Local AI Video Generation: What Works in 2026

February 5, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Local AI video generation went from 'barely functional' to 'genuinely useful' in about 12 months. The best options right now: Wan 2.1/2.2 (best quality, 1.3B runs on 8GB, 14B needs 12GB+ with GGUF quantization), LTX-Video (fastest — 5-second clip in 4 seconds on an RTX 4090), and HunyuanVideo 1.5 (best faces, needs 16GB+). You need at least 12GB VRAM for anything worth keeping. 24GB (RTX 3090/4090) is the sweet spot — every major model runs. Quality is behind Runway Gen-4 and Sora 2, but Wan 2.2 14B is competitive for social media and B-roll. Everything runs through ComfyUI. It's slow — expect 4-15 minutes per clip except LTX-Video — but it's free after hardware costs and improving fast.

📚 **More on this topic:** ComfyUI vs Automatic1111 vs Fooocus · VRAM Requirements Guide · What Can You Run on 24GB VRAM · Stable Diffusion Locally · Planning Tool

A year ago, local AI video generation was a novelty — 2-second clips at 480p with visible artifacts, taking 30 minutes to render. You'd show someone and say "isn't that cool?" and they'd politely agree while looking at a melting face.

That's not where we are anymore. Wan 2.2 generates coherent 5-second clips with smooth human motion. LTX-Video produces clips faster than real-time. HunyuanVideo 1.5 handles faces better than most cloud services. And all of it runs on hardware you can buy for under $2,000.

It's still early. It's still slow (mostly). The quality gap with top-tier cloud services is real. But if you have a GPU with 12GB+ VRAM and some patience, you can generate video locally that would have cost hundreds of dollars in cloud credits a year ago.

Here's what actually works, what you actually need, and what's still painful.

The Models: What's Available Right Now

Seven models matter for local video generation in early 2026. They vary wildly in quality, speed, VRAM needs, and what they're good at.

Wan 2.1 / 2.2 (Alibaba) — Best Overall

The current king of open-source video generation. Apache 2.0 licensed, meaning full commercial use.

Spec	1.3B Model	14B Model
Resolution	480p	480p / 720p
Frame rate	24 FPS	24 FPS
Duration	~5 seconds	~5 seconds
VRAM (standard)	~8GB	~24GB+
VRAM (GGUF Q5)	N/A	~11GB
VRAM (GGUF Q3/Q4)	N/A	~7-10GB
Speed, RTX 4090	~4 minutes	~4-8 minutes
Capabilities	Text-to-video	Text-to-video, image-to-video

The 1.3B model is the entry point — genuine video generation on 8GB VRAM. The results are limited to 480p and lack fine detail, but the motion is smooth and coherent. For a model that fits on an RTX 4060, that's remarkable.

The 14B model is where Wan gets serious. At full precision it needs 24GB+, but GGUF quantization (thank you, llama.cpp community) brings it down to 11GB at Q5 with minimal quality loss. That means a 14-billion-parameter video model runs on an RTX 3060 12GB. A year ago this would have been absurd.

Wan 2.2 adds a Mixture-of-Experts variant (A14B) and improved image-to-video. Quality leads the open-source field in human motion consistency, texture detail, and prompt adherence.

Best for: Everything. If you're picking one model, pick this one.

LTX-Video / LTX-2 (Lightricks) — Fastest

The speed champion. Nothing else comes close.

Spec	LTX-Video 2B	LTX-2
Resolution	768x512	Up to 4K (with upscaler)
Frame rate	24 FPS	24-50 FPS

Spec	LTX-Video 2B	LTX-2
Duration	~5 seconds	~5 seconds
VRAM (FP16)	~12GB	~16GB
VRAM (FP8)	~8GB	~12GB
Speed, RTX 4090 (768x512)	~4 seconds	~4-11 seconds
Speed, RTX 4090 (1216x704)	N/A	~2 minutes

Read that speed line again. A 5-second video clip generated in 4 seconds. On a single consumer GPU. That's faster than real-time.

This changes the workflow completely. With every other model, you type a prompt, wait 5-15 minutes, and hope you got it right. With LTX-Video, you iterate. Try a prompt, see the result in seconds, adjust, try again. It's closer to how you'd use Stable Diffusion for images.

The quality trade-off is real — LTX-Video's output doesn't match Wan 14B or HunyuanVideo for fine detail. But the built-in 4K spatial upscaler in LTX-2 helps close the gap. NVIDIA announced NVFP4/NVFP8 optimizations specifically for LTX-2 at CES 2026, promising 3x faster generation and 60% less VRAM on RTX 40/50 series cards.

