ACTIVANT RESEARCH

Is Video AI Having Its GPT-3 Moment?

Inside the Models, the Economics, and What Comes Next

June 2026

AUTHORS:

Emma Rowand
Research Analyst

Jono Vickery
Vice President, Research

About Activant

Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

Seven months after launch, OpenAI pulled the plug on Sora's consumer app.[1] The product was reportedly burning $15 million a day in compute costs, had generated just $1.4 million in total in-app revenue, and was draining GPU capacity the company needed elsewhere.[2,3] For a product meant to showcase AI's creative frontier, the numbers painted a more sobering picture.

But Sora's death isn't really about one product at one company. It's a signal about where the entire video AI market sits right now: technically impressive, economically punishing, and far less like the LLM race than it appears. Today's leading video models can generate striking short clips, but quality degrades quickly past fifteen seconds, the top models change every month, and a single ten-second clip costs roughly $1.30 in compute.[4]

On the surface, that puts video AI somewhere around where LLMs were in 2020. But the resemblance ends there. Language models were always converging toward the same capability: more fluent, more accurate, and more general-purpose. Video models are doing the opposite. Each is developing a distinct creative identity, with real differences in motion quality, stylistic control, speed, and precision.

To understand the market better, we looked at the architectures driving that divergence, the business models being tested against punishing unit economics, and the legal fault lines forming around training data. *We think the orchestration layer is where durable value will concentrate.*

The Model Landscape: Divergence, Not Convergence

For the last five years, the LLM race has been a fight over the same turf. OpenAI, Anthropic, Google, have all converged on capability, competing over benchmarks and pricing while serving largely similar use cases.

Fal founder Batuhan Taskaya noted the half-life of the top five video models was roughly 30 days through mid-2025, with rankings reshuffling constantly.[5] That pace of churn might look like the early LLM race, but the underlying dynamic is different. Language models converged because language is language. There's one kind of fluent English paragraph but there are many kinds of good video.

Sora produces naturalistic motion with a painterly quality. Kling prioritizes structured filmmaking with frame-level control. Veo leans into photorealism. Seedance handles complex multi-subject scenes. These stylistic variations reflect fundamentally different training data, architectural choices, and optimization targets. Because creative preference is subjective and context-dependent, this divergence is likely permanent. Raw quality will converge as models approach a realism ceiling. But creative character will not, for the same reason that Spielberg and Nolan both make films but are not interchangeable.

Each model stakes out different ground on what it does best, the inputs it accepts, and how it handles audio, physical realism, and scene continuity. There is no single leader here, only different answers to different creative problems

AI Video Generation Models: Head-to-Head (2026)

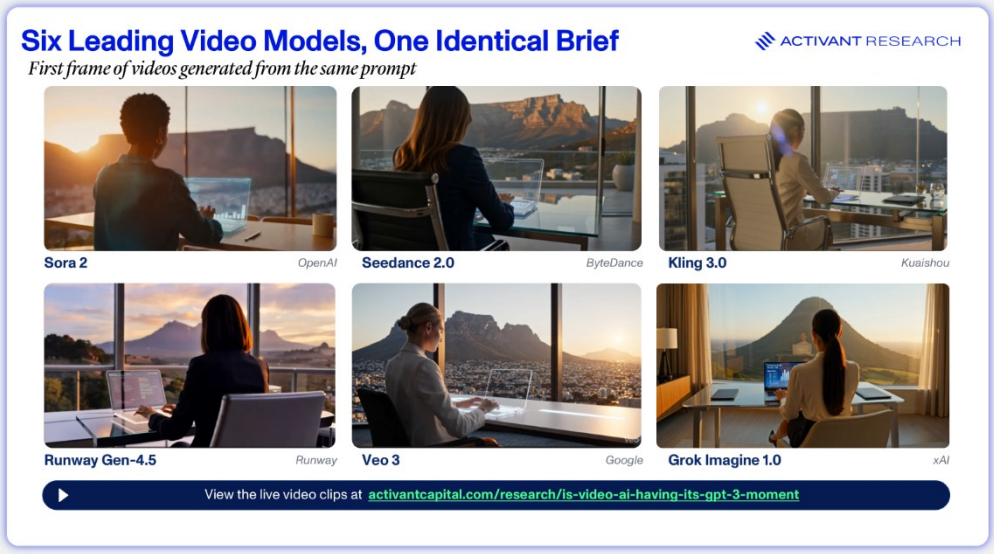
ACTIVANT RESEARCH

Model	Ideal For	Strengths	API Cost/ Min	Monthly Subscription Cost
Veo 3 (Google) Veo 3	Short, polished video content from cinematic clips to trailers with synchronized audio (music, dialogue, effects, ambient sound).	• Native audio • Strong prompt adherence • Realistic physics & natural movement • High-resolution • Character consistency	$9.00 - $24.00	$19.99 (Pro) to $249.99 (Ultra)
Grok Imagine 1.0 (xAI)	Rapid prototyping, concept art iteration, and all-in-one audiovisual synthesis.	• Native audio generation • Cross-modal architecture • Rapid generation speed	$4.20	$30.00 (SuperGrok) to ~$40.00 (Premium+)
Seedance 2.0 (ByteDance) ByteDance	Cinematic continuity, advanced video to video generation, multi-shot storytelling, and structured control using a unified @-tagging reference system.	• All-in-one reference system • Work across text, image, video, audio • Seamless scene extensions • Consistent scene throughput • HD output	$0.80 - $4.82	$29.90 (Basic), $49.90 (Standard), $99.90 (Pro) and $199.90 (Max)
Kling 3.0 (Kuaishou) KlingAI	Multi-shot scenes, dialogue-driven videos, action sequences, and branded content that require consistent characters, dynamic camera movement, and synchronized audio.	• AI Director workflow • Storyboard Controls • Advanced reference handling • Native audio • 4K cinematic output	$7.56 - $25.20	$6.99 (Standard), $25.99 (Pro), $64.99 (Premier) and $127.99 (Ultra)
Runway Gen-4.5 runway	Photorealism, physical realism, product demonstrations, and visual effects integration.	• Accurate physical dynamics • Detail retention during motion	$7.20	$15.00 (Standard) to $95.00 (Unlimited)
Sora 2* (OpenAI)	Storytelling, trailers, short films, or social videos that demand realism and creative control.	• Multi-modal input • Integrated audio • Physics-aware motion • Multi-shot scene handling	$6	$20.00 (Plus) to $200.00 (Pro)

