What Can You Actually Run on 16GB VRAM?

January 30, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** 16GB runs 13B-14B models at Q4-Q6 as the sweet spot (22-53 tok/s), 7B-8B at near-lossless Q8 with massive context, and 20B-22B models at Q4 with short context. Flux runs well at FP8, SDXL is comfortable with the refiner. You can't run 32B models practically or 70B at all. It's a meaningful upgrade over 12GB but noticeably behind 24GB. The RTX 5060 Ti 16GB (~$429) is the best new card at this tier; a used RTX 3090 (~$700-850) is better if you can stretch the budget.

📚 **More on this topic:** VRAM Requirements · What Can You Run on 12GB · What Can You Run on 24GB · Quantization Explained

You have 16GB of VRAM. Maybe it's an RTX 5060 Ti, a 4060 Ti 16GB, a 4080 Super, or an AMD RX 7800 XT. You know 16GB is more than 12GB. The question is: how much more?

The honest answer: meaningfully more, but not as much as you might hope. 16GB is the awkward middle child of local AI — clearly better than 12GB, clearly behind 24GB, and in a strange position where the models that fit well are excellent but the next tier up is just out of reach. This guide covers exactly what that means in practice.

Who This Is For

GPU	VRAM	Bandwidth	Street Price (Jan 2026)
RTX 5060 Ti 16GB	16GB GDDR7	448 GB/s	~$429-459
RTX 4060 Ti 16GB	16GB GDDR6	288 GB/s	~$270 used, ~$430 new
RTX 4080 Super	16GB GDDR6X	736 GB/s	~$1,000-1,200
AMD RX 7800 XT	16GB GDDR6	~500-530 GB/s	~$430-557

These cards share the same VRAM ceiling but have very different performance. The RTX 4080 Super generates tokens roughly 2.3x faster than the 4060 Ti 16GB thanks to bandwidth — they just happen to load the same models. The RTX 5060 Ti 16GB sits in between, offering 50% more bandwidth than the 4060 Ti for nearly the same price.

If you're choosing between these cards, bandwidth matters more than you'd think for LLM inference. But VRAM capacity determines what you can run at all, and that's what this guide focuses on.

The Awkward Middle Tier

Here's the honest positioning of 16GB:

What you gain over 12GB:

- 13B-14B models at Q6-Q8 instead of just Q4
- 20B-22B models fit (they don't at 12GB)
- Longer context windows — 16K+ on 14B models vs. 8K at 12GB
- Flux image generation at FP8 with actual headroom (12GB is razor-thin)
- SDXL with base + refiner comfortably, with room for LoRAs

What you still can't do that 24GB can:

- Run 27B-32B models (weights alone exceed 16GB at Q4)
- Use 70B models at any usable quality
- Fine-tune anything larger than ~7B
- Run multiple large models simultaneously
- Use Flux at full FP16 precision

Sixteen gigabytes is where you stop squeezing 7B-8B models and start running genuinely larger ones. It's not where you run everything — that's 24GB.

What Runs Well on 16GB

7B-8B at Q8/FP16: Maximum Quality

With 16GB, you can stop quantizing 7B-8B models aggressively and run them at near-original quality. At Q8, a 7B model uses ~8GB for weights, leaving 8GB for the KV cache — enough for 32K-65K+ token context windows.

Model	Quant	VRAM (Weights)	Context Room	Speed (RTX 5060 Ti)
Llama 3.1 8B	Q8	~8 GB	~8 GB free → 32K-53K tokens	~38 tok/s
Qwen 2.5 7B	Q8	~7.5 GB	~8.5 GB free → 32K+ tokens	~40 tok/s
Mistral 7B	Q8	~7 GB	~9 GB free → 40K+ tokens	~42 tok/s
Llama 3.1 8B	FP16	~14 GB	~2 GB free → 4K-8K tokens	~22 tok/s

At Q8, quality is essentially lossless — perplexity increase is roughly 0.01 points versus FP16. This is where 16GB shines for 7B models: no quality compromise and massive context windows.

FP16 fits but leaves almost no headroom. Use it only if you need bit-perfect output and can live with short context.

13B-14B at Q4-Q8: The Sweet Spot

This is the reason to own 16GB. The 13B-14B class is a major quality jump over 7B-8B, and 16GB handles it comfortably at Q4-Q6 — something 12GB can do at Q4 but struggles with at higher quants.