Best for: Rapid iteration, concept prototyping, quantity over maximum quality.

HunyuanVideo 1.5 (Tencent) — Best Faces

The cinematic quality leader, especially for anything involving human faces.

Spec	HunyuanVideo 1.0	HunyuanVideo 1.5
Parameters	13B	8.3B
Resolution	720p	720p
Duration	5-10 seconds	5-10 seconds (up to 121 frames)
VRAM (standard)	40GB+	~24GB
VRAM (with offloading)	~24GB	~14GB
Speed, RTX 4090	5-10 minutes	~75 seconds (distilled) / 3-12 min (standard)

HunyuanVideo 1.0 was impressive but impractical — 40GB+ VRAM meant consumer GPUs were out. Version 1.5 was a breakthrough: they cut parameters from 13B to 8.3B, added sparse attention for 1.87x speedup, and brought VRAM down to 14GB with offloading.

The quality, especially on faces, is exceptional. Multiple characters in a scene, natural expressions, minimal artifacts. If your video involves people, HunyuanVideo is the model to try first.

The step-distilled variant generates in about 75 seconds on an RTX 4090 — not LTX-Video fast, but far from the 10-minute waits of earlier models.

Best for: Cinematic content, anything with human faces, multi-character scenes.

CogVideoX (Tsinghua/ZhipuAI) — Best Image-to-Video

Spec	2B Model	5B Model
Resolution	720x480	720x480
Frame rate	8 FPS	8 FPS
Duration	6 seconds	6 seconds
VRAM (standard)	~12GB	~16-18GB
VRAM (with offloading)	~8GB	~12GB
Speed, RTX 4090	~5-8 minutes	~15 minutes

CogVideoX has the best image-to-video mode among open models. Generate a hero image with Flux or SDXL, then animate it with CogVideoX. The 3D Causal VAE technology delivers strong detail preservation.

The downsides: 8 FPS output looks noticeably choppy compared to 24 FPS competitors, and the 15-minute generation time for the 5B model tests your patience. CogVideoX 1.5 pushes resolution to 1360x768 but needs even more VRAM.

Excellent ComfyUI support through the `kijai/ComfyUI-CogVideoXWrapper` node, which handles T2V, I2V, LoRA, and GGUF quantization.

Best for: Image-to-video animation, scientific/technical content.

AnimateDiff — Most Flexible (but Aging)

Spec	SD 1.5	SDXL
Resolution	512x512 typical	1024x768
Frames	16 (up to 24)	16

Spec	SD 1.5	SDXL
VRAM (minimum)	~8GB (with tricks)	~14GB
Speed, RTX 4090	1-3 minutes	2-5 minutes

AnimateDiff doesn't generate video from scratch — it adds a temporal motion module on top of existing Stable Diffusion checkpoints. This means you can use any of the thousands of community SD 1.5 or SDXL models and LoRAs, and AnimateDiff adds motion.

That flexibility is its superpower and its limitation. You get access to the massive SD ecosystem of styles, but the output quality is fundamentally limited by the base SD architecture. Motion can look jittery or looping rather than cinematic. The technique is showing its age against purpose-built video diffusion models.

Still worth knowing about if you're already deep in the SD ecosystem and want to animate your existing workflows without learning a new model.

Best for: Stylized animations, leveraging existing SD checkpoints and LoRAs.

Mochi 1 (Genmo) — The Pioneer

Spec	Value
Resolution	480p (848x480)
Frame rate	30 FPS
Duration	~5.4 seconds (163 frames)
VRAM (ComfyUI optimized)	~20GB
VRAM (quantized)	~17-18GB
Speed, RTX 4090	~5 min (49 frames) / ~20-30 min (163 frames)

Mochi 1 was the first truly capable open-source text-to-video model (Apache 2.0). It broke new ground in late 2024 and early 2025. Strong prompt adherence, good motion quality, and the first model that made people take local video generation seriously.

It's been surpassed. Wan 2.2 produces better quality at lower VRAM. HunyuanVideo handles faces better. LTX-Video is orders of magnitude faster. Mochi's 480p resolution and 20GB+ VRAM requirement make it hard to recommend over the newer options.

Best for: Historical interest. Use Wan 2.2 instead.