*Sora developer API remains available until September 24, 2026.
Sources: Pingoy, InVideo, DataCamp, Runware, fal.ai, MindStudio, official model documentation | Prices and features as of April 2026

To test how consistently today's leading models interpret the same instructions, we ran a single structured prompt across multiple models, keeping parameters as uniform as possible: 16:9 aspect ratio, 8 seconds duration and 720p resolution.[6]

The prompt: *A professional woman types on a transparent laptop, then swipes a glowing holographic chart mid-air. The window in front of her shows Table Mountain in golden hour light. Camera dollies in from behind, then orbits as holo-charts rise from screen. Cinematic, shallow depth of field, anamorphic flare.*

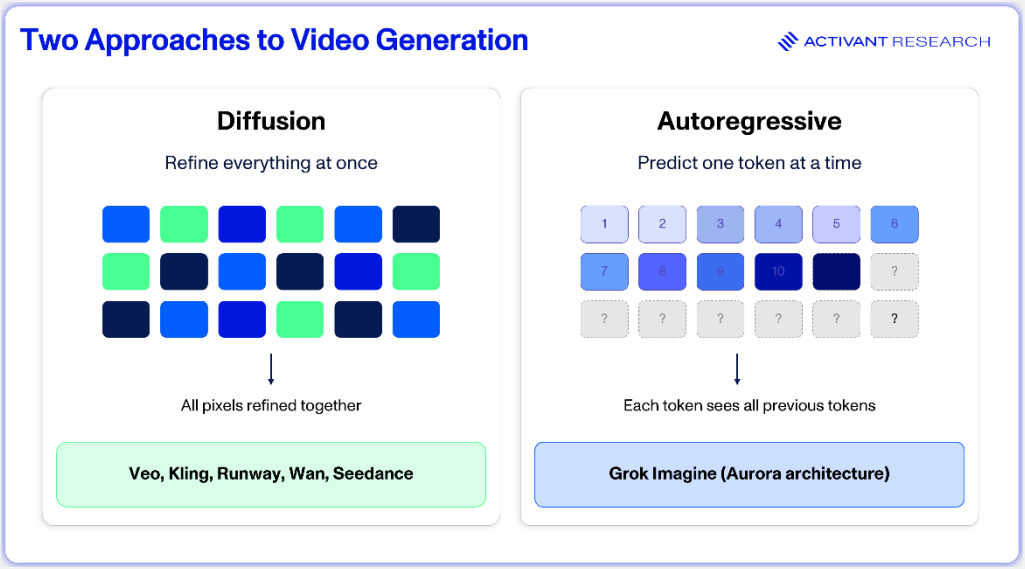


To understand why these models differ so greatly, we need to look at the underlying architecture. Like LLMs, video models have converged on transformers, but they split sharply on the generation paradigm layered on top and the implementation choices that follow.

The Technical Stack: Diffusion vs. Autoregressive

Transformers replaced the U-Net convolutional designs that dominated earlier video generation. Transformers capture relationships across both space and time more effectively, translating into more coherent motion and more consistent scenes frame-to-frame. What separates the leading models is the generation paradigm layered on top: most use **diffusion**, while a smaller camp uses **autoregressive** generation.

Before either paradigm can run, raw video needs to be compressed into something manageable. The quality of this step has a disproportionate impact on the final output. A weak compressor leads to blurry frames, jerky motion, or flickering, no matter how good the rest of the model is. Most leading models now compress along height, width, and time simultaneously, shrinking the data dramatically before generation begins.

