What Can You Actually Run on 8GB VRAM?

January 28, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Qwen 3.5 9B at Q4_K_M is now the default pick for 8GB GPUs — 5.68GB file size, ~7GB VRAM with 8K context, native vision, and it beats models 3x its size on reasoning benchmarks. Stable Diffusion 1.5 runs great, SDXL is tight but doable. You won't fit anything above 13B. Start with `ollama run qwen3.5:9b`, close Chrome, and keep context to 4-8K. When you hit the wall, a used RTX 3090 (24GB, ~$700-850) is the best upgrade.

More on this topic: VRAM Requirements · What Can You Run on 12GB · What Can You Run on 4GB · Quantization Explained · Qwen 3.5 9B Setup Guide

You have 8GB of VRAM. Maybe it's an RTX 4060, a 3060 Ti, a 3070, or even an older 2080. You've seen people running AI chatbots locally and you're wondering: can I actually do that with my card?

The short answer is yes — with limits. 8GB is the floor for local AI, not the sweet spot. But "the floor" doesn't mean useless. It means you need to know exactly what fits, what doesn't, and how to squeeze every megabyte. That's what this guide covers.

Who This Is For

If you own any of these cards, this guide is for you:

GPU	VRAM	Architecture
RTX 4060	8GB	Ada Lovelace
RTX 3070	8GB	Ampere
RTX 3060 Ti	8GB	Ampere
RTX 2080	8GB	Turing
RTX 2070	8GB	Turing
RTX 2060 Super	8GB	Turing

This is one of the largest GPU audiences out there. The RTX 4060 alone is the best-selling current-gen card. If you already own one, you don't need to spend another dime to start running AI locally. You just need to know what's realistic.

The Honest Truth About 8GB

Here's the deal: 8GB is limiting, but it's limiting the way a studio apartment is limiting. You can absolutely live there — you just can't spread out.

A 9B model at Q4_K_M quantization uses roughly 5.5-6GB for weights, leaving 2-2.5GB for the KV cache (working memory) and system overhead. That's a tight fit, and it means you'll be making tradeoffs — shorter context windows, no room for multiple models, and nothing bigger than 9B running comfortably.

But here's what matters: Qwen 3.5 9B at Q4 on an 8GB GPU produces genuinely useful output at ~38 tokens per second, handles images natively, and scores 82.5 on MMLU-Pro — numbers that were 70B territory a year ago. That's fast enough for interactive chat, coding assistance, and real work. You're not watching paint dry.

What Runs Well on 8GB

7B-9B Models at Q4: The Sweet Spot

This is where 8GB cards shine. A quantized 7B-9B model at Q4 fits with room to breathe, runs fast, and retains ~90-95% of the original model's quality.

Model	VRAM Used (Q4_K_M)	Speed (RTX 4060)	Best For
Qwen 3.5 9B	**~7 GB**	**~38 tok/s**	**Everything — new default pick**
Llama 3.1 8B	~5.5 GB	~42 tok/s	General assistant, writing
Mistral 7B	~4.8 GB	~45 tok/s	Fast chat, summarization
DeepSeek R1 8B	~5.5 GB	~38 tok/s	Reasoning, math

Qwen 3.5 9B changed the math here. It's a dense 9B model (all parameters active), the 5.68GB Q4_K_M file needs ~7GB VRAM with 8K context, and it scores 82.5 on MMLU-Pro and 81.7 on

GPQA Diamond — numbers that beat models three times its size. It also handles images and video natively from the same weights. No separate vision model needed.

The tradeoff: it's a tighter fit than the older 7B models. Llama 3.1 8B and Mistral 7B leave more headroom for context and background apps. If you need longer conversations or can't close Chrome, those are still good picks. But for raw capability per VRAM byte, Qwen 3.5 9B is the new king.

These speeds are real-world numbers with full GPU offload. At 35+ tokens per second, responses feel like a fast typist. You'll be reading slower than the model generates.

Smaller Models at Higher Quality

If you want more VRAM headroom — for longer conversations or running other apps alongside — smaller models punch above their weight:

Model	VRAM Used (Q4_K_M)	Speed (RTX 4060)	Best For
Qwen 3.5 3B	**~2.5 GB**	**~65 tok/s**	**Fast chat, coding, multilingual**
Qwen 3.5 4B	~3.0 GB	~58 tok/s	Step up from 3B, still light
Llama 3.2 3B	~2.5 GB	~65 tok/s	Quick Q&A, simple tasks
Phi-3 Mini (3.8B)	~2.8 GB	~60 tok/s	Reasoning, coding (compact)

Qwen 3.5 3B is part of the same March 2026 small model drop as the 9B. It punches well above its weight for a 3B model — noticeably better at instruction following and coding than Llama 3.2 3B. If you need headroom for long context or want to keep other apps running, start here.

