GPU Buying Guide for Local AI: Pick the Right Card

January 25, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For most people getting into local AI, a used RTX 3060 12GB ($180-250) is the best entry point. If you want to run larger models, a used RTX 3090 ($700-900) offers the best VRAM-per-dollar on the market.

📚 **More on this topic:** VRAM Requirements · Used RTX 3090 Guide · Used GPU Buying Guide · AMD vs NVIDIA

When I first started exploring AI, I experimented with image generation and quickly ran up against real barriers—my graphics card wasn't up to the task and my motherboard couldn't hold enough memory. My image generation took extremely long to render. I quickly found out my GPU's VRAM was too small and was a major bottleneck.

After trying several configurations, I was able to make some things work, but with limitations. Wanting to upgrade, I was intimidated by how expensive the cards were. Later, after upgrading my motherboard, CPU, and memory, I bought a 3060. This opened up the ability to run several of the smaller LLM models and generate bigger images.

You can do a lot with a 12GB GPU card, but if you want to use even larger models, it's all about VRAM. That's why I created this guide. We'll cover recommended cards by budget, where to buy used cards, how to buy safely, and auction guidelines.

Why VRAM Matters More Than Anything Else

Here's the brutal truth about running AI locally: if your model doesn't fit in VRAM, it doesn't run. There's no graceful degradation. You either have enough memory or you get an out-of-memory error.

This is why a used RTX 3090 with 24GB of VRAM at $800 is a better buy than a new RTX 4070 Ti with 12GB at the same price. For LLM work, VRAM capacity trumps architectural improvements in almost every case.

The VRAM Cliff

The rule of thumb: roughly 2GB of VRAM per billion parameters at FP16 precision. A 7B model needs ~14GB. A 13B model needs ~26GB. A 70B model needs ~140GB.

But here's where quantization saves us. A 13B model that needs 26GB at full precision can run in 8-10GB when quantized to 4-bit. You lose some quality, but the model actually runs.

VRAM Requirements Table

VRAM	LLMs You Can Run	Image Generation	Real-World Examples
8GB	7B models (Q4 only)	SD 1.5, limited SDXL	Llama 3.1 8B Q4, Mistral 7B Q4
12GB	7B full precision, 13B Q4-Q8	SDXL, Flux (tight)	Llama 3.1 8B, Qwen 2.5 14B Q4
16GB	13B full, 20B Q4-Q6	SDXL + ControlNet, Flux comfortably	Llama 3.1 8B + long context, Mixtral 8x7B Q4
24GB	30B full, 70B Q4	Everything + batching	Llama 3.3 70B Q4, Qwen 2.5 32B, any SD model
48GB+	70B Q6-Q8, 100B+ Q4	Multiple models loaded	Llama 3.1 70B near-full precision

→ Use our Planning Tool to check exact VRAM for your setup.

Recommended GPUs by Budget Tier

Entry Level ($150-$300): Getting Started

Best Pick: NVIDIA RTX 3060 12GB — $180-250 used

This is the card I started with, and it's still the best tight-budget choice for local AI. That 12GB of VRAM is its secret weapon—many newer cards (RTX 4060, RTX 4070) ship with only 8GB, which severely limits what models you can run.

Spec	RTX 3060 12GB
VRAM	12GB GDDR6
Memory Bandwidth	360 GB/s
Used Price	$180-250

Spec	RTX 3060 12GB
Power Draw	170W
Token Speed (8B)	~38-40 t/s

What you can run (from my experience):

- **LLMs:** Llama 3.1 8B runs great at ~38 t/s, Qwen 2.5 7B hits 40 t/s. 13B models work with Q4-Q6 quantization but you'll feel the limits.
- **Image Gen:** Stable Diffusion XL runs well, Flux works with careful VRAM management. SD 1.5 is no problem.
- **Limitations:** 70B models are completely out of reach. Context length gets limited fast on anything above 7B.