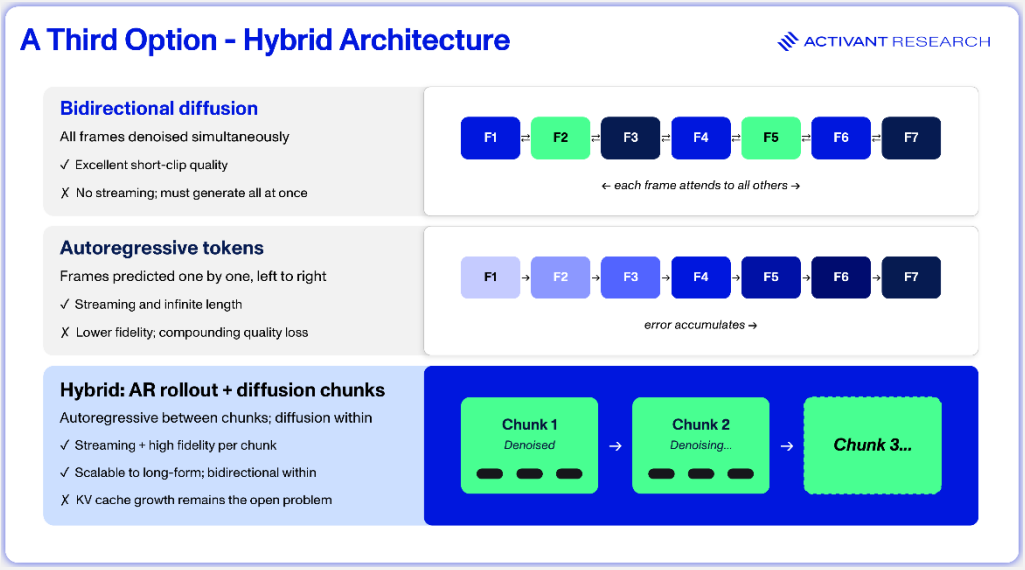


Diffusion starts with pure noise and gradually refines it into a coherent image or video.[7] Because each generation begins from randomness, it follows a slightly different path, even when given the same prompt. Run the same instructions twice and you'll get noticeably different results. Run them across different models and the gap widens further: different motion, visual style, color, and interpretation of what you asked for. Google's Veo, Kuaishou's Kling, ByteDance's Seedance, and Runway's Gen-4.5 all fall into this camp, each layering on its own variations like flow matching or cascaded diffusion.

Autoregressive models take a different approach. Instead of denoising an entire frame at once, they predict visual tokens sequentially, much like a language model generates text word by word.[8] xAI's Grok Imagine is the clearest example, built on what they call the Aurora architecture. This sequential method allows tighter frame-to-frame control and is a big part of why Grok Imagine has a noticeably different visual character from its diffusion-based competitors.

Even with transformers as a shared backbone, implementations vary. Models compress data at different ratios along height, width, and time, and layer on architectural choices like Mixture of Experts (MoE).

Diffusion dominates video generation today because it builds on a mature image generation ecosystem and because generating an entire video at once produces exceptional visual coherence in short clips.[9] But compute and memory costs scale steeply with longer sequences, making long videos prohibitively slow and expensive. Autoregressive architectures avoid this by generating video chunk by chunk, naturally supporting longer outputs and interactive applications like game engines and robotics world models. The trade-off is that small errors accumulate over time, gradually degrading quality the longer a video runs.[10]



The most promising recent work combines both paradigms: **autoregressive rollout for temporal structure with diffusion-based refinement at each step**. Early hybrid models have already matched state-of-the-art quality benchmarks while enabling real-time streaming on a single GPU but significant challenges remain.[11] Memory usage grows as sequences get longer, and the models tend to drift because they train on clean data but must build on their own imperfect outputs at generation time.[12,13] Despite their architectural differences, both camps are converging on the same goals: more quality from less compute, and finer control over what appears on screen and when.

The good news is that memory capacity and bandwidth are growing with each generation of NVIDIA chips, which should eventually ease these constraints. But for now, all models are bottlenecked by the hardware they run on.

Commercial Viability: The Unit Economics Problem

If no single model can do everything, and better models require more compute, the question becomes whether anyone can build a viable business around this technology at current costs.

The early answer appears to be no. Sora generated $1.4 million in lifetime global net in-app revenue while ChatGPT generated $1.9 billion in the same period.[14] Despite 11 million users and two billion views, Pika reportedly directs over 60% of its funding to compute.[15] Every user splicing themselves into a Hollywood chase is burning money the company can't recoup.

The reflex is to call this a compute cost problem, which is partly true. Cantor Fitzgerald estimated each ten-second Sora clip cost roughly $1.30 to generate, factoring in about 40 minutes of GPU time across four parallel GPUs at nearly $2/hour rental rates.[16] At $20/month, a ChatGPT Plus subscriber only needed to generate 15 clips to exceed their subscription cost, and that $20 also covered image generation, Codex, and Deep Research. GPU costs will come down naturally and what's uneconomical today may look different in two years. So, was Sora really a compute problem, or rather a business model problem?

The Business Model Problem

In our view, Sora failed because it targeted the wrong customer at the wrong time.

OpenAI launched Sora as a consumer product, likely because that's the playbook they know. But the compute costs associated with video generation are dramatically higher than text generation. For a consumer paying $20/month, the math simply doesn't work. Either the product must be heavily rate-limited (which kills the user experience), or OpenAI must subsidize usage at a loss.

Beyond unit economics, B2C doesn't have strong product-market fit. Most individual users don't have a recurring, high-value need for AI video. They experiment a few times, share something on social media, and move on. There's no deep workflow integration, no daily habit and no retention hook comparable to what ChatGPT offers as a general-purpose assistant. And $20/month is a tough sell when Netflix delivers a massive library of professional content through a personalized recommendation engine for roughly the same price. At this stage of the cost curve, B2C is the wrong surface to monetize.

Enterprise buyers are different. Large companies already spend heavily on video production for advertising, marketing content, product demos, training materials, and localization. Traditional enterprise video costs often exceed $12,000 per hour of final content.[17] When Artlist.io aired their Super Bowl ad this year, they paid the same $8 million for airtime as every other brand.[18] But while most advertisers spent another $5 to $15 million on production and talent, Artlist built their spot in

five days for a few thousand dollars using AI tools.[19] For these buyers, an imperfect AI tool that cuts production timelines from weeks to days or halves the cost of generating video variants is a straightforward ROI calculation. They don't need to be convinced that video is worth paying for; they need to be convinced that AI can do it well enough to displace some portion of their existing spend.

Enterprise customers are also more tolerant of the technology's current limitations. A marketing team using AI to generate first-draft storyboards, rough cuts for internal review, or high volumes of social media variants doesn't need perfect cinematic output. They have internal creative teams who can refine the results. Consumers, by contrast, judge the product against professional content on YouTube and TikTok, where any visible artifact or inconsistency feels like a failure.

Put simply, the value proposition doesn't work for consumers yet because the technology is neither cheap enough nor good enough to serve casual, low-willingness-to-pay users at scale.