Stable Video Diffusion (SVD / SVD-XT) — Legacy

Spec	SVD	SVD-XT
Type	Image-to-video only	Image-to-video only
Resolution	1024x576	1024x576
Frames	14	25
VRAM (ComfyUI)	<10GB	<10GB

SVD was Stability AI's video model. Image-to-video only — no text-to-video. 14 or 25 frames of subtle camera motion and scene animation. Stability removed it from their API in August 2025 and shifted focus elsewhere. The weights are still on HuggingFace if you want to try it, but the community has moved on.

Best for: Simple image animation if you're already on a very tight VRAM budget.

What You Can Actually Run: The VRAM Reality Check

This is the section most guides skip. Here's the honest truth about what works at each VRAM tier.

8GB VRAM (RTX 4060, RTX 3060 8GB)

Model	Works?	What You Get
Wan 2.1 1.3B	Yes	480p, 5 sec, ~4-6 min
LTX-Video (FP8)	Yes	512x512, ~50 frames, seconds
AnimateDiff (SD 1.5)	Yes, with tricks	512x512, 8-16 frames, 15-20 min
Wan 14B (GGUF Q2/Q3)	Barely	480p, very slow, noticeable quality loss
Everything else	No	Not enough VRAM

The reality: You can generate video on 8GB. Wan 1.3B and LTX-Video are the viable options. The results are real but modest — 480p with limited detail. This is the absolute floor for local video generation. If you're on 8GB and want to experiment, start here. If you want results you'd actually use for something, you need more VRAM.

12GB VRAM (RTX 3060 12GB, RTX 4070)

Model	Works?	What You Get
Wan 14B (GGUF Q5/Q6)	Yes	480p, good quality, 10-15 min
CogVideoX 2B	Yes	720x480, 6 sec
LTX-Video (FP8)	Yes	768x512, fast
Wan 1.3B	Yes, comfortably	480p, 5 sec
AnimateDiff	Yes	512x512, 16+ frames
SVD-XT	Yes	1024x576, 25 frames
HunyuanVideo 1.5	Marginal	Needs aggressive offloading, slow

The reality: A major step up from 8GB. The GGUF-quantized Wan 14B is the standout — a 14-billion-parameter video model producing real quality on a $180 used GPU. This is the tier where local video generation starts being genuinely useful rather than just a tech demo. The RTX 3060 12GB remains the best budget entry point.

16GB VRAM (RTX 4060 Ti 16GB, RTX 4080)

Model	Works?	What You Get
HunyuanVideo 1.5	Yes (with offloading)	720p, 121 frames, 3-12 min
CogVideoX 5B	Yes	720x480, 6 sec, ~15 min
LTX-2 (FP16)	Yes	Full quality, near real-time
Wan 14B (GGUF Q6/Q8)	Yes	720p, good quality
All 8GB/12GB models	Yes, comfortably	Better quality, faster

The reality: This is where things get genuinely good. HunyuanVideo 1.5 becomes accessible — 720p with the best face rendering in open source. LTX-2 runs at full quality. Higher-quality GGUF quants of Wan 14B are available. If you're buying hardware specifically for video generation, 16GB is the minimum to aim for.

24GB VRAM (RTX 3090, RTX 4090)

Model	Works?	What You Get
Everything	Yes	Full quality, best speeds

Model	Works?	What You Get
Wan 14B (FP16)	Yes	Full precision 720p
HunyuanVideo 1.5	Yes, comfortably	720p, 121 frames
Mochi 1	Yes	480p, 163 frames
LTX-2	Yes	4K with upscaler, near real-time

The reality: The sweet spot. Every model runs without painful compromises. You get full-precision weights, higher resolutions, and reasonable generation times. The RTX 4090 is about 40-70% faster than the RTX 3090 despite the same 24GB — faster CUDA cores and better memory bandwidth. But the used RTX 3090 at ~$700 is the bang-for-buck winner.

48GB+ (Multi-GPU, Mac Unified Memory)

At 48GB+ you run unquantized 14B models at full precision, generate longer clips, and push higher resolutions without offloading overhead. Mac M4 Max with 64-128GB unified memory handles everything but with slower compute than a dedicated NVIDIA GPU. This tier is for people who want zero compromises.

Honest Comparison: Local vs Cloud Services

Let's not pretend local video generation matches the best cloud services. It doesn't. But let's also not pretend the gap is as wide as it was six months ago.