These leave 5GB+ free, which means longer context windows, no VRAM pressure, and snappy responses. They're less capable than the 9B models, but for quick questions and simple coding tasks, they're surprisingly good.

What's Possible But Painful

13B Models at Low Quantization

You can technically squeeze a 13B model into 8GB at Q2 or Q3 quantization. Should you? Usually not.

The numbers tell the story: GPU utilization drops to 25-42% because the model barely fits, inference crawls, and Q2-Q3 quantization degrades quality noticeably. You're running a larger model badly instead of a smaller model well.

When it's worth trying: If you need a specific 13B model for a task where the 7B version doesn't cut it — and you can tolerate slow, lower-quality output. Think of it as a proof-of-concept, not a daily driver.

When to skip it: For everything else. A Llama 3.1 8B at Q4 will outperform a Llama 2 13B at Q2 in most practical tasks, and it'll do it 5x faster.

Context Length Limits

Here's the part nobody warns you about: context windows eat VRAM. With a 7B model at Q4_K_M, the KV cache for an 8K context window needs 2-3GB. On 8GB total, that leaves almost nothing.

In practice, expect:

- **2048-4096 tokens**: Comfortable, fast, no issues
- **8192 tokens**: Tight. May work but monitor for slowdowns
- **16K+**: Forget it. You'll hit OOM (out of memory) errors

For most conversations, 2-4K context is enough. But if you need to process long documents or maintain very long chat histories, 8GB will frustrate you.

What Won't Work

Let's save you the troubleshooting time:

- **30B+ models**: Doesn't fit. Period. Not even at Q2. A 30B model at Q2 still needs ~12GB.
- **70B models**: Even with CPU offloading, inference drops to 1-3 tokens per second. That's not usable — it's a screensaver.
- **Fine-tuning**: Needs 16GB minimum for LoRA on 7B models. Training is out of scope for 8GB.
- **Multiple models simultaneously**: One model at a time. Loading a second will OOM your first.

If any of these are dealbreakers, skip ahead to the upgrade section.

The “runs on anything” option: 1.58-bit models

Two models worth knowing about if you’re exploring the extreme low end:

BitNet b1.58 2B4T — Microsoft’s ternary model uses weights of just {-1, 0, +1}. The non-embedding memory footprint is 0.4GB. It runs on CPU with no GPU required, 2-6x faster than comparable models on x86 processors. The catch: it requires Microsoft’s bitnet.cpp runtime. No Ollama, no LM Studio, no GGUF. Setup is manual. Quality is comparable to other 2B models — useful for simple tasks, not a replacement for 7B+.

Falcon-Edge 1.58-bit — TII’s ternary models come in 1B (665MB) and 3B (999MB). Same idea as BitNet: extreme compression for edge deployment. Benchmarks are lower than standard quantized models at the same parameter count, and they also need a custom runtime (no Ollama). Experimental, but the direction matters — ternary architectures may eventually make 8GB VRAM concerns irrelevant.

Neither model is ready for daily use today. But if you’re curious about where inference efficiency is headed, both are worth a look.

Image Generation on 8GB

Good news: 8GB handles image generation better than you might expect.

Stable Diffusion 1.5: Runs Great

SD 1.5 is the sweet spot for 8GB cards. The model uses ~4GB VRAM, leaving plenty of headroom. Generation is fast and the community ecosystem (LoRAs, checkpoints, ControlNet) is massive.

Resolution	Time (RTX 4060)	Notes
512x512	~5 seconds	Fast, plenty of headroom
768x768	~10 seconds	Still comfortable
1024x1024	~18 seconds	Pushing it, but works

SDXL: Tight But Doable

SDXL produces noticeably better images but uses ~7-8GB VRAM for the base model alone. On 8GB, it works — with some conditions:

- Use **ComfyUI** (more memory-efficient than AUTOMATIC1111)
- Enable a **FP16 VAE** (drops VAE VRAM from ~6GB to under 1GB)
- Enable **xformers** (25-30% speedup, slight memory savings)
- Expect ~30-35 seconds per 1024x1024 image

The refiner model won't fit alongside the base model. Run them sequentially, not simultaneously. For most people, the base model alone produces great results.

Flux: Limited

Flux's schnell variant can technically run on 8GB with aggressive optimization, but it's not a good experience. If Flux is your priority, you need 12GB+.

