What Can You Actually Run on 24GB VRAM?

January 29, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** 24GB is where local AI stops compromising. You can run 7B-14B models at maximum quality (FP16, 100-130 tok/s), 27B-32B models at Q4-Q5 (20-38 tok/s), and even squeeze in 70B at Q3 with limited context. Image generation is wide open — Flux FP16, SDXL with refiner, LoRA training all fit comfortably. The used RTX 3090 (~$700-850) is the value king; the RTX 4090 (~$1,800-2,200) is 14-17% faster at generation but 2.5x the price. Start with Qwen 3.5 27B at Q4 — it's replaced the older Qwen 2.5 32B as the sweet spot for this tier, using less VRAM while scoring higher on every benchmark.

📚 **More on this topic:** VRAM Requirements · GPU Buying Guide · Used RTX 3090 Guide · Quantization Explained

If 8GB is the floor and 12GB is the sweet spot, 24GB is where you stop counting megabytes and start choosing models based on what you actually want to do.

With 24GB, you run 32B models at interactive speeds. You run 7B-14B models at maximum quality with massive context windows. You generate Flux images at full precision without optimization hacks. And you can fine-tune your own models — something no smaller VRAM tier allows comfortably.

This is the “buy once, cry once” tier. Whether it’s a used RTX 3090 for $700 or an RTX 4090 for $2,000, 24GB of VRAM opens a different class of local AI. Here’s exactly what fits.

Who This Is For

GPU	VRAM	Memory Bandwidth	Street Price (Jan 2026)	Notes
Used RTX 3090	24GB	936 GB/s	~$700-850	The VRAM-per-dollar king
RTX 3090 Ti	24GB	1,008 GB/s	~$950	Marginal upgrade, not worth the premium
RTX 4090	24GB	1,008 GB/s	~$1,800-2,200	Faster compute, same VRAM

GPU	VRAM	Memory Bandwidth	Street Price (Jan 2026)	Notes
RTX 5090	32GB	1,792 GB/s	~$2,000 (MSRP)	More VRAM + bandwidth, if you can find one
RTX A5000	24GB	768 GB/s	~$500-1,000	Cheaper but slower bandwidth

The RTX 3090 and 4090 are the two you'll see most. Same 24GB of VRAM, same model compatibility — the difference is speed and price. We'll break that down later in this guide.

The RTX 5090's 32GB is a meaningful step up, but at $2,000+ MSRP with spotty availability, the used 3090 at $700-850 remains the pragmatic recommendation for most people.

What Runs Well on 24GB

7B-14B at Maximum Quality

On 8GB and 12GB, you run these models at Q4 or Q6 because VRAM is tight. On 24GB, you can run them at Q8 or even FP16 — near-lossless to lossless quality — with enormous context windows and VRAM to spare.

Model	Quant	VRAM	Speed (3090)	Speed (4090)	Context Headroom
Llama 3.1 8B	FP16	~14 GB	~46 tok/s	~54 tok/s	16K+ easy
Llama 3.1 8B	Q8_0	~8 GB	~67 tok/s	~80 tok/s	32K+ easy
Qwen 2.5 14B	Q8_0	~15 GB	~35 tok/s	~45 tok/s	16K comfortable
Qwen 2.5 14B	Q4_K_M	~9 GB	~60 tok/s	~83 tok/s	64K possible

Why run small models at max quality? For coding, precise reasoning, and long-context tasks, Q8 and FP16 retain nuances that Q4 quantization rounds away. If your work involves exact code generation or complex instruction following, the quality difference is measurable. And on 24GB, you're not sacrificing anything to get it.

30B-34B at Q4-Q6: The Sweet Spot

This is the tier that makes 24GB worth it. 32B models are noticeably smarter than 14B — better reasoning, better coding, longer coherent output. On 12GB, they barely fit at Q4 with minimal context. On 24GB, they run fast with room for 8-16K context windows.

Model	Quant	VRAM	Speed (3090)	Speed (4090)	Best For
Qwen 3.5 27B	Q4_K_M	~17 GB	~25-35 tok/s	~38 tok/s	Best overall — replaced Qwen 2.5 32B
Qwen 2.5 32B	Q4_K_M	~19 GB	~22 tok/s	~35 tok/s	Still solid, but superseded
DeepSeek R1 32B	Q4_K_M	~19 GB	~22 tok/s	~38 tok/s	Reasoning, math
QwQ 32B	Q4_K_M	~19 GB	~20 tok/s	~34 tok/s	Extended reasoning
CodeStral 22B	Q4_K_M	~13 GB	~35 tok/s	~50 tok/s	Coding

Qwen 3.5 27B at Q4_K_M is the new king of 24GB. At ~17GB for weights, it leaves ~7GB for KV cache — enough for 64-131K context with room to spare. It scores 72.4 on SWE-bench Verified and ranks #1 among open-weight models in the 4B-40B class on Artificial Analysis. It's 5B parameters lighter than the Qwen 2.5 32B it replaced, so it actually fits better on 24GB while being meaningfully smarter.

