What Can You Actually Run on 24GB VRAM?

January 29, 2026 · Updated: November 4, 2026

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 ($1,400-1,550) is the value king; the RTX 4090 ($2,900-3,100) is 14-17% faster at generation but about twice the price. Start with Qwen 3.6-27B or 3.8-27B at Q4 (about 41 tok/s on our 3090; the two tie on our tests), switch to Qwen 3.6-35B-A3B when speed matters (157.66 tok/s), and use Nemotron 3.5 Bolt for long context: it loads at 128K on one 3090 with 2,310 MiB spare.

📚 **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 around $1,500 or an RTX 4090 for around $3,000, 24GB of VRAM opens a different class of local AI. Here’s exactly what fits.

Update, 2026-10-02: this paragraph originally said around $2,250 for a 4090. The page summary originally gave $2,100-2,550, nearly twice a 3090 at $1,000-1,400. The card is $2,575-3,010 used as of October 2026 (median $2,860, from 216 sold eBay listings). Against a $1,200-1,400 3090 it is more than twice the price.

Updated 2026-11-04: picks refreshed from our own RTX 3090 runs. The 24GB pick is now Qwen 3.6-27B or 3.8-27B (a tie on our tests), with Qwen 3.6-35B-A3B for speed and Nemotron 3.5 Bolt for 128K context. Originally recommended Qwen 3.5 27B as the main model and Qwen 3.5 35B-A3B at ~110 tok/s, with a caveat about a llama.cpp CUDA slowdown that was fixed in February. Also, the flash-attention command is `-fa on`: current llama.cpp’s `-fa` takes a value (`on`, `off` or `auto`), and the page originally said `--flash-attn -fa 1`, which fails.

Updated 2026-10-15: prices refreshed from eBay US sold listings (used, excluding shipping). Originally quoted ~$1,200 and $1,000-1,400 for a used RTX 3090 (July 2026), ~$2,860 and $2,575-3,010 for an RTX 4090 (October 2026), and ~$2,000 (MSRP) for the RTX 5090 row (January 2026).

Who This Is For

GPU	VRAM	Memory Bandwidth	Street Price	Notes
Used RTX 3090	24GB	936 GB/s	~$1,500	The VRAM-per-dollar king
RTX 3090 Ti	24GB	1,008 GB/s	~$950 (unverified)	Marginal bandwidth gain, 450W draw
RTX 4090	24GB	1,008 GB/s	~$3,000	Faster compute, same VRAM
RTX 5090	32GB	1,792 GB/s	~$5,900-6,350 used	More VRAM + bandwidth, if you can find one
RTX A5000	24GB	768 GB/s	~$500-1,000	Cheaper but slower bandwidth
Update, 2026-09-24: the RTX 3090 Ti cell originally said ~$950. Our September eBay sold-listing pull puts it at $1,300-1,499 used, above the vanilla 3090's $1,200-1,400. The RTX 5090 sells for $3,900-4,400 used, not MSRP.				

Update, 2026-10-02: the RTX 4090 row originally said ~$2,250. The card is $2,575-3,010 used as of October 2026 (median $2,860, from 216 sold eBay listings).

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 $1,400-1,550 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.

27B Dense at Q4: The Sweet Spot

This is the tier that makes 24GB worth it. The 27B-class models are noticeably smarter than 14B — better reasoning, better coding, longer coherent output. On 12GB they don't fit without offloading. On 24GB they run entirely on the card with room for context.

Measured on our own RTX 3090 (llama.cpp, `llama-bench`, headless; peak VRAM from the card):

Model	Quant	Peak VRAM	Generation (3090)	At 8K depth	Source
Qwen 3.6-27B	UD-Q4_K_XL	18.5 GB	41.77 tok/s	40.67 tok/s	3.8 vs 3.6 test
Qwen 3.8-27B	UD-Q4_K_XL	18.8 GB	41.49 tok/s	40.48 tok/s	3.8 vs 3.6 test

We don't have our own 4090 numbers for these models, so there's no 4090 column. For coding-specific picks, see the coding models roundup.

Qwen 3.6-27B or Qwen 3.8-27B at Q4 is the 24GB pick, and it doesn't much matter which. On our 3090 test the two generate within a percent of each other, 3.8 needs 254 MiB more VRAM, and on HumanEval they're a statistical tie (3.6 at 82.32%, 3.8 at 80.49%, McNemar p = 0.70). Alibaba reports bigger gains for 3.8 on agentic coding benchmarks we can't run on a consumer card. If you already run 3.6-27B, there's no measured reason to re-download; if you're starting fresh, 3.8-27B is the newer weights under the same Apache 2.0 license.

