FEB 29, 2024

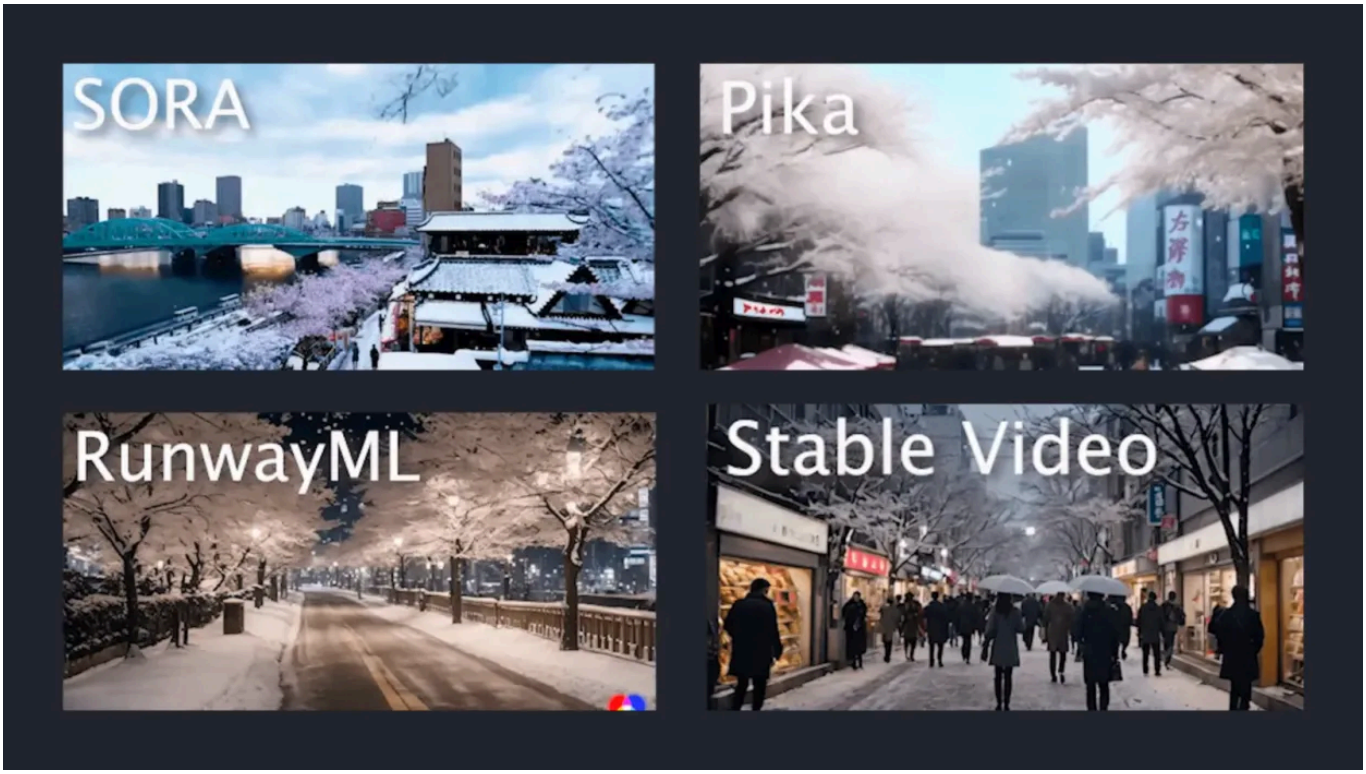
# Video Generation at Scale: Breaking Down Sora's Cost and Compute Requirements

SOTA video-generation models decoded.



By Jenny Xiao and Jay Zhao

Leonis Capital · leoniscap.com/research/sora-and-the-future-of-ai-video-generation/




[Full webinar deck available here.]

As we dive into the fascinating world of AI video generation, it's exciting to see the momentum and curiosity surrounding Sora, a groundbreaking model that's pushing the boundaries of what's possible in this space. In this blog post, we'll explore the core aspects of Sora, its implications for the future, and what it means for investments in AI technology.

# What is Sora?

Sora stands as a pinnacle in the evolution of AI video generation, captivating the tech community with its remarkable capabilities. Developed by OpenAI, Sora represents a fusion of cutting-edge technologies, including a diffusion model and a transformer. But what exactly sets Sora apart from its predecessors?

## Capability 1: High Visual Quality

LEONIS CAPITAL

- Some hypotheses about **why Sora is so much better** than other models:
  a. **Data:** large datasets + specialized data
  b. **Compute:** OpenAI + Microsoft vs. startups
  c. **Architecture:** Sora (diffusion model + transformer) vs. Stable Video (latent diffusion model)




Comparison of videos generated by different models with the same prompt.