The 3060 12GB opened local AI for me. It's where I recommend everyone start.

Alternative: Intel Arc B580 — $249 new

The Arc B580 punches above its weight at 62 tokens/second for 8B models—faster than the 3060. The catch? It requires Intel's IPEX-LLM or OpenVINO stack instead of CUDA. If you're comfortable with Linux and some tinkering, it's excellent value. If you want plug-and-play, stick with NVIDIA.

Ultra-Budget Builds: The Radeon VII Route

If you're handy and want to go even cheaper, check out Country Boy Computers. He does wild budget AI builds—like pairing a Radeon VII (16GB HBM2 with massive bandwidth) with old Dell workstations for under $400 total. It requires ROCm and Linux, so it's not for beginners, but if you want maximum VRAM per dollar and don't mind tinkering, it's worth watching his builds for inspiration. For a step-by-step approach, see our budget AI PC build guide.

Skip: RTX 4060 (8GB), RTX 3060 Ti (8GB), any card under 12GB VRAM

Mid-Range ($300-$700): The Sweet Spot

Best Pick: NVIDIA RTX 3090 — $700-900 used

After hitting the VRAM wall on my 3060, I upgraded to a 3090. The difference was night and day. The 24GB of VRAM and 936 GB/s memory bandwidth make this card competitive with GPUs twice its price. If you're considering the same move, read our used RTX 3090 buying guide for detailed tips.

Spec	RTX 3090	RTX 4060 Ti 16GB	Intel Arc A770
VRAM	24GB GDDR6X	16GB GDDR6	16GB GDDR6
Memory Bandwidth	936 GB/s	288 GB/s	560 GB/s
Price	$700-900 used	$499 new	$250-350
Token Speed (8B)	~87-111 t/s	~34 t/s	~45-55 t/s

The RTX 4060 Ti 16GB looks good on paper, but its narrow 128-bit memory bus cripples LLM performance. The 3090's bandwidth advantage translates to 2-3x faster inference despite being two generations older.

What you can run (from my experience):

- **LLMs:** 30B models run great. 70B models with Q4 quantization are usable (15-25 t/s)—not fast, but functional. Any 7B-13B model flies with room for massive context windows.
- **Image Gen:** Everything works. SDXL, Flux, ControlNet stacks, multiple LoRAs—no compromises.
- **The jump from 12GB to 24GB:** This is where local AI gets genuinely useful. Models that were impossible suddenly just work.

Skip: RTX 4070 (12GB), RTX 4070 Ti (12GB)—the VRAM is too limiting for the price

---

High-End ($700-$1,500): Serious Local AI

Best Pick: NVIDIA RTX 4090 — $1,400-1,800 used / $1,599 new

If budget allows, the RTX 4090 is the best single consumer GPU for local AI. It's roughly 50% faster than the 3090 with the same 24GB VRAM.

Spec	RTX 4090	RTX 3090
VRAM	24GB GDDR6X	24GB GDDR6X
Memory Bandwidth	1,008 GB/s	936 GB/s
Token Speed (8B)	~128 t/s	~87-111 t/s
Power Draw	450W	350W
Price	$1,400-1,800	$700-900

Is it worth 2x the price of a 3090? For most people, no. The 3090 hits the sweet spot. If you're running inference constantly or time is money, the 4090 makes sense.

Buying Used GPUs: Where and How

Best Places to Buy Used Cards

eBay — Largest selection, best buyer protection

- Pros: Huge inventory, Money Back Guarantee, easy returns
- Cons: Prices slightly higher, some scam attempts
- Best for: RTX 3090, 3060, 4090

r/hardwareswap — Best prices, more risk

- Pros: 10-20% cheaper than eBay, direct negotiation
- Cons: Less protection, requires PayPal Goods & Services
- Best for: Experienced buyers who know fair prices

Facebook Marketplace — Local deals, cash transactions

- Pros: Can inspect before buying, no shipping damage risk
- Cons: Limited selection, potential safety concerns
- Best for: Local pickup in metro areas