RTX 3090 note: our peak VRAM figures above were read with up to 8K of context filled. Longer contexts grow the KV cache, so watch `nvidia-smi` and lower `num_ctx` (Ollama) or `-c` (llama.cpp) if usage climbs past ~22GB.

```
ollama pull qwen3.6:27b
# or the newer weights
ollama pull qwen3.8:27b
```

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 have large total parameter counts but only activate a fraction per token, giving you big-model knowledge at smaller-model speeds. Two of them are worth knowing on 24GB, and we've measured both on our 3090.

Model	Total Params	Active Params	VRAM (Q4)	Fits 24GB?
Qwen 3.6-35B-A3B	35B	3B	22.7 GB peak (UD-Q4_K_M)	Yes — 157.66 tok/s on our 3090
Nemotron 3.5 Bolt 74B-A9B 24GB	74B	9B	22.02 GB file (NVIDIA's 24GB build)	Yes — at 128K, about 90 tok/s on our 3090
Mixtral 8x7B	47B	13B	~24-26 GB	Barely — partial offload needed
Mixtral 8x22B	141B	44B	~73-80 GB	No — needs 4x 24GB GPUs

Qwen 3.6-35B-A3B is the speed pick. Only 3B parameters activate per token, so it ran at 157.66 tok/s on our 3090 at UD-Q4_K_M, 153.24 tok/s with 8K of context filled, with the whole model on the card (22.7 GB peak).

Nemotron 3.5 Bolt is the long-context pick. NVIDIA built this 74B-total, 9B-active model to fit 24GB at 128K, and on our 3090 it does: 128K context loads with 2,310 MiB to spare at f16 KV cache, decode runs about 90 tok/s, and it still does 82.42 tok/s with 102,400 tokens of context filled. Keep the KV cache at f16 on this model; q8_0 dropped deep-context decode to 60.61 tok/s. We've checked its quality on one task only (a tie on a 47-item intent split), so treat it as a long-context tool, not a general upgrade.

```
ollama pull qwen3.6:35b-a3b

# Nemotron 3.5 Bolt, as we ran it (llama.cpp v0.7.0; the server listens on 9931 unless you pass
```

```
llama-server -m NVIDIA-Nemotron-3.5-Bolt-74B-A9B-24GB.gguf \
  -ngl 99 -fa on -c 131072 -np 1 -ctk f16 -ctv f16 --fit off -lm mmap
```

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 3.6-27B or 3.8-27B — The daily driver. About 41 tok/s on our 3090 at Q4 in 18.5-18.8 GB, with thinking mode for reasoning and math. The two tied on our speed and HumanEval tests; pick 3.8 if you're starting fresh.

```
ollama pull qwen3.8:27b
```

2. Qwen 3.6-35B-A3B — The speed pick. 157.66 tok/s on our 3090 with only 3B parameters active. Great for interactive chat and agents where response time matters more than the last bit of reasoning depth.

```
ollama pull qwen3.6:35b-a3b
```

3. Nemotron 3.5 Bolt — The long-context pick. A 74B model that loads at 128K on one 3090 and still decodes 82.42 tok/s with 102,400 tokens filled. Run it with llama.cpp at f16 KV cache (command above).

For coding, the 27B pick above is where to start; the coding models roundup covers the alternatives.

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 $1,400-1,550 — expect around $1,500 after the 2026 memory-shortage surge — you get the same 24GB of VRAM and 85% of the generation speed at about 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 delivers almost identical AI performance for 100W more draw, and the bandwidth improvement is marginal. Our used-price figure for the Ti predates the 2026 surge, so check both cards live before assuming either is cheaper.

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 -fa on

# 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 27B 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 `qwen3.8:27b` (or keep the `qwen3.6:27b` you have; they tied on our tests). That's your main model, about 41 tok/s on a 3090.
2. Pull `qwen3.6:35b-a3b` for fast interactive chat — 157.66 tok/s on our 3090 with only 3B active parameters. For 128K context, run Nemotron 3.5 Bolt in llama.cpp.
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 $1,400-1,550 is the value play. The 4090 is faster but costs about twice as much.

Update, 2026-10-02: this line originally said a 3090 at $1,000-1,400 and a 4090 at nearly double. The card is $2,575-3,010 used as of October 2026 (median $2,860, from 216 sold eBay listings).

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

A weekly email with every new guide and measured benchmark.

Subscribe — free, no spam

Source: https://insiderllm.com/guides/what-can-you-run-24gb-vram/

Free guides for running AI locally