The Cloud Landscape

Service	Quality	Pricing	Best Feature
Runway Gen-4	Top tier	$12-28/month (52 sec - 3 min of video)	Camera/scene controls, professional tooling
Sora 2 (OpenAI)	Top tier	Requires ChatGPT Plus ($20/month)	Physics understanding, photorealism
Kling 2.6	Excellent	Free tier (66 credits/day), $10-92/month	Synchronized audio, generous free tier
Pika	Good (stylized)	Free tier (150 credits/month), $8-76/month	Best value entry point

Where Cloud Wins

- **Peak quality.** Sora 2, Runway Gen-4, and Kling 2.6 produce more consistently photorealistic results than any local model. The gap is narrowing but real.
- **Audio.** Kling 2.6 generates synchronized voiceover, dialogue, and sound effects natively. No local model does this.
- **Speed of use.** Cloud services return results in 10-60 seconds. Local models (except LTX-Video) take 4-15 minutes.
- **No hardware investment.** No $700-1,600 GPU purchase required.

Where Local Wins

- **Cost at volume.** After hardware, generation is free. A creator making 50+ clips per day would spend hundreds per month on Runway. Locally: electricity.
- **Privacy.** Nothing leaves your machine. No content policy filters. No terms of service.
- **No limits.** No monthly credit caps. No watermarks. No content restrictions.
- **Customization.** LoRA training, model merging, ComfyUI workflow pipelines — you control every parameter.

The Honest Verdict

Wan 2.2 14B and HunyuanVideo 1.5 are competitive with mid-tier cloud offerings. For social media clips, B-roll footage, concept prototyping, and creative experiments, local is already “good enough.” For professional commercial content or anything requiring photorealistic human faces with complex interactions, cloud services still lead — but the margin is shrinking fast.

ComfyUI: The Video Generation Hub

Every serious local video model runs through ComfyUI. It’s become the universal interface for video generation, with official or community-maintained nodes for every model.

Which Models Have ComfyUI Support

Model	ComfyUI Node	Quality of Support
Wan 2.1/2.2	Native + Wan2GP	Excellent — official workflows, best-supported
LTX-Video / LTX-2	`ComfyUI-LTXVideo` (official)	Excellent — day-1 support, NVIDIA optimized

Model	ComfyUI Node	Quality of Support
HunyuanVideo	Community nodes	Good
CogVideoX	`ComfyUI-CogVideoXWrapper`	Mature — T2V, I2V, LoRA, GGUF
AnimateDiff	`ComfyUI-AnimateDiff-Evolved`	Very mature ecosystem
Mochi 1	Official nodes	Works, less actively maintained

Useful Workflows

The rapid iteration pipeline (LTX-2): Generate at 768x512 in seconds, iterate on prompts until you have what you want, then upscale winners with the built-in 4K spatial upscaler. The fastest path from idea to watchable video.

The quality pipeline (Wan 2.2 + upscaling): Generate at 480p/720p with Wan 2.2, upscale with RealESRGAN 4x, interpolate frames with RIFE or GIMM-VFI. More steps, better results.

The image-to-video pipeline (Flux + CogVideoX): Generate a hero image with Flux or SDXL, then animate it with CogVideoX's I2V mode. Great for bringing still images to life with controlled motion.

Beyond ComfyUI

- **Wan2GP** (deepbeepmeep): Standalone web UI supporting Wan 2.1/2.2, HunyuanVideo, and LTX-Video. Five memory profiles for different hardware tiers. Simpler than ComfyUI if you just want to generate video without building node graphs.
- **Pinokio**: One-click installer for ComfyUI, Wan2GP, and other tools. Good for beginners who don't want to deal with git clones and Python environments.

Speed Expectations: Don't Sugarcoat It

This is where most guides get dishonest. Here are real generation times for a ~5-second clip.

RTX 4090 (24GB)

Model	Resolution	Time
LTX-Video 2B	768x512	**4-11 seconds**
HunyuanVideo 1.5 (distilled)	720p	**~75 seconds**

Model	Resolution	Time
Wan 2.1 1.3B	480p	**~4 minutes**
Wan 2.1 14B	720p	**~4-8 minutes**
CogVideoX 5B	720x480	**~15 minutes**
Mochi 1	480p	**~5-20 minutes**

RTX 3090 (24GB)

Add 40-50% to all RTX 4090 times. A Wan 14B clip that takes 6 minutes on a 4090 takes about 9 minutes on a 3090.