Model	Quant	VRAM	Context	Speed (5060 Ti)	Speed (4080 Super)
Qwen 2.5 14B	Q4_K_M	~9 GB	8K-16K	~33 tok/s	~53 tok/s
Qwen 2.5 14B	Q6_K	~11 GB	4K-8K	~28 tok/s	~45 tok/s
Qwen 2.5 14B	Q8_0	~13 GB	2K-4K	~22 tok/s	~38 tok/s
Phi-4 14B	Q4_K_M	~9 GB	8K-16K	~33 tok/s	~53 tok/s
Gemma 3 12B	Q4_K_M	~8 GB	8K-16K	~35 tok/s	~55 tok/s
DeepSeek R1 14B	Q4_K_M	~9 GB	8K-16K	~32 tok/s	~50 tok/s

The sweet spot here is Q4_K_M with 8K-16K context. That gives you the best balance of quality, speed, and usable context length. If you want higher quality and can accept shorter conversations, Q6 is excellent — and it's something 12GB cards can't do comfortably with 14B models.

20B-22B at Q4: The Upper Limit

This is what 16GB unlocks that 12GB can't touch. Models like Codestral 22B and MoE architectures become accessible:

Model	Quant	VRAM	Context	Notes
Codestral 22B	Q4_K_M	~13 GB	2K-4K	Best code model at this tier
Codestral 22B	Q5_K_M	~15 GB	1K-2K	Barely fits, tight context
GPT-OSS 20B (MoE)	Q4	~12 GB	16K-65K	Sparse MoE — faster than expected

Codestral 22B at Q4 is a highlight. It's one of the strongest open-source coding models, and 16GB is the minimum to run it on GPU. Context is limited to a few thousand tokens, so it's best for code completion and function generation rather than processing entire codebases.

MoE (Mixture-of-Experts) models deserve special mention. They have high total parameters but only activate a fraction per token. The GPT-OSS 20B runs at 82 tok/s on the RTX 5060 Ti at 16K context — faster than dense 14B models despite being "bigger."

What's Possible But Painful

32B at Q3: Technically Loads, Barely Usable

A 32B model at Q4_K_M needs ~19-20GB — it doesn't fit. At Q3_K_S (~14GB) or IQ3_M (~13-14GB), the weights fit, but you're left with 2-3GB for the KV cache.

That means:

- Context limited to roughly 1K-4K tokens
- Quality degraded by aggressive quantization
- Speeds of ~10-15 tok/s (GPU-only) or ~5-10 tok/s (with CPU offloading)

Verdict: A 14B model at Q4-Q6 will produce better output, faster, with longer context. The 32B squeeze is a party trick, not a daily driver. For real 32B inference, you need 24GB.

70B: Not Happening

Even at the most aggressive 2-bit quantization (IQ2_XS), a 70B model's weights approach 15-17GB — leaving nothing for context. With heavy CPU offloading, you might see 1-3 tok/s at severely degraded quality.

A Qwen 2.5 14B at Q4 will beat a 70B at Q2 on most tasks and run 10x faster. Don't chase parameter counts when the quantization destroys the quality advantage.

What Won't Work

- **32B+ models at Q4+**: Weights alone exceed 16GB. Not without heavy CPU offloading.
- **70B models**: Even at Q2, it's not practical.
- **Fine-tuning 14B+ models**: LoRA on 7B fits at FP16. LoRA on 14B is tight. Anything larger needs 24GB.
- **Multiple models simultaneously**: One large model at a time. Loading a second will OOM.
- **128K+ context on 14B**: The KV cache alone would need more than 16GB.

Image Generation on 16GB

16GB is a genuine upgrade over 12GB for image generation. Several models that are tight or impossible at 12GB become comfortable.

SDXL: Comfortable

SDXL base uses ~7-8GB. On 16GB, you can run base + refiner simultaneously (~10-12GB total) with room left for LoRAs. No memory-management hacks needed — just generate.

Setup	VRAM Used	Works on 12GB?	Works on 16GB?
SDXL base only	~7-8 GB	Yes	Yes
SDXL base + refiner	~10-12 GB	Tight	Comfortable
SDXL + refiner + LoRAs	~12-14 GB	No	Yes

Flux: FP8 Is the Move

Flux Dev at FP16 needs ~24GB — it won't fit. At FP8, it drops to ~12GB and runs well on 16GB with headroom to spare.