For the players straddling both B2C and B2B, defensibility increasingly comes down to distribution. Sora launched as a standalone app with no built in distribution or workflow to anchor it. Meanwhile, ByteDance has TikTok (an estimated 900m DAUs), Kling has Kuaishou (441m DAUs), and Google has YouTube (2.5bn MAUs).[20,21,22] Runway appears to understand this dynamic as well, which is why it bolted itself onto Adobe.[23]

Among the B2C survivors, Kling looks the most commercially sound, crossing $240m in ARR with 60 million users by December 2025.[24] Its $6.99/month standard tier sits in a sweet spot individual creators can justify.[25] But what makes Kling's position durable is that Kuaishou doesn't need it to stand on its own. Unlike Runway or Pika, which must cover compute costs entirely through generation revenue, Kuaishou can subsidize inference costs across a business running at 55% gross margins.[26] Kling feeds back into that engine: advertisers use it to produce video at more than 60% lower cost than traditional production, driving ad spend on the platform, while the added content boosts user traffic and engagement.[27] Priced at 20-30% of Veo 3, with 70% of revenue from international markets, Kling is building a global subscription base while simultaneously strengthening the core business that funds it.[28]

Runway and Pika operate without that safety net. Runway has positioned itself as the premium option for professional users, with advanced camera controls, world consistency, and HD output. Its recent $315m raise at a $5.3bn valuation is funding a pivot toward world models for robotics and gaming, higher-value verticals that could better justify compute spend, while a CoreWeave deal expands infrastructure capacity.[29] Pika, valued at $470m on $135m raised, has taken the opposite approach, doubling down on B2C as a consumer brand for viral content. But its last known fundraise dates back to mid-2024, and the silence since may be telling.[30]

It's also possible we haven't found the right consumer business model for video AI yet. Chatbots proved consumers will pay for text-based tools, but video may not follow the same pattern. For now, the economics only hold for professional and enterprise buyers, customers already spending real money on video and able to measure real savings.

If enterprise buyers are the only segment where the economics work, the obvious next question is whether the biggest buyers can cut out the middleman entirely.

The Production Gap: Why Hollywood Isn't There Yet

In late 2024, Lionsgate licensed its film library to Runway to build a custom model, one of the first major studio-to-model-provider deals. It looked like a template for the industry: a clean licensing arrangement that sidestepped the legal minefield of scraping the open internet. Within a year, it was reportedly failing.[31]

The core problem was **data asymmetry**. Foundation models train on internet-scale datasets orders of magnitude larger than any single studio library, so a studio-trained model lacks the diversity to generate novel content and instead ends up resembling existing material. Fine-tuning a broad model on studio data doesn't fully solve the issue either. You either get **underfitting** (the model ignores the studio content) or **overfitting** (it memorizes rather than learns). Meanwhile, legacy archives stored in outdated formats and siloed systems make converting decades of video into model-ready datasets expensive and slow.

The Lionsgate experiment matters because it tested the most intuitive path for enterprise adoption: own the data, own the model, own the output. But while it's tempting to blame the failure on limited data, even foundation models trained on orders of magnitude more footage are hitting hard technical limits of their own. The constraints are driven not only by what goes into the model but also by what the current architectures can get out.

Temporal Coherence and Long-Form Consistency: Every major diffusion model is effectively capped at 5-10 seconds of reliable output. Full spatiotemporal attention carries quadratic memory costs, meaning doubling video length roughly quadruples compute.[32] Autoregressive generation compounds small errors across frames, producing characters that morph, physics that break, and scenes that drift.[33] Techniques like PackForcing (compressing generation history to extend short-clip video models to long-form output) and Context Forcing (extending effective context beyond 20 seconds) are promising, but they manage the degradation curve rather than eliminate it.[34]

Controllability and Creative Precision: Even perfectly coherent long-form video is useless without precise control over camera movement, lighting, blocking, and timing. Runway currently has the deepest controllability toolkit and is embedding it into Adobe's suite. Kling 3.0 approaches the problem from a structured

filmmaking angle, with automated storyboarding and frame-level precision. And Idomoo's Strata produces multi-layer blueprints instead of flat pixel files, so individual elements can be edited independently.[35]

On resolution, the field is converging toward 1080p as the baseline, with several models pushing into 4K. Frame rates have standardized around 24 fps, though Kling supports up to 48 fps. Duration remains the biggest constraint: most models generate 8-25 second clips, with Kling the clear outlier at up to three minutes of continuous output through its Extend feature.[36]

Audio-Visual Alignment: Most leading models now generate synchronized audio natively, but lip-synced dialogue remains the weakest link. Many phonemes look nearly identical visually, and standard evaluation metrics perversely reward less lip movement, meaning near-static faces outscore realistic speech animation.[37] [38] Veo 3 and Kling 3.0 have pushed the frontier forward, while avatar platforms like Synthesia and Tavus now drive micro-expressions and head movement from vocal tone. But beyond two minutes, emotional range degrades, body consistency breaks down, and frozen eyes give the game away.[39]

Inference speed ranges from roughly 10 seconds to several minutes, which matters for creative workflows where iteration speed determines how many ideas can actually be explored. Meanwhile, **input modalities** are broadening quickly: text-to-video and image-to-video are standard, but models like Seedance now accept up to 12 reference files spanning text, image, video, and audio.[40]