RTX 3060 12GB

GGUF-quantized models only. Wan 14B Q5 at 480p: roughly 10-15 minutes per clip. LTX-Video FP8: still seconds.

The Reality

LTX-Video is the outlier — genuinely fast enough for iterative workflows. Everything else requires patience. You're not going to sit there generating clip after clip like you would with cloud services. The workflow is more like: carefully craft your prompt, hit generate, go make coffee, come back and evaluate.

For comparison, cloud services return results in 10-60 seconds. The only local model that approaches cloud speed is LTX-Video.

This is a genuine limitation, not something you can optimize away. Video generation is computationally heavier than image generation by orders of magnitude. A 5-second video at 24 FPS is 120 frames — each requiring denoising passes through a multi-billion parameter model. The physics are what they are.

Best Setup for the Money

Budget: Under $300

- **GPU:** Used RTX 3060 12GB (~$180)
- **Models:** Wan 2.1 1.3B, Wan 14B GGUF Q4-Q5, LTX-Video FP8

- **What you get:** Functional 480p video generation, fast iteration with LTX-Video
- **Honest assessment:** You can generate video. Results are usable for social media and experiments. Fine detail is limited.

Sweet Spot: $700-800

- **GPU:** Used RTX 3090 24GB (~$700)
- **Models:** Everything runs — Wan 14B, HunyuanVideo 1.5, LTX-2, CogVideoX 5B
- **What you get:** Full-quality 720p from every model, reasonable generation times
- **Honest assessment:** This is the setup to buy. Every current model runs without painful compromises. The 3090 is 3+ years old but 24GB of VRAM is 24GB of VRAM.

Maximum Performance: $1,600+

- **GPU:** RTX 4090 (~$1,600)
- **Models:** Everything, 40-70% faster than RTX 3090
- **What you get:** Near real-time with LTX-2, 75-second HunyuanVideo distilled clips, faster iteration across the board
- **Honest assessment:** The speed bump over the 3090 is significant if you're generating a lot of video. If you're experimenting occasionally, the 3090 is a better value.

Where This Is Heading

The pace of improvement in 2025 was extraordinary:

- **January 2025:** Best local option was CogVideoX at 720x480/8fps. Experimental at best.
- **February 2025:** Wan 2.1 released, immediately set a new quality standard.
- **Mid 2025:** GGUF quantization brought 14B models to 12GB cards.
- **Late 2025:** HunyuanVideo 1.5 cut VRAM by 40% while improving quality. LTX-Video proved real-time generation was possible.
- **January 2026:** NVIDIA announced NVFP4/NVFP8 optimizations at CES, promising 60% less VRAM and 3x speed on RTX 40/50 series.

In 12 months, local video generation went from "technically possible but painful" to "genuinely useful on mainstream hardware."

What’s Coming

- **Longer videos.** Current models cap at 5-10 seconds. LTX-Video 13B and Wan 2.2 extensions are pushing toward 30-60 seconds.
- **Native audio.** Cloud models like Kling already do synchronized audio. Local models will follow.
- **RTX 5090.** 32GB VRAM with NVFP4 support — will run full-precision 14B models comfortably and make quantization less necessary for most users.
- **Better distillation.** HunyuanVideo’s 75-second generation time shows what distilled models can do. Expect every major model to get a fast variant.

When Will Local Be “Good Enough”?

It depends on what you’re making:

Use Case	Ready Now?
Social media clips, memes	Yes — Wan 2.2, LTX-2
B-roll, concept visualization	Yes — Wan 14B, HunyuanVideo
Product demos, prototypes	Mostly — may need cloud for final render
Professional ads, commercial content	Not yet — late 2026 with next-gen models
Cinematic/film quality	Not yet — 2027-2028 realistically

The most likely future: local models handle iteration, drafting, and high-volume generation. Cloud services handle final renders for premium content. Hybrid workflows, just like how professionals use local Stable Diffusion for concept art and Midjourney for client-ready finals.

The Bottom Line

Local AI video generation is real, useful, and improving faster than any other area of consumer AI. A used RTX 3090 and ComfyUI gives you access to every major model. Wan 2.2 for quality, LTX-Video for speed, HunyuanVideo for faces.

Is it as good as Runway or Sora? No. Not yet. But it’s free after hardware, private, unlimited, and uncensored. And the gap is closing fast enough that what’s “not quite there” today will likely be competitive by the end of the year.

If you’ve been waiting for local video generation to be worth trying — it is now.

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/local-ai-video-generation/

Free guides for running AI locally