Variant	Precision	VRAM	Speed (4080 Super)
Flux Dev	FP8	~12 GB	~55 seconds (50 steps)
Flux Schnell	FP8	~12 GB	A few seconds (4 steps)
Flux Dev	FP16	~24 GB	Doesn't fit

FP8 delivers nearly identical quality to FP16 with ~40% faster generation. Flux Schnell at FP8 is particularly impressive — 1-4 steps for near-instant images. On 12GB, Flux FP8 works but with zero headroom; 16GB gives you room to breathe.

SD 3.5 Large: Fits with Quantization

SD 3.5 Large at full FP16 needs ~18GB — too much. But with NVIDIA's TensorRT FP8 optimization, it drops to ~11GB with minimal quality loss and a 2.3x speed boost. GGUF Q8 versions also fit at ~16GB.

This is a clear 16GB advantage over 12GB: SD 3.5 Large at Q8 simply doesn't fit on a 12GB card. For the best Stable Diffusion output quality available, 16GB is the practical minimum.

Best Models for 16GB GPUs (Ranked)

1. Qwen 2.5 14B Instruct — The all-rounder for this tier. Strong at everything, excellent multilingual and coding. This is your default.

```
ollama pull qwen2.5:14b
```

2. Llama 3.1 8B Instruct (Q8) — When you want maximum quality on a smaller model with long context. Q8 on 16GB gives you lossless quality and 32K+ context.

```
ollama pull llama3.1:8b
```

3. DeepSeek R1 14B — Best for reasoning and math at this tier. Uses chain-of-thought to improve complex answers.

```
ollama pull deepseek-r1:14b
```

4. Codestral 22B — The best coding model you can run on 16GB. Context is limited (~2-4K), so use it for function-level work rather than full-file analysis.

```
ollama pull codestral:22b
```

5. Phi-4 14B — Microsoft's strong reasoning model. Competitive with Qwen 2.5 14B on benchmarks, slightly better on some reasoning tasks.

```
ollama pull phi4:14b
```

6. Gemma 3 12B — Google's efficient option. Strong instruction following and a slightly smaller footprint than 14B models, giving more context headroom.

```
ollama pull gemma3:12b
```

New to local AI? Ollama handles everything — one command to install, one to run. If you prefer a visual interface, LM Studio works just as well.

→ Check what fits your hardware with our Planning Tool.

Tips to Maximize 16GB

1. Pick the Right Quantization

On 16GB, you have real choices:

Model Size	Recommended Quant	Why
7B-8B	Q8_0	Essentially lossless, leaves 8GB+ for context
13B-14B	Q4_K_M or Q5_K_M	Best balance of quality and context room
20B-22B	Q4_K_M	Only option that fits with usable context

Don't default to Q4 for everything. If the model is small enough, use Q6 or Q8. The quality difference is noticeable, and you have the VRAM for it.

2. Use Flash Attention

Flash Attention reduces KV cache memory usage significantly. Most modern inference engines (llama.cpp, vLLM) support it. On 16GB, this can mean the difference between 8K and 16K context on a 14B model.

3. Try KV Cache Quantization

Newer versions of llama.cpp support Q8 and Q4 KV cache, which roughly halves or quarters the memory used by context. With Q8 KV cache on a 14B model at Q4 weights, you can push context from ~16K to ~32K tokens.

```
# llama.cpp with Q8 KV cache
./llama-server -m model.gguf -ctk q8_0 -ctv q8_0
```

4. Close VRAM Hogs

The same advice from every other tier: Chrome's hardware acceleration, video players, and game launchers consume VRAM. Close them before loading large models.

5. Monitor VRAM Usage

```
nvidia-smi -l 1
```

If you're hovering near 15-16GB, you're on the edge. Drop one quantization level or reduce context length.

Is 16GB Worth It? (Honest Take)

This is the question everyone with a 16GB card — or considering one — needs answered.

16GB vs. 12GB: What You Gained

Capability	12GB	16GB
7B-8B at Q8	Yes, but tight context	Yes, with massive context

Capability	12GB	16GB
13B-14B at Q4	Sweet spot	Sweet spot, with room for Q5-Q6
13B-14B at Q8	Doesn't fit	Fits (2-4K context)
20B-22B at Q4	Doesn't fit	Fits (2-4K context)
Flux FP8	Razor-thin	Comfortable
SD 3.5 Large Q8	Doesn't fit	Fits

The jump from 12GB to 16GB is real but incremental. You don't unlock a new class of models the way 24GB does — you get more headroom on the same class (13B-14B) and access to a few larger models (20B-22B) with tight context.

