


# 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 (\$180-220) is the best entry point. If you want to run larger models, a used RTX 3090 (\$800-900) offers the best VRAM-per-dollar on the market. In 2026, the RTX 5090 (32GB, \$2,000 MSRP) is the speed king but nearly impossible to find at retail. Intel's Arc Pro B70 (32GB, \$949) is a wildcard with software risks. The RTX 5060 Ti 16GB (\$429 MSRP) is decent but overpriced at current \$570 street.

 **More on this topic:** [VRAM Requirements](#) · [Used RTX 3090 Guide](#) · [Used GPU Buying Guide](#) · [AMD vs NVIDIA](#) · [RTX 5090 Benchmarks](#) · [Intel 32GB GPU](#)

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, Mistral 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
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 12GB — \$249-299 new

The Arc B580 has been out long enough now for real community benchmarks. On llama.cpp's Vulkan backend, it hits 62 tok/s on Qwen 2.5 7B Q4 — faster than the RTX 3060's ~45 tok/s. Same 12GB VRAM, 456 GB/s bandwidth (vs 360 on the 3060).

The catch hasn't changed: no CUDA. Ollama still lacks native Arc support as of March 2026 (models fall back to CPU). You need the Vulkan backend in llama.cpp, which works but isn't plug-and-play. Intel archived the IPEX-LLM repo in January 2026, so the software story is Vulkan or bust. If you're on Linux and comfortable with llama.cpp directly, it's genuinely the best value at \$4.50 per tok/s. If you want Ollama to just work, 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](#) — \$800-900 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.

Used 3090 prices have crept up slightly in early 2026 — expect \$800-900 on eBay now, up from \$700-800 a year ago. The RTX 5090's extreme street markup (\$3,800 vs \$2,000 MSRP) has kept demand high for 3090s. It's still the best VRAM-per-dollar play.

Spec	RTX 3090	RTX 5060 Ti 16GB	RTX 4060 Ti 16GB
VRAM	24GB GDDR6X	16GB GDDR7	16GB GDDR6
Memory Bandwidth	936 GB/s	448 GB/s	288 GB/s
Price	\$800-900 used	\$429 MSRP (~\$570 street)	\$499 new
Token Speed (8B)	~87 t/s	~51-60 t/s	~34 t/s

### New contender: [RTX 5060 Ti 16GB](#) — \$429 MSRP (~\$570 street)

The 5060 Ti 16GB shipped in early 2026 with GDDR7 pushing 448 GB/s bandwidth — a 55% jump over the 4060 Ti's 288 GB/s. That translates to ~51-60 tok/s on 8B models and ~32 tok/s on 14B Q4. At MSRP, it's a decent mid-range option. At the current \$570 street price, it's overpriced — you're paying 4060 Ti money for the same 16GB VRAM with a speed bump. Wait for supply to normalize, or put that \$570 toward a used 3090 with 8GB more VRAM.

### What you can run on the 3090 (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. RTX 5060 Ti at \$570 street — wait for MSRP or buy a used 3090 instead.

## High-End (\$700-\$2,000): Serious Local AI

**Best Pick: [NVIDIA RTX 4090](#) — \$1,400-1,800 used / \$1,599 new**

The RTX 4090 remains the best value high-end GPU because the 5090 is nearly impossible to buy at MSRP.

Spec