CONFIDENTIAL 7

- **High Visual Quality:** One of Sora's standout features is its high visual quality, surpassing other AI models like Pica, Runway ML, and Stable Video by orders of magnitude. This superiority is attributed to several key factors, including its extensive training on a large and high-quality dataset, substantial computational resources from OpenAI, and a unique model architecture combining diffusion and transformer elements.
- **Long-Range Coherence and Object Permanence:** Sora excels in maintaining long-range coherence and object permanence, overcoming challenges faced by earlier models that often resulted in glitchy or unrealistic outputs. This improvement signifies Sora's advancements in simulating real-world physics dynamics, a crucial aspect for achieving more lifelike and immersive video generation.
- **Versatility in Applications:** Beyond generating new videos, Sora demonstrates prowess in text-to-video generation, image-to-video generation, and video editing. These capabilities hold immense value for content creators and editors, enabling seamless creation, transformation, and enhancement of visual content.

- **Understanding of Physics:** Sora's simulation of real-world physics dynamics is particularly noteworthy, resembling the realism seen in gaming engines like Unreal Engine 5. While speculations about its training data on Unreal Engine 5 exist, Sora's ability to mimic complex physics scenarios marks a significant leap in AI's capacity for realistic simulations.

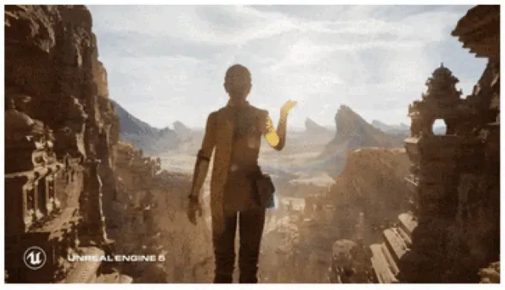
## Capability 4: Simulating the Physical World

LEONIS CAPITAL

**Sora vs. Unreal Engine 5**



Prompt: Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.




Example video generated by Unreal Engine 5.



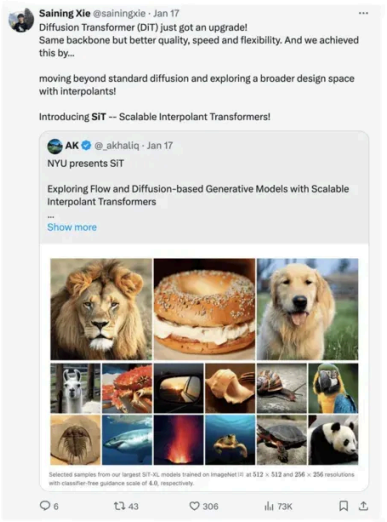
The technical foundation of Sora lies in its innovative architecture, a combination of diffusion and transformer models. This architecture, called a diffusion transformer (DIT), leverages time-space patches as tokens, enhancing scalability and effectiveness in processing vast amounts of data. Despite initial skepticism due to its perceived lack of novelty, Sora's simplicity and scalability have proven instrumental in achieving groundbreaking results.

## Sora: Diffusion Model + Transformer

LEONIS CAPITAL

TL;DR:

- **DiT** = VAE encoder + ViT + DDPM + VAE decoder
  - Prioritized **scalability and simplicity** (much faster than UNet).
  - Demonstrates the viability of the **scaling law** to videos.
  - Builds on Google's work on **ViT** (vision transformers, 2021) -- images treated as tokens, unites vision and language.
- Original paper by NYU team rejected from CVR2023 for "lack of novelty."
- NYU team released **SiT** (self-supervised vision transformer), same architecture but faster and better.