Qwen 2.5 32B is still a fine model if you're already using it, but for new setups, Qwen 3.5 27B is the one to pull.

RTX 3090 note: At Q4_K_M with context under 131K, the 27B fits entirely in VRAM with no spillover. Push past 131K context and VRAM starts spilling to system RAM — if that happens, reduce `num_ctx` or drop to Q3_K_M. For daily use, 64K context is the comfortable sweet spot on the 3090.

```
ollama pull qwen3.5:27b
ollama pull deepseek-r1:32b
```

70B Quantized: The Big Unlock (With Caveats)

Here's the headline: a 70B model can run on 24GB — with significant tradeoffs.

At Q3_K_M, a 70B model uses ~20-25GB for weights. That leaves barely anything for context, and quality at Q3 is noticeably degraded. At Q4_K_M, the weights alone are ~35-40GB — it doesn't fit without CPU offloading, which drops speeds to 2-3 tok/s.

Model	Quant	VRAM	Speed (3090)	Fits?	Notes
Llama 3.1 70B	IQ2_XS	~18 GB	~12 tok/s	Yes	Aggressive 2-bit, notable quality loss
Llama 3.1 70B	Q3_K_M	~22 GB	~8-10 tok/s	Barely	Minimal context (~2-4K)
Llama 3.1 70B	Q4_K_M	~38 GB	~2 tok/s	No	Needs CPU offload, very slow

The honest take: 70B at Q3 on 24GB is a proof of concept, not a daily driver. You get a smarter model than 32B, but at half the speed, worse quantization quality, and almost no context window. A 32B model at Q4-Q5 is a better experience for most tasks — faster, higher quantization quality, and room to breathe.

When 70B at Q3 makes sense: One-off complex reasoning tasks where you need the smartest model possible and speed doesn't matter. Set a short context, ask your question, get your answer.

When it doesn't: Interactive chat, coding assistance, anything requiring long context or back-and-forth conversation.

What's Possible But Tight

MoE Models

Mixture-of-Experts models like Mixtral have large total parameter counts but only activate a fraction per token, giving you big-model quality at smaller-model speeds.

Model	Total Params	Active Params	VRAM (Q4)	Fits 24GB?
Qwen 3.5 35B-A3B	35B	3.5B	~21-22 GB	Yes — fast (~110 tok/s on 3090)
Qwen3 30B-A3B	30B	3B	~18 GB	Yes — fast (~110 tok/s on 4090)
Mixtral 8x7B	47B	13B	~24-26 GB	Barely — partial offload needed
Mixtral 8x22B	141B	44B	~73-80 GB	No — needs 4x 24GB GPUs

Qwen 3.5 35B-A3B is the successor to Qwen3 30B-A3B and the MoE model to try on 24GB. Only 3.5B parameters activate per token, so it runs at small-model speeds (~110 tok/s on RTX 3090) with much larger model knowledge. At ~21-22GB Q4_K_M, it fits on 24GB with less headroom than the older 30B-A3B, but still leaves enough for 32K context.

One caveat: The 35B-A3B uses Gated DeltaNet attention layers (a linear attention variant) that currently have a 35% speed regression on CUDA in llama.cpp. This is an implementation issue being worked on — if you're using llama.cpp directly and speed matters, the older Qwen3 30B-A3B is faster right now. Ollama and other backends may or may not be affected depending on their llama.cpp version.

Mixtral 8x7B barely squeezes in — you'll need partial GPU offloading, which cuts speed. Mixtral 8x22B doesn't fit at any quantization.

Fine-Tuning with LoRA

24GB is the entry point for training your own model adaptations.

Training Task	Method	VRAM Needed	Fits 24GB?
7B LoRA	FP16 base	~20 GB	Yes — comfortable
7B QLoRA	4-bit base	~8-10 GB	Yes — plenty of room
13B QLoRA	4-bit base	~15-20 GB	Yes — tight but works
70B QLoRA	4-bit base	~35 GB	Needs CPU offload, very slow

What's realistic: QLoRA on 7B-13B models works well on 24GB. You can train a custom LoRA adapter in hours to a few days depending on dataset size. Use Unsloth for 2x faster QLoRA training, or Hugging Face PEFT for production-grade setups.