Best Models for 8GB GPUs (Ranked)

Here's what to install first, in order:

1. Qwen 3.5 9B — The new default. Native vision, 262K context support, beats models 3x its size on reasoning. Tight fit at ~7GB but worth it.

```
ollama run qwen3.5:9b
```

2. Llama 3.1 8B Instruct — More VRAM headroom than Qwen 3.5 9B (~5.5GB). Still a strong all-rounder, especially if you need longer context on 8GB.

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

3. DeepSeek R1 8B — Best for reasoning and math. Uses a "thinking" approach that improves complex answers.

```
ollama pull deepseek-r1:8b
```

4. Mistral 7B — Fastest of the bunch at ~45 tok/s. Great for quick chat when speed matters more than benchmarks.

```
ollama pull mistral
```

5. Qwen 3.5 3B — When you need VRAM headroom. Leaves 5GB+ free for context, background apps, or running alongside image generation.

```
ollama pull qwen3.5:3b
```

New to Ollama? It's one command to install, one command 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 Squeeze Every Megabyte

1. Close Chrome (Seriously)

Chrome with a few tabs open can eat 500MB-1GB of VRAM for hardware acceleration. Close it before running models, or disable hardware acceleration in Chrome settings (`chrome://settings/system`).

2. Use Q4_K_S Instead of Q4_K_M

Q4_K_S is slightly smaller than Q4_K_M (about 200MB less for a 7B model) with minimal quality loss. On 8GB, that 200MB matters. It can be the difference between fitting with headroom and running out of memory.

For the difference between quantization formats, see our quantization guide.

3. Reduce Context Length

If you're hitting memory limits, explicitly set a shorter context window:

```
# Ollama: set context to 2048 tokens
ollama run llama3.1:8b --num-ctx 2048
```

2048 tokens is plenty for most single-turn conversations. Only increase it if you actually need longer context.

4. Monitor VRAM Usage

Keep an eye on what's happening:

```
# Check VRAM usage in real time
nvidia-smi -l 1
```

If you see VRAM hovering near 7.5-8GB, you're on the edge. Consider a smaller model or lower quantization.

5. Kill Background GPU Processes

Video players, game launchers (Steam, Epic), and even some desktop compositors use VRAM. Check `nvidia-smi` for surprise consumers and close what you don't need.

When to Upgrade (And What To)

You've outgrown 8GB when:

- You constantly adjust context length to avoid OOM errors
- You need 13B+ models for your work
- You want to run image generation and LLMs without swapping
- You're waiting on CPU offloading and it's painfully slow

Here's the upgrade path, ranked by value:

GPU	VRAM	Street Price (Jan 2026)	What It Unlocks	Best For
Used RTX 3060 12GB	12GB	~$200	13B at Q4, comfortable 7B at Q6+	Cheapest meaningful upgrade
RTX 5060 Ti 16GB	16GB	~$429-500	30B quantized, 7B-13B at high quants	New card sweet spot (if you can find one)
Used RTX 3090	24GB	~$700-850	70B quantized, fine-tuning small models	VRAM king on a budget
RTX 4060 Ti 16GB	16GB	~$450-500 (used)	Same as 5060 Ti 16GB	Poor value vs. new 5060 Ti

The RTX 3060 12GB at ~$200 is the cheapest step up — 50% more VRAM for the price of a nice dinner. But if you're serious about local AI and can stretch the budget, the used RTX 3090 at $700-850 remains the best VRAM-per-dollar card available. Twenty-four gigabytes opens up an entirely different tier of models.

The Bottom Line

8GB of VRAM is real. Not a demo. Not a toy. You can run capable 7B-8B language models at interactive speeds, generate images with Stable Diffusion, and do genuine local AI work — all without spending another dollar.

The practical advice:

1. Install Ollama and run `ollama run qwen3.5:9b`. You'll be chatting with a vision-capable model in under five minutes.
2. Use Q4_K_M quantization and keep context to 4-8K tokens.
3. Close Chrome and other VRAM hogs before running models.
4. When 8GB isn't enough — and you'll know when — a used RTX 3090 is the move.

Your card is more capable than you think. Start using it.

Related Guides

- How Much VRAM Do You Need for Local LLMs?

- GPU Buying Guide for Local AI
- What Quantization Actually Means

Sources: DatabaseMart RTX 4060 Ollama Benchmark, LocalLLM.in 8GB VRAM Guide, XDA Developers Used RTX 3090, SDXL System Requirements, BestValueGPU Price Tracker

Get notified when we publish new guides.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/what-can-you-run-8gb-vram/

Free guides for running AI locally