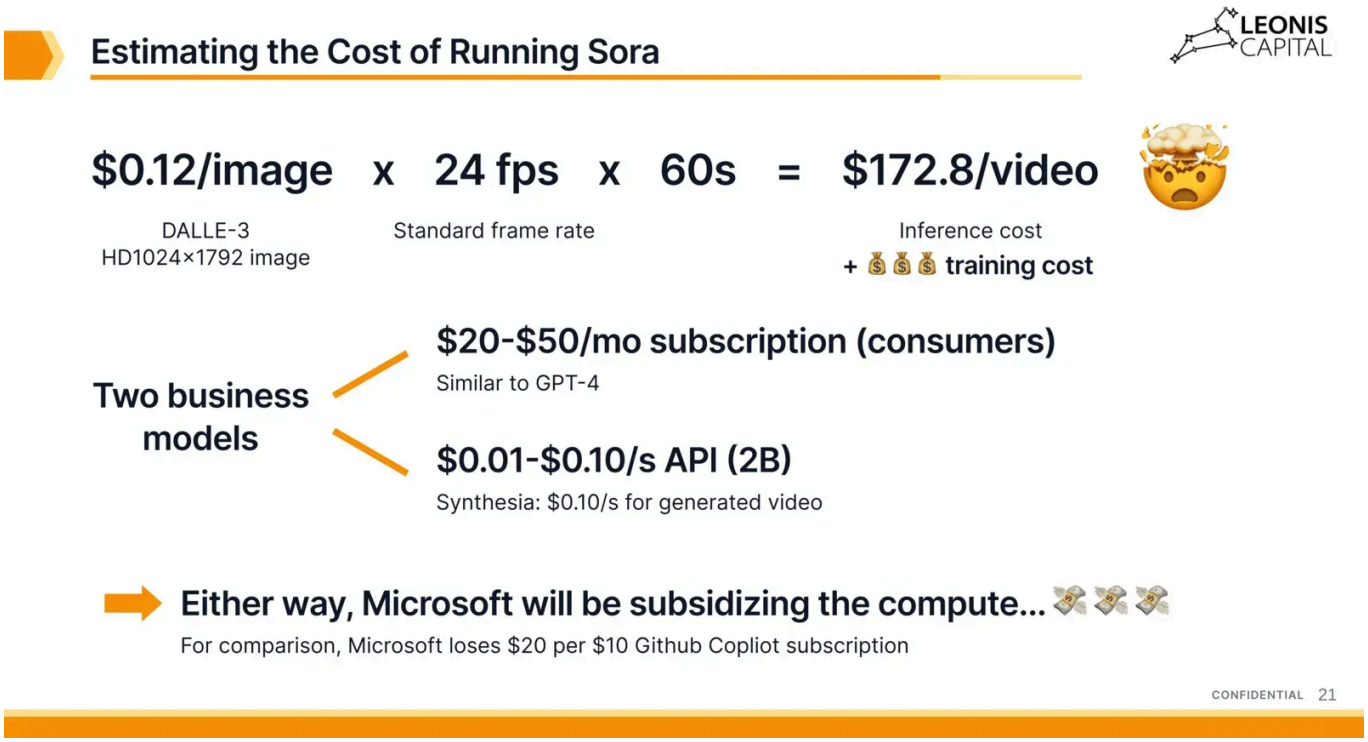



In essence, Sora's emergence represents more than just a leap in video generation technology; it signifies a stride towards Artificial General Intelligence (AGI) by bridging the gap between AI models and real-world simulations.

## Compute and AI Video Generation

One often underestimated aspect of AI video generation is its profound implications for compute costs. Unlike text generation, video generation is an immensely compute-heavy task, demanding substantial resources at every stage of the process.

**In a simple exercise, we estimated the cost of running Sora and it was surprisingly high.** For instance, OpenAI charges 12 cents per image for a single HD image using DALLE-3. Assuming a standard frame rate of 24 frames per second and a 60-second video duration, the cost balloons to $172.8 per video. Even if we conservatively estimate the image generation cost within videos to be one-tenth of the price, it still amounts to $17 per video. When factoring in training costs for models like Sora, the expenses escalate rapidly.




Considering business models for monetizing Sora, two strategies emerge: subscription-based and usage-based. The former, akin to the GPT-4 model, involves charging consumers a monthly fee with a cap on video generations. The latter, catering more to businesses and influencers with high video generation needs, could adopt a pay-per-usage model, potentially ranging from one cent to 10 cents per API call.

However, these pricing models raise concerns about affordability and sustainability. Given the high compute costs, providers like OpenAI may need to heavily subsidize services, akin to how Microsoft subsidizes GitHub Copilot. This dynamic underscores a discrepancy between user willingness to pay and actual costs, challenging the viability of such models without substantial backing or cost optimizations.

**The compute-intensive nature of AI video generation might give rise to specialized chips, such as Groq's LPU.** Unlike GPUs, LPUs offer speed advantages for deterministic tasks like LLM inference. While promising, specialized chips face challenges of flexibility and long-term cost-effectiveness compared to established GPU solutions like Nvidia's, which have proven adaptable and durable over time.

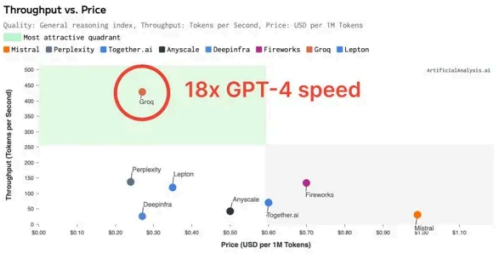
## Groq and New Compute Paradigms?