For 70B fine-tuning, the practical path is two 24GB GPUs using Answer.AI's FSDP+QLoRA method — not realistic on a single card.

What Still Won't Work

- **70B at Q4 or higher (single GPU)**: The weights alone need 35-40GB. Doesn't fit. CPU offloading drops to 2 tok/s — usable for batch work but not interactive.
- **70B at FP16**: Needs ~140GB. Not happening on any consumer GPU.
- **Full fine-tuning of anything above 7B**: A full 7B fine-tune needs ~100-120GB. LoRA/QLoRA is the only path on consumer hardware.
- **Multiple large models simultaneously**: Loading two 32B models at Q4 would need ~40GB. One model at a time.
- **Mixtral 8x22B or DeepSeek R1 671B**: These need 80-160GB+. Multi-GPU or cloud territory.

If you need 70B at Q4 or higher at full speed, you need either two 24GB GPUs, a 48GB card (used A6000 ~$1,100-2,300), or the RTX 5090 (32GB, which helps but still doesn't fully fit 70B Q4).

Image Generation on 24GB

24GB is where image generation becomes unrestricted.

Flux: Full Precision, No Compromises

Flux at FP16 uses ~18-20GB — a tight squeeze on 24GB but it works without quantization workarounds. This is the full-quality experience that 8GB and 12GB cards can't touch without NF4 quantization.

Setup	Speed (RTX 3090)	Speed (RTX 4090)	VRAM Used
Flux Dev FP16, 1024x1024	~15 seconds	~8 seconds	~18-20 GB
Flux Dev FP8, 1024x1024	~26 seconds*	~11 seconds	~12-14 GB
Flux Schnell FP16, 1024x1024 (4 steps)	~4 seconds	~2 seconds	~18-20 GB

*RTX 3090 lacks hardware FP8 — it runs via software emulation, so FP16 is actually faster on the 3090.

SDXL: Everything Works

SDXL base (~7-8GB) plus refiner (~7-8GB) both fit in VRAM simultaneously — no model swapping between passes. Add ControlNet on top and you're still under 24GB. This is the full creative pipeline that smaller cards have to run sequentially.

Training LoRAs

24GB is comfortable for Stable Diffusion LoRA training. SDXL LoRA training peaks at ~20GB VRAM — no gradient checkpointing required. You can train custom styles, characters, or concepts from as few as 10-20 images.

Best Models for 24GB GPUs (Ranked)

1. Qwen 3.5 27B — The best dense model that runs fast on 24GB. At ~17GB Q4_K_M, it leaves more headroom than the Qwen 2.5 32B it replaced while scoring higher on every benchmark. Your daily driver.

```
ollama pull qwen3.5:27b
```

2. Qwen 3.5 35B-A3B — Only 3.5B active parameters, so it runs at ~110 tok/s on a 3090 — absurdly fast for a "35B" model. Great for interactive chat where speed matters more than raw reasoning depth.

```
ollama pull qwen3.5:35b-a3b
```

3. DeepSeek R1 32B — Best for complex reasoning and math. Uses chain-of-thought "thinking" to solve hard problems. Still the go-to for tasks where you want the model to show its work.

```
ollama pull deepseek-r1:32b
```

4. Qwen 2.5 14B at Q8 — Near-lossless quality with fast speeds and huge context headroom. The best option when you need quality + speed + long context simultaneously.

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

5. CodeStral 22B — The coding specialist. At ~13GB Q4, it leaves tons of headroom for long code contexts and runs at 35-50 tok/s.

New to local AI? Start with our Ollama setup guide.

→ Check what fits your hardware with our Planning Tool.

RTX 3090 vs RTX 4090: Same VRAM, Different Experience

Both have 24GB. Both run the same models. The difference is speed, power, and price.

Spec	RTX 3090	RTX 4090
VRAM	24GB GDDR6X	24GB GDDR6X
Memory Bandwidth	936 GB/s	1,008 GB/s
CUDA Cores	10,496	16,384
Architecture	Ampere	Ada Lovelace
TDP	350W	450W
Used Price (Jan 2026)	~$700-850	~$1,800-2,200

LLM Speed Comparison

Workload	RTX 3090	RTX 4090	Difference
8B Q4 generation	~112 tok/s	~128 tok/s	4090 is **14% faster**
8B FP16 generation	~46 tok/s	~54 tok/s	4090 is **17% faster**
8B Q4 prompt processing	~3,865 tok/s	~6,899 tok/s	4090 is **78% faster**
32B Q4 generation	~22 tok/s	~35 tok/s	4090 is **59% faster**
Flux Dev FP8 (1024x1024)	~26 seconds	~11 seconds	4090 is **2.4x faster**

Token generation (the “typing” speed you see) is memory-bandwidth-bound. The 3090 and 4090 have similar bandwidth (936 vs 1,008 GB/s), so generation speed differs by only 14-17% on small models. On 32B models, the 4090’s compute advantage matters more, widening the gap to ~59%.