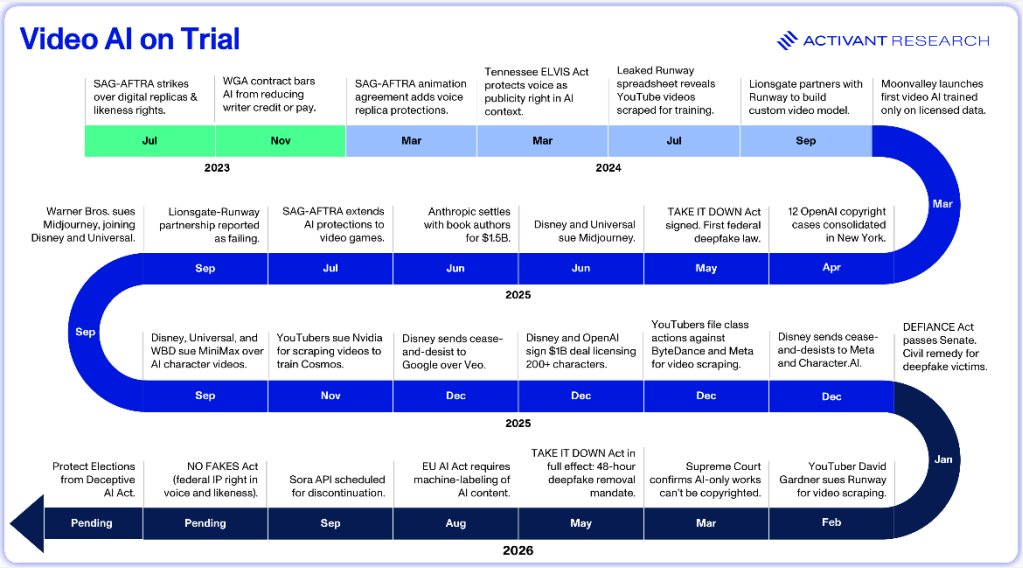
None of this means video AI is useless to studios and brands today. A company already spending seven figures on shoots, VFX, and post-production can meaningfully reduce costs and timelines through partial automation, even at current quality levels. Nestlé, Unilever, and WPP are leaning into AI-assisted production, with some industry estimates suggesting as much as 70% of content could become AI-assisted within the next few years.[41] Short-form social content and corporate training videos already work. With Synthesia, Zoom cut video production time by 90% and saved up to $1,500 per employee, while Heineken reached 70,000 employees across 170 countries in their local languages. Knowlify, a unified AI video platform, is used by teams for internal comms, onboarding and training, and customer education. It produces explainer videos in multiple formats, from animations to talking avatars, with one financial services firm saving $220K a year by moving production in-house.[42] Advertising production is emerging fast, with Fiverr reportedly producing commercials at roughly 10% of traditional costs.[43] Full film production and real-time interactive video remain more speculative but increasingly plausible.

Enterprise workflows also depend on making existing footage usable. TwelveLabs indexes and searches video libraries at scale through a single API, letting teams query years of archived material in natural language. For studios with decades of footage, brands managing past campaigns, or companies updating training libraries, retrieval is as much a part of the stack as generation.

Closing the remaining gaps requires better models. Better models require more and better training data. And more and better training data means wading deeper into legally contested territory.

Copyright, Ethics, and the Legal Landscape

The video AI industry sits in an uncomfortable paradox: the models most useful to independent creators tend to be trained on the broadest, and legally riskiest, datasets. The safest models produce the most generic output. Where a company falls on this spectrum increasingly determines its legal exposure, output quality, and viability as a professional tool.



At one end sit ByteDance and Kuaishou, with access to effectively limitless user-generated video, much of it uploaded under terms of service that predate generative AI. In December 2025, YouTubers filed class actions against both ByteDance and Meta, alleging they bypassed YouTube's technical protections to scrape millions of training videos.[44]

At the other end is Adobe, which built Firefly entirely on licensed data and offers commercial users legal indemnification. The trade-off is constrained creative output: user reviews and Adobe's own community forums cite results that are generic, distorted, and creatively sterile.[45,46] Moonvalley, founded by former DeepMind researchers, launched its Marey model in 2025 as "the first fully clean AI model," trained exclusively on licensed footage. About 80% of its data is B-roll licensed from independent filmmakers, with gaps filled by footage shot in-house.[47] Moonvalley was betting that building cleanly would be a structural advantage once the legal dust settled, even if it meant training on roughly one-fifth of the data available to competitors. Last week, however, Moonvalley joined forces with Reka, an AI lab focused on the physical world, shifting its research team toward robotics, simulation, and world models rather than filmmaking tools.[48] The pivot suggests the clean-data thesis, however principled, was a harder business to sustain on its own, a reminder of the brutal economics of the industry.

Meanwhile, regulation is accelerating. The EU AI Act's transparency requirements take effect in August 2026, while the C2PA coalition (backed by Adobe, Microsoft, Google, Meta, Sony, the BBC, and 200+ others) has already built an open standard for Content Credentials: cryptographically signed records that function like nutrition labels for digital media.

In between sits a growing middle ground where many of the industry's most interesting experiments, and most instructive failures, are unfolding.

Hollywood's Contradictions

Hollywood's major studios are simultaneously plaintiffs and customers. In December 2025, Disney sent Google a cease-and-desist over AI-generated use of its characters while simultaneously announcing a $1 billion licensing deal with OpenAI covering 200+ characters for Sora.[49] **Studios want control over how their IP is used in AI, not to stop it being used at all.**

Then came the talent question. Lionsgate may own its films' IP, but actors' likenesses and contractual rights don't transfer as simply. If a John Wick-trained model generates output resembling Keanu Reeves, who gets paid?