LEONIS CAPITAL

- **Faster computation** → real-time inference + cheaper video generation
- Groq is an AI chip company founded in 2016 by ex-Google TPU execs.
- Groq's LPU uses **SRAM** (Static Random Access Memory), while GPUs use HBM (High-Bandwidth Memory).
  - LPUs are ideal for deterministic computing tasks like transformer-based model inference.
- BUT:
  - **Specialized AI chips can be brittle** and can quickly become obsolete (e.g., Telsa's CNN chips)
  - Groq might be significantly subsidizing the inference cost and **might not be sustainable**




Groq's promotional materials



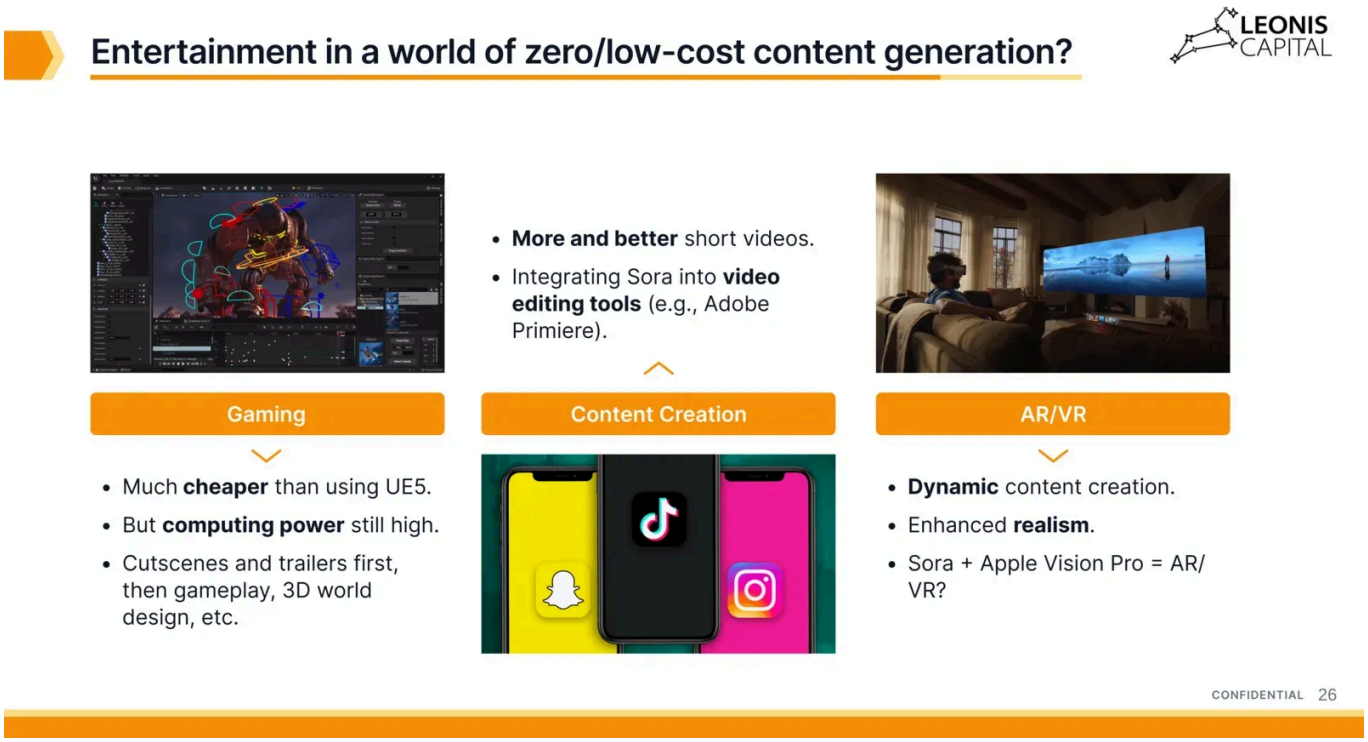

Comparing the speed and price of different AI models

CONFIDENTIAL 24

Ultimately, navigating the compute landscape in AI video generation requires balancing innovation, cost efficiency, and sustainable business models, offering intriguing prospects and challenges for the future of AI-driven content creation.

## Potential Applications and Implications

The applications of AI video generation models like Sora extend across various domains, promising transformative changes and posing intriguing questions for builders and investors alike. Gaming stands out as an early adopter, where dynamic gaming experiences integrating VR and AR technologies could redefine player interactions and immersion. Content creation platforms such as TikTok and YouTube are also poised for significant shifts, as lower content generation costs challenge creators to compete on creativity and editing quality.