How to Buy Safely

Before you buy:

1. Check completed auctions on eBay to see actual sold prices—not asking prices
2. Verify the listing category is Consumer Electronics or Computers/Tablets (Business & Industrial has no Money Back Guarantee)
3. Look for real photos of the actual card, not stock images
4. Read the full description for signs of mining use or defects
5. Check seller feedback—look for GPU-specific reviews if possible

Red flags to avoid:

- Stock photos instead of real photos

- Price significantly below market (too good to be true)
- Seller has multiple identical cards listed (likely miner)
- Vague descriptions ("works great!" with no details)
- New account with no feedback

After you buy:

1. Inspect the card for physical damage, dust buildup, or thermal paste residue
2. Check the serial number matches the listing photos
3. Install and verify it's detected correctly in device manager
4. Stress test for 24 hours using Unigine Valley or FurMark
5. Run actual AI workloads to verify stability under sustained load

eBay Auction Guidelines

Sniping strategy:

- Don't bid early—it just drives up the price
- Set your maximum bid and use a sniping tool (or manual snipe) in the last 10 seconds
- Your max bid should be what you'd actually pay, not a lowball

Current fair prices (Jan 2025):

Card	Fair Auction Price	Fair BIN Price
RTX 3060 12GB	$150-180	$200-250
RTX 3090	$650-750	$800-900
RTX 3090 Ti	$750-850	$900-1,000
RTX 4090	$1,300-1,500	$1,500-1,800

Before You Buy: System Requirements

Power Supply Considerations

GPU	Minimum PSU	Recommended PSU
RTX 3060 12GB	500W	550W

GPU	Minimum PSU	Recommended PSU
RTX 3090	750W	850W
RTX 4090	850W	1000W
Dual 3090	1000W	1200W

Don't cheap out here. A quality 80+ Gold PSU from Corsair, EVGA, or Seasonic is worth the investment.

Motherboard Compatibility

For single GPU: Any motherboard with a PCIe x16 slot works.

For dual GPU:

- Need two physical x16 slots (they'll usually run at x8/x8)
- Verify your board supports this in the manual
- Consumer boards often have issues; workstation boards (like Threadripper) are more reliable

---

My Recommendations by Budget

Budget	Recommended Card	What You Can Run
$200	RTX 3060 12GB (used)	7B models great, 13B with quantization
$500	RTX 3060 12GB + save for 3090	Same as above—save the difference
$800	RTX 3090 (used)	30B models, 70B with Q4, all image gen
$1,500	RTX 4090 or dual 3090	70B usable, fast everything
$2,000+	RTX 5090 or dual 4090	70B comfortable, future-proofed

My honest take: Start with a used RTX 3060 12GB to learn—our beginner's guide to running your first local LLM walks you through it—then upgrade to a used RTX 3090 when you hit the VRAM wall. This is the path I took, and it costs under $1,100 total to get you to 24GB of VRAM where you can run almost anything.

Don't buy a new mid-range card. The RTX 4060, 4070, and their Ti variants are bad values for AI work—limited VRAM at high prices. The used market is your friend.

Conclusion

The local AI GPU market has a clear hierarchy:

1. **Entry point:** RTX 3060 12GB ($180-250 used)
2. **Sweet spot:** RTX 3090 24GB ($700-900 used)
3. **Premium:** RTX 4090 24GB ($1,400-1,800)
4. **Overkill:** Multi-GPU or workstation cards

Buy used, stress test thoroughly, and don't overpay for VRAM you can get cheaper on older cards. The RTX 3090 exists, and it's the best deal in local AI hardware.

Related Guides

- How Much VRAM Do You Actually Need for Local LLMs?
- Used RTX 3090 Buying Guide for Local AI
- Build a Local AI PC for Under $500

Get notified when we publish new guides.

Subscribe — free, no spam
