectators must be cautious not to assign undue meaning or truth-value to AI-generated footage.

There is also the psychological impact. As AI content becomes more prevalent, viewers might develop a general skepticism ("seeing is no longer believing") or conversely an indifference to whether visuals are real or AI-made. We might enter an era of hyperreality in media, where the origin of images is obscured and it matters less whether a clip was recorded or generated, as long as it's convincing or engaging. This ties back to the question of what cinema becomes when the "camera" can be an algorithm – possibly a new art form where human and AI co-create narratives. In reflecting on the overall journey from research to current models, one is struck by how interwoven the technical and cultural threads are. The technical triumph – machines creating moving imagery from text – forces a cultural and theoretical reckoning with the nature of visual truth, creativity, and the role of human skill.

Yet, history suggests that new technologies in art (photography, film, digital editing) have always caused initial alarm, only to become new tools in the artist's palette. It is likely the same will be true for generative video, rather than replacing human filmmakers, it will become a powerful tool for them. It may enable entirely new genres (perhaps AI-generated virtual reality experiences tailored to each viewer's prompts) or allow niche communities to produce content at scale (imagine fan-fiction videos generated for any scenario).

The broader implication is that storytelling itself could become democratized to an unprecedented degree, but also that we will need to educate society about how to critically evaluate visual media in this new age. Ultimately, the future of generative video will be what we collectively make of it – guided by research insights, shaped by creative experimentation, and hopefully grounded in thoughtful reflection on its impact.

6. AI TOOLS DECLARATION

Throughout the process of researching and writing this paper, AI tools – notably OpenAI's ChatGPT and NoteboolLM – were employed as valuable assistants.

I used NoteboolLM in the early stages to brainstorm and outline the structure of the report, ensuring it followed the MIETR framework. For instance, I used NotebookLM's podcast generation function to summarize the key points of the two core papers, which helped me draft the Ideate section by comparing methodologies and goals in a concise way. This provided a foundation that I then augmented with direct quotes and details from the papers.

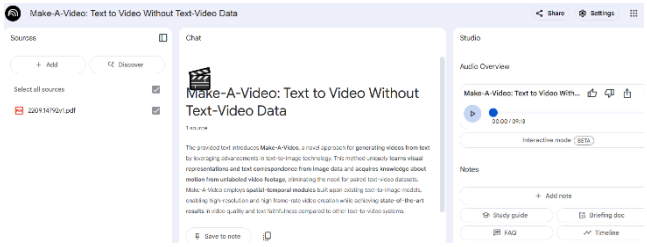

Figure 16
I use NotebookLM to quickly capture the key points of articles and books, and generate podcasts to better understand the possible issues discussed in the text.

NoteboolLM was also helpful in summarization of Daniel Chávez Heras's theoretical concepts: by asking it to explain ideas like "datamatic time" or "machine vision spectatorship" in simple terms, I clarified my understanding before referencing the book.

During the Explore section, I queried ChatGPT about the features of *Sora* and *Hailuo*. It gave me quick bullet points about length,

resolution, and usage which I cross-verified with the official sources. This accelerated the gathering of comparative points.

All AI-provided content was treated as a draft: I carefully verified any factual statements against the original papers or credible sources. Importantly, ChatGPT's ability to generate well-structured text helped in refining the wording and coherence of the final report. For example, if a paragraph was too verbose, I asked ChatGPT to rephrase it more succinctly. The result was then edited by me to maintain accuracy and citation integrity.

In this paper, AI tools functioned as a writing and research aid, handling repetitive summarization tasks and offering stylistic suggestions, while I, as the author, performed the critical tasks of fact-checking, analysis, and synthesis of the content. This collaboration with AI allowed me to focus more on higher-level comparisons and reflections, improving the efficiency of the report-writing process.

7. REFERENCES

[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, Karsten Kreis. (2023). Align your Latents: High-Resolution Video

[2] Edwards, B. (2024, 2 23). *Tyler Perry puts $800 million studio expansion on hold because of OpenAI's Sora*. Retrieved from arsTechnica: https://arstechnica.com/information-technology/2024/02/i-just-dont-see-how-we-survive-tyler-perry-issues-hollywood-warning-over-ai-video-tech/

[3] Franzen, C. (2024, 10 8). *Hailuo gets feature competitive, launching image-to-video AI generation capability*. Retrieved from VentureBeat: https://venturebeat.com/ai/hailuo-gets-feature-competitive-launching-image-to-video-ai-generation-capability/#:~:text=A%20fast,video%20generation

[4] Heras, D. C. (2024). *Cinema and Machine Vision: Artificial Intelligence, Aesthetics and Spectatorship.* Edinburgh: Edinburgh University Press

[5] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, David J. Fleet. (2022). Video Diffusion Models. *arXiv:2204.03458*.

[6] Leffer, L. (2024, 3 4). *Everything to Know About OpenAI's New Text-to-Video Generator, Sora*. Retrieved from Science American: https://www.scientificamerican.com/article/sora-openai-text-video-generator/#:~:text=In%20terms%20of%20the%20duration,He%E2%80%99s%20not

[7] Levy, S. (2024, 2 15). *OpenAI's Sora Turns AI Prompts Into Photorealistic Videos*. Retrieved from WIRED: https://www.wired.com/story/openai-sora-generative-ai-video/

[8] Manovich, L., & Arielli, E. (2022). *Artificial aesthetics: A critical guide to AI, media and design* [Manuscript]. Retrieved April 3, 2025, from https://manovich.net/content/04-projects/175-artificial-aesthetics/manovich_and_arielli.artificial_aesthetics.all_chapters_final.pdf

[9] Mauran, C. (2024, February 16). *What was Sora trained on? Creatives demand answers.* Mashable. https://mashable.com/article/openai-sora-ai-video-generator-training-data

[10] Meta, A. (2023). *Make-A-Video Studio*. Retrieved from Make-A-Video: https://makeavideo.studio/#:~:text=Image

[11] Metz, C. (2024, 2 15). *OpenAI Unveils A.I. That Instantly Generates Eye-Popping Videos*. Retrieved from The New York Times: https://www.nytimes.com/2024/02/15/technology/openai-sora-videos.html

[12] OpenAI. (2024, 2 15). *Creating video from text*. Retrieved from OpenAI: https://openai.com/index/sora

[13] OpenAI. (2024). *Sora system card*. https://openai.com/index/sora-system-card/

[14] Sharwood, S. (2025, March 17). China announces plan to label all AI-generated content with watermarks and metadata. *The Register*. https://www.theregister.com/2025/03/17/asia_tech_news_roundup/

[15] OpenAI. (2024, 2 15). *Video generation models as world simulators*. Retrieved from OpenAI Sora: https://openai.com/index/video-generation-models-as-world-simulators/

[16] Richie Cotton, Matt Crabtree. (2024, December 9). *What is OpenAI's Sora? How it works, examples, features*. DataCamp. Synthesis with Latent Diffusion Models. *arXiv:2304.08818*.

[17] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, Robert Geirhos. (2025). Do generative video models understand physical principles? *arXiv:2501.09038*.

[18] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, Aditya Ramesh. (2024, 2 15). *Video generation models as world simulators*. Retrieved from OpenAI: https://openai.com/index/video-generation-models-as-world-simulators/

[19] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman. (2022). Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792*. doi:https://doi.org/10.48550/arXiv.2209.14792

[20] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, Yu-Gang Jiang. (2024). A Survey on Video Diffusion Models. *arXiv:2310.10647*.