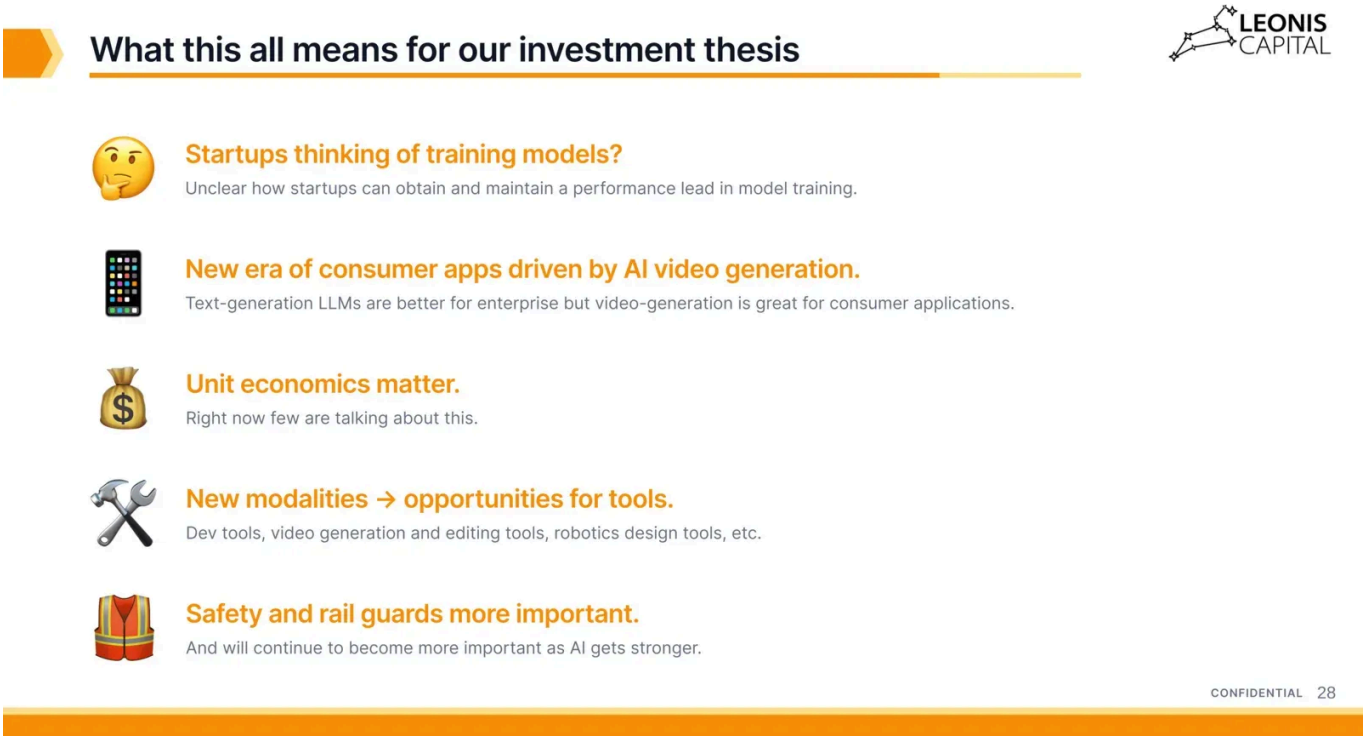



Moreover, the impact extends to consumer entertainment, where low-cost or free internet content generation could become a reality, reshaping how content is consumed and produced. However, questions about unit economics and commercial value remain pivotal, especially as it currently costs around $170 to generate a 60-second video. This cost-effectiveness will be a crucial factor in determining the widespread adoption of these technologies.

In the realm of robotics, advancements in AI video generation could revolutionize development processes. Virtual testing and prototyping could reduce reliance on physical experimentation, accelerating progress toward AGI-powered robotics. Synthetic data generated through AI simulations could streamline training processes, making robotics development faster, cheaper, and more effective.

## Conclusion

The journey through AI video generation models like Sora unveils a world of possibilities and challenges. While the applications span gaming, content creation, and robotics, the underlying question of data quality and its impact on model outputs remains paramount. The interplay between data, model architecture, and real-world applications underscores the dynamic nature of AI innovation.




As builders and investors navigate this landscape, considerations around cost-effectiveness, commercial viability, and safety take center stage. The convergence of AI technologies with consumer demands and industrial needs opens avenues for groundbreaking innovations, but it also requires thoughtful exploration and strategic investment.