16GB vs. 24GB: What You're Missing

Capability	16GB	24GB
27B-32B at Q4	Doesn't fit	Sweet spot
70B at Q3-Q4	Doesn't fit	Tight but works
Fine-tuning 7B	Tight	Comfortable
Fine-tuning 14B	Doesn't fit	Fits
Flux FP16	Doesn't fit	Native
14B at Q4, 32K+ context	Doesn't fit	Yes

The jump from 16GB to 24GB is the bigger one. Eight extra gigabytes unlocks 27B-32B models — the quality tier where local models start competing with commercial APIs. If your budget allows it, 24GB is the better investment.

When 16GB Is the Right Call

- **You already own a 16GB card.** Don't upgrade for the sake of upgrading. 16GB runs excellent 14B models at interactive speeds.
- **You want a new card under $500.** The RTX 5060 Ti 16GB at ~$429 is the best value in this range. A used RTX 3090 is better for AI but costs more, draws 350W, and is five years old.
- **You need a card for gaming AND AI.** The 5060 Ti and 4080 Super are excellent gaming cards. A used 3090 is powerful but loud, hot, and huge.

- **You run 7B-14B models daily.** These run perfectly at 16GB with the quality and context you need. If you're not constantly reaching for 32B+ models, 16GB is enough.

When 16GB Is Not Enough

- **You need 32B+ models.** Qwen 2.5 32B, DeepSeek R1 32B, Gemma 3 27B — these need 24GB.
- **You want to fine-tune models.** LoRA on 14B barely fits at 16GB. Real fine-tuning needs headroom.
- **You run complex image generation workflows.** Multi-model ComfyUI pipelines with ControlNet, IP-Adapter, and LoRAs can exceed 16GB.
- **You want to future-proof.** Models keep getting bigger. 16GB is comfortable today but may feel tight in a year.

When to Upgrade (And What To)

You've outgrown 16GB when:

- You're constantly choosing smaller models or lower quants to fit
- You need 32B models for your work
- You want longer context on 14B+ models
- Image generation workflows keep hitting OOM

The upgrade path:

GPU	VRAM	Street Price (Jan 2026)	What It Unlocks	Best For
Used RTX 3090	24GB	~$700-850	32B at Q4-Q5, 70B at Q3, fine-tuning	VRAM king on a budget
RTX 4090	24GB	~$1,800-2,200	Same models, 40-70% faster	Best single consumer GPU
RTX 5090	32GB	~$1,999+	70B at Q4, 32B at Q8	Overkill for most, future-proof

The used RTX 3090 remains the value play. For the price of two 16GB cards, you get 24GB of VRAM and 936 GB/s of bandwidth — over 3x what the 4060 Ti 16GB offers. If AI work is your priority, it's the best dollar-per-capability upgrade available.

The Bottom Line

16GB is a capable, slightly awkward tier. It runs 13B-14B models at quality levels that 12GB can't match, handles Flux and SDXL without workarounds, and gives you just enough room for 20B-22B models in a pinch. It doesn't reach 32B, and it doesn't pretend to.

The practical advice:

1. Install Ollama and pull `qwen2.5:14b`. That's your workhorse at Q4 with 8K-16K context.
2. Run 7B-8B models at Q8 instead of Q4 — you have the VRAM, use it for better quality.
3. Try Codestral 22B for coding tasks. It's tight but worth the squeeze.
4. When 16GB isn't enough — and 32B models are calling — a used RTX 3090 is the move.

Your card is more capable than 12GB and honest about not being 24GB. Use what it does well.

Related Guides

- What Can You Actually Run on 12GB VRAM?
- What Can You Actually Run on 24GB VRAM?
- GPU Buying Guide for Local AI
- What Quantization Actually Means

Sources: Hardware Corner GPU Ranking for LLMs, Hardware Corner RTX 5060 Ti Analysis, LocalScore.ai RX 7800 XT, Puget Systems LLM Inference Benchmarks, Civitai Flux FP8 vs FP16, BestValueGPU Price Tracker

---