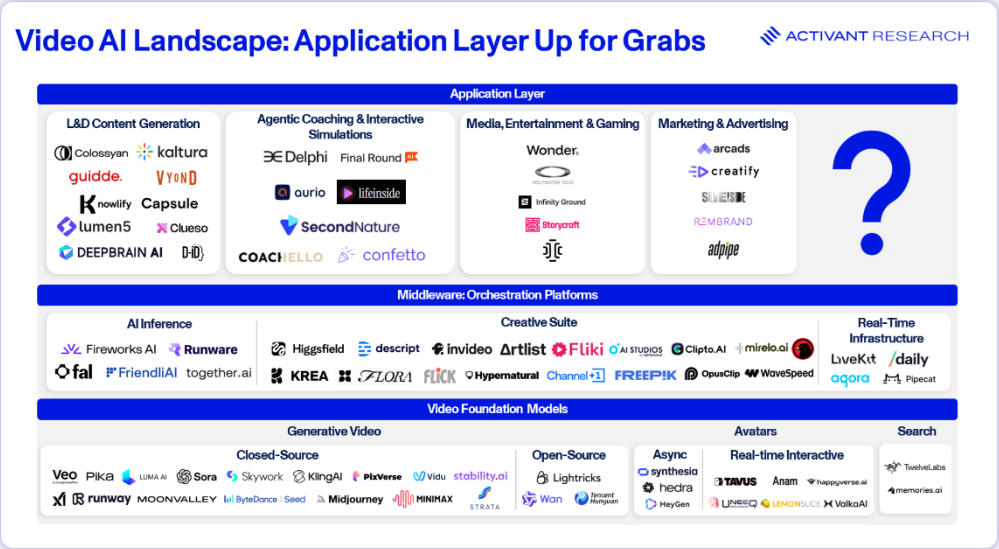
Since the 2023 strike, SAG-AFTRA has steadily expanded AI protections across its contracts, requiring consent for digital replicas and "independently created" AI likenesses.[50,51] The union recently filed charges over the AI recreation of James Earl Jones's voice in a videogame.[52] In parallel, deepfake legislation has moved quickly. The federal TAKE IT DOWN Act (May 2025) and DEFIANCE Act (January 2026) now sit alongside deepfake laws in 46 states, with 146 bills introduced in 2025 alone.[53]

So Where Does That Leave The Market?

The fundamental tension remains unresolved. The most capable models carry the greatest legal risk, while the cleanest datasets produce the most constrained output. At the same time, the model landscape isn't consolidating, and stylistic preferences mean it likely *never* will. Compute and memory constraints make consumer video AI commercially unviable, and enterprise buyers, the only segment where the economics work, are the least willing to tolerate legal or reputational exposure.

In a market this fragmented, where the regulatory picture is only getting more complex, value accrues to whoever can abstract that complexity away and offer enterprise customers a single, defensible workflow they can trust.

The Case for Orchestration

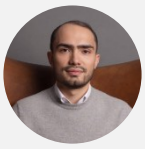


Our thesis in voice AI pointed to a natural ceiling: you can't get more real than real. Once voice providers like ElevenLabs achieved human-sounding output, the next move was owning the full agent workflow, keeping users on-platform and cutting out the orchestration layer.

Video is fundamentally different. Voice had one optimization target: sound human. Video has many. A brand might want cinematic polish for a product launch, an anime aesthetic for a game trailer, and photorealistic talking heads for corporate training, all in the same week. No single model excels at all of these, no single provider can realistically maintain thousands of style-specific models, and no consumer wants five separate $20-40/month subscriptions just to access them.

In an inherently creative industry where choice of model matters enormously, the orchestration layer becomes essential. Orchestration players like Higgsfield aggregate multiple models under one roof, offering access to Sora, Kling, Veo and others for just $15/month.[54] For consumers and small teams, this is a no-brainer. For model providers, orchestration platforms become critical distribution channels that extend reach beyond direct subscribers.

The same dynamic is emerging in adjacent markets. ElevenLabs, best known for voice synthesis, recently launched ElevenCreative, aggregating Sora, Veo, Kling, Wan, and Seedance alongside its own audio models, doubling down on its platform play. LiveKit started as a WebRTC backbone for voice agents and has since expanded into a platform supporting plugin integrations across major LLM, STT, TTS, and avatar providers. Daily, the WebRTC layer behind Tavus, HeyGen, and Lemonslice, has paired its transport with Pipecat, an open-source framework that orchestrates the same model stack in real time. Even Runway now hosts its competitors inside its own subscription, shipping Gen-4.5 alongside the leading closed and open-source models. If even the model providers are conceding that no single model is enough, the orchestration layer is where the customer relationship lives.



Video AI is at the inflection point AI-assisted coding was a year ago. The strong adoption we're seeing now isn't a big surprise. The creator economy is a $250 billion market – value in this market was always going to originate at the application layer, and it's barely started."

Alex Mashrabov
CEO and Co-founder, Higgsfield

The open-weight vs. closed-weight divide is accelerating this trend. The most capable open video models (Wan 2.1, Hunyuan Video) are directly competitive with closed alternatives and free to run. If Runway releases something marginally better than Wan but at 20x the price, production teams will route around it. This is already playing out in image generation, where Flux displaced Midjourney in many professional workflows by being good enough and dramatically cheaper at scale.[55]

That compresses the window in which closed models can command premium pricing on raw generation quality alone. Value is already migrating downstream into fine-tuning, deployment infrastructure, and the application layer. Synthesia trains proprietary model stacks for its core product while relying on general-purpose video models for B-roll.[56] Higgsfield's latest release, the Supercomputer, is a cloud-native, self-learning AI agent that turns a plain-language brief into a finished asset, picking the right model for each step, and shipping the output autonomously. And Runway just launched a $10M fund to back startups building on its models.[57]

On the infrastructure side, serverless inference providers like fal.ai have restructured compute economics by charging per token rather than per GPU hour, eliminating cold starts, and offering API access to over 1,000 models spanning image, video, audio, and 3D through a single integration. Its model-agnostic approach means the inference layer increasingly resembles the orchestration layer, just built from the infrastructure up rather than the application down.

As models converge on quality and diverge on use case, competitive advantage shifts to orchestration: the system that picks the right model for the right task, manages the pipeline end to end, integrates with existing tools, and manages downstream liability. A public-facing ad campaign carries very different legal exposure from an internal training video. We believe orchestrators hold the largest moat precisely because video AI is not a winner-takes-all market. New models will keep emerging from unexpected places, and that disruption only reinforces the value of sitting above the model layer.