Prompt processing (the “thinking” phase before the first token) is compute-bound. The 4090 crushes here — 78% faster on 8B models. If you process long documents or run large batch prompts, this matters.

The Verdict

Get the RTX 3090 if: Budget matters. At $700-850, you get the same 24GB of VRAM and 85% of the generation speed at less than half the price. For most local AI work — chat, coding, image generation — the 3090 is the rational choice.

Get the RTX 4090 if: You do heavy prompt processing (RAG, long documents), need faster image generation (especially Flux), or also game at 4K. The 4090 earns its premium in compute-bound workloads and offers DLSS 3 for gaming.

Skip the RTX 3090 Ti. It costs ~$100-150 more than the standard 3090 for almost identical AI performance. The bandwidth improvement is marginal.

Tips to Maximize 24GB

1. Enable Flash Attention

Flash attention reduces VRAM usage for the KV cache and speeds up inference. Most modern inference backends support it:

```
# llama.cpp
./llama-cli -m model.gguf --flash-attn -fa 1

# Ollama enables it automatically on supported models
```

2. Quantize Your KV Cache

For longer contexts on 32B models, quantize the KV cache from FP16 to Q8 or Q4:

```
# llama.cpp
./llama-cli -m qwen2.5-32b-q4_k_m.gguf \
  -c 16384 \
  --cache-type-k q8_0 \
  --cache-type-v q4_0
```

This roughly doubles your usable context length with minimal quality impact. A 32B model at Q4 with quantized KV cache can handle 16K+ context on 24GB.

3. Use Q4_K_M as Your Default for 32B Models

On 24GB with 32B models, Q4_K_M is the sweet spot — it leaves enough headroom for useful context windows. Q5_K_M fits but limits you to ~4-8K context. Q6 and above won't fit.

For 7B-14B models, you can afford Q8 or FP16 — use the highest quality that leaves room for your context needs.

4. Monitor VRAM Usage

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

On 24GB, you have more room than smaller cards, but 32B models still push limits. Watch for usage above 22GB — that's when you're close to the edge and long conversations might trigger OOM.

5. Close Background GPU Apps

Same advice as smaller VRAM tiers but less critical. Chrome's hardware acceleration, game launchers, and Discord with screen sharing all consume VRAM. On 24GB this rarely matters for 7B-14B models, but it can push 32B Q4 sessions over the edge.

The Bottom Line

24GB is the tier where you pick models based on what you need, not what fits. You run 32B models at interactive speeds, generate Flux images at full precision, and fine-tune your own LoRAs — all on a single GPU.

The practical advice:

1. Install Ollama and pull `qwen3.5:27b`. That's your main model — smarter than the Qwen 2.5 32B it replaced, and it fits better at ~17GB Q4_K_M.
2. Pull `qwen3.5:35b-a3b` for fast interactive chat — 110 tok/s with only 3.5B active parameters.
3. Use Q4_K_M for 27B-32B models, Q8 for 14B models, FP16 for 7B models — you have the VRAM for quality.
4. For image generation, use Flux at FP16 (RTX 4090) or FP16/FP8 (RTX 3090). SDXL with refiner runs without workarounds.
5. If you're choosing between a 3090 and 4090, the 3090 at $700-850 is the value play. The 4090 is faster but costs 2.5x more.

You're in the big leagues now. Use the hardware.

Related Guides

- What Can You Run on 8GB VRAM?
- What Can You Run on 12GB VRAM?
- GPU Buying Guide for Local AI
- Used RTX 3090 Buying Guide

Sources: Hardware Corner GPU Ranking for LLMs, Hardware Corner RTX 4090 LLM Benchmarks, Hardware Corner RTX 3090 24GB Guide, Hardware Corner RTX 5090 LLM Benchmarks, RunPod RTX 5090 Benchmarks, Puget Systems Consumer GPU LLM Analysis, DatabaseMart RTX 4090 Ollama Benchmark, BestValueGPU Price Tracker, XDA Used RTX 3090 Value King

Get notified when we publish new guides.

Subscribe — free, no spam