

# Helios: A 14B Video Model That Runs at 19.5 FPS on a Single GPU

Kabui, Charles

2026-03-06

---

[Read at ToKnow.ai](#)

---

The graphic features a dark blue background with a grid pattern. At the top left, the title 'Helios: 14B Video Real-Time on 1 GPU' is displayed in white and light blue. Below the title, the text 'Minute-scale video generation without standard acceleration tricks' is written in a smaller white font. Three key performance indicators are highlighted in separate boxes: '19.5 FPS' (Real-time on single H100, No KV-cache or quantization), '14B' (Parameters, cost of 1.3B, 4 models fit in 80GB VRAM), and '60s' (Minute-scale coherent video, No anti-drifting heuristics). A faint line graph is visible in the upper right corner. The 'ToKnow.ai' logo is located in the bottom right corner.

## Helios: 14B Video Real-Time on 1 GPU

Minute-scale video generation without standard acceleration tricks

- 19.5 FPS**  
Real-time on single H100  
No KV-cache or quantization
- 14B**  
Parameters, cost of 1.3B  
4 models fit in 80GB VRAM
- 60s**  
Minute-scale coherent video  
No anti-drifting heuristics

ToKnow.ai

Researchers at ByteDance and Peking University released [Helios](#), a 14-billion parameter autoregressive diffusion model that generates video at 19.5 frames per second on a single NVIDIA H100 GPU. The model produces minute-scale clips (up to 60 seconds at roughly 24 FPS output) and natively handles text-to-video, image-to-video, and video-to-video tasks through a unified input representation. What makes Helios unusual is what it does not use: there is no

KV-cache, no sparse or linear attention, no quantization, and no anti-drifting heuristics like self-forcing or keyframe sampling. Instead, it compresses historical and noisy context aggressively and reduces sampling from 50 steps to 3 via adversarial hierarchical distillation. The result is inference costs comparable to models one-tenth its size. Training is similarly lean: four 14B models fit within 80 GB of GPU memory without parallelism or sharding frameworks.

That speed changes what video generation can be used for. Previous models at this quality level took minutes to produce seconds of footage, limiting them to offline workflows. Helios at real-time speeds opens the door to interactive applications: live content creation tools, game engines fed by generative video, and on-the-fly storyboarding. The [code](#), [base model](#), and [distilled model](#) are all released under Apache 2.0 with day-one support from Diffusers, vLLM-Omni, and SGLang.

This is ByteDance’s second major video generation release in recent weeks, following [Seedance 2.0’s unified multimodal approach](#). Where Seedance focused on multi-input control and audio-video fusion, Helios targets raw generation speed and long-form coherence, two problems the field has treated as fundamentally at odds with scale.

Sources:

- [Helios: Real Real-Time Long Video Generation Model \(arXiv\)](#)
- [Helios GitHub Repository](#)
- [Helios on Hugging Face Papers](#)
- [Helios Project Page](#)

---

*Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. **Read more:** [/terms-of-service](#)*