Looking ahead, we expect the serverless inference and orchestration layers to converge. With real-time interactive video emerging as the next major test, durable value will concentrate with providers that can handle workloads far more

demanding than today's batch generation while offering intelligent routing based on use case, cost, style, and legal constraints.

The Next Frontier: Real-Time Interactive Video

Everything discussed so far assumes a familiar workflow: prompt a model, wait, get a clip back. A separate frontier is forming around video that responds in real time, where the model isn't a generator you call but a participant you talk to.

Tavus has spent the last four years building this, framing the next wave of computing as a human interface, where talking to a machine feels like talking to a coworker. Their stack, running six models over Daily, shows just how hard this is: STT, LLM, TTS, Sparrow-1 for turn-taking, Raven-1 for perception, and Phoenix-4, a 1080p Gaussian diffusion renderer. Anam uses a similar cascaded pipeline of outsourced STT, LLM, and TTS components, while its Cara-3 model consists of a diffusion transformer that maps audio to motion embeddings and a separate renderer that applies them to a reference image.

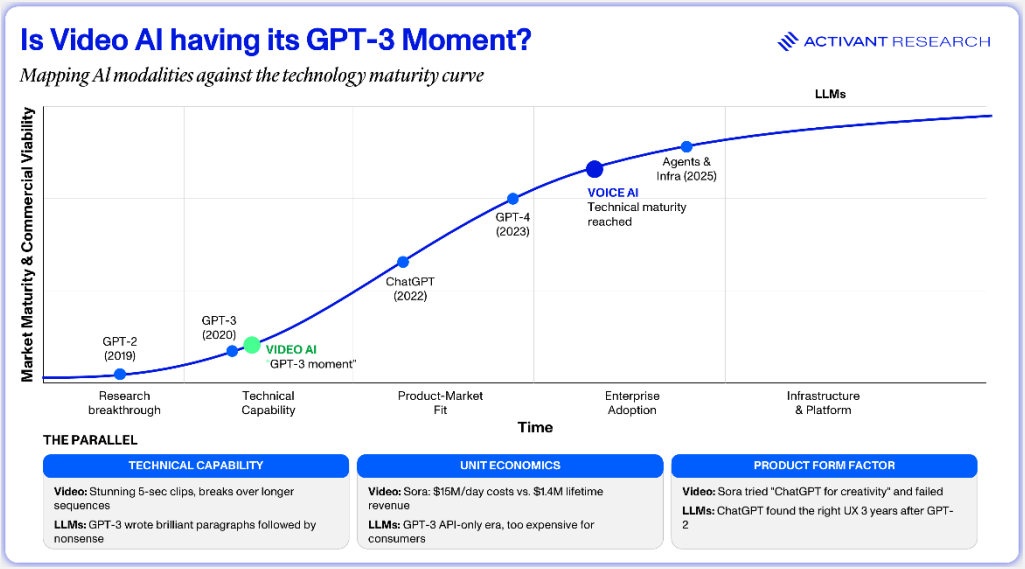
This is the same architectural bottleneck we wrote about in voice AI: stacking models compounds latency, strips paralinguistic signals and multiplies vendor costs. Surprisingly, the video layer itself only adds 100 to 200ms of latency, with most of the lag sitting in the LLM and TTS layers.[58] Thinking Machines Lab's recent interaction model preview argues that to scale, interactivity must be built into the model itself. Today's models experience reality in stages: they wait while the user speaks or types, then freeze perception once generation begins. To ensure real-time interaction is native to the model, Thinking Machines trains its models on continuous two-way streams of audio, video, and text in 200ms micro-turns.[59]

Real-time interactive video and video generation serve different needs, with different costs, latency limits, and buyers. Tavus serves customers ranging from startups to Fortune 10 companies including Amazon, Better, and Alibaba, with use cases across healthcare intake, recruiting, sales, training, and even elderly companionship.[60] The economics work for the same reason they work in enterprise video generation: the buyer is already paying for the workflow being replaced.

Tavus customers are finding personalized video use cases the company never imagined, and Synthesia's founder is rethinking education entirely.[61] Once a model is always listening and always watching, video moves from creative tool to interface, unlocking entirely new possibilities. Getting there depends on model quality, cost, and legal infrastructure. So where does video AI stand today, and what can the LLM race tell us about where it's headed?

Mapping the Maturity Curve Against the LLM Playbook

Video AI in 2026 is roughly where LLMs were in 2020. The technology works, but only in flashes. Generation costs are too high for consumers and not yet reliable enough for enterprise production at scale. Top models reshuffle monthly, multiple architectures are still competing for dominance, and nothing has meaningfully consolidated.



But video is unlikely to follow the LLM playbook. Language models consolidated around whoever built the best model because “better” meant roughly the same thing to everyone: more fluent, more accurate, and more capable across the same tasks. Video is fundamentally different. Better means different things to different users in different contexts. That makes this a market where providers compete on character rather than raw capability, where open-weight alternatives erode the pricing power of closed models, and where the binding constraint isn't who builds the best model but who can route the right model to the right job while balancing cost, quality, and legal exposure.

Sora's collapse is the market's first real correction, a $15 million-a-day proof that raw generation aimed at consumers isn't a viable business. Studios licensing their libraries are learning that limited datasets produce limited models. And the open-weight community is learning that free models still require infrastructure, fine-tuning, and workflow integration to be commercially useful.

These lessons all point in the same direction. The value in video AI does not sit in the model alone. It sits in the system that decides which pixels to generate, how to generate them, and how to guarantee the output is something a paying customer can actually use. The model layer will continue churning and someone we've never heard of will top the leaderboard next month. As pure generation becomes commoditized, the winners will be the orchestration and workflow layers: platforms that stitch multiple models together, plug into real production

pipelines, and deliver lower costs, better UX, and the connective tissue between raw generation and actual work. And perhaps the most exciting part is that the application layer is still wide open. From video as a new computing interface to world models that simulate physical environments for robotics, we're betting the most interesting applications haven't even been built yet.

***

If you're building in this space, we'd love to connect.

[1] The Sora developer API remains available until September 24, 2026, after which all Sora endpoints stop.

[2] BBC News, OpenAI closes Sora video-making app and cancels $1bn Disney deal, 2026

[3] The Wall Street Journal, The Sudden Fall of OpenAI's Most Hyped Product Since ChatGPT, 2026

[4] Forbes, OpenAI Could Be Blowing As Much As $15 Million Per Day On Silly Sora Videos, 2025

[5] Sequoia Capital, The Rise of Generative Media: Fal's Bet on Video Infrastructure and Speed, 2025

[6] Grok's available duration parameters (1, 3, 6, 9, 12, or 15 seconds) precluded an exact 8-second match, so 9s was used.

[7] MIT Technology Review, How do AI models generate videos?, 2025

[8] Microsoft Research, AR-Videos, 2025

[9] arXiv, From Slow Bidirectional to Fast Autoregressive Video Diffusion Models, 2024

[10] Ibid

[11] Ibid

[12] Ibid

[13] arXiv, Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion, 2025

[14] BBC News, OpenAI closes Sora video-making app and cancels $1bn Disney deal, 2026

[15] Activant Expert Network

[16] Medium, OpenAI Sora Shutdown: $15M/Day Costs, $2.1M Revenue — The Full Story, 2026

[17] Reezo AI, Enterprise AI Video Revolution: Fortune 500 Savings 2025, 2025

[18] Yahoo Finance, How Much Does A Super Bowl Commercial Cost In 2026? 30-Second Ad Prices Explained, 2026

[19] Artlist Blog, How Artlist made a Big Game commercial in 5 days, 2026

[20] DemandSage, TikTok User Statistics, 2025

[21] EQS News, AI emerges as a mid to long-term growth engine: Kling 2.1 sets new standard for cost-efficient video generation, 2025

[22] Limelight Digital, YouTube Statistics, 2025

[23] Adobe, Adobe and Runway Partner, 2025

[24] PR Newswire, Kling AI Annualized Revenue Run Rate Hits USD240 Million in December 2025, 2025

[25] Kling AI, Membership Plan, 2026

[26] JP Morgan, Beijing Kuaishou Technology, An AI application play; initiate on the '31s/'36s at N/OW (23 March 2026)

[27] Ibid

[28] Ibid

[29] TechCrunch, AI video startup Runway raises $315M at $5.3B valuation, eyes more capable world models, 2026

[30] Pika, Fundraising Announcement: Pika raises $80M, so anyone can make video on command, 2024

[31] Futurism, Lionsgate's Attempt to Create Movies Using AI Has Crumbled Into Disaster, 2025

[32] arXiv, PackForcing: Short Video Training Suffices for Long Video Sampling and Long Context Inference, 2026

[33] Ibid

[34] Ibid

[35] Business Wire, Idomoo Introduces Strata: The First AI Foundation Model for Layered Video, 2026

[36] Kling AI, AI Video Extension, 2025

[37] arXiv, Dr. SHAP-AV: Decoding Relative Modality Contributions via Shapley Attribution in Audio-Visual Speech Recognition, 2026

[38] arXiv, THEval: Evaluation Framework for Talking Head Video Generation, 2025

[39] arXiv, Lights, Camera, Consistency: A Multistage Pipeline for Character-Stable AI Video Stories, 2025

[40] WaveSpeed AI, Seedance 2.0 Complete Guide: Multimodal Video Creation, 2025

[41] Activant Expert Network

[42] Knowlify, A global financial services firm brought L&D production in-house and saved $220K annually., 2026

[43] Adweek, Fiverr's AI-Generated Mascot Is Built to Be the Internet's Punching Bag, 2025

[44] Bloomberg Law, Meta, ByteDance Hit With YouTubers' AI Copyright Scraping Suits, 2025

[45] G2, Adobe Firefly Reviews, 2026

[46] Adobe Community, Firefly is terrible, 2026

[47] TIME, This AI Video Startup Is Trying to Win Hollywood's Trust, 2025

[48] PR Newswire, Reka and Moonvalley Join Forces to Advance Models and Infrastructure for Physical AI, 2026

[49] TechCrunch, Disney hits Google with cease and desist, claiming massive copyright infringement, 2025

[50] SAG-AFTRA, Digital Replicas, 2023

[51] SAG-AFTRA, 2025 Interactive Media Video Game Agreement, 2025

[52] The Wrap, SAG-AFTRA Files Unfair Labor Practice Charge Over AI Replica of Darth Vader's Voice in 'Fortnite', 2025

[53] Las Vegas SUN, Explicit AI image creation increasingly a legal issue amid crackdown on deepfakes, 2026

[54] Higgsfield, Pricing, 2026

[55] Magai, Introducing Flux, 2024

[56] Synthesia, Media, 2026

[57] TechCrunch, Exclusive: Runway launches $10M fund, Builders Program to support early-stage AI startups, 2026

[58] Activant Expert Network

[59] Thinking Machines, Interaction Models: A Scalable Approach to Human-AI Collaboration, 2026

[60] YouTube YC Root Access, Tavus: The AI Human Platform, 2025

[61] Financial Times, Synthesia's Victor Riparbelli: forget everything you know about video, 2026

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see "full list of investments" at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes "forward-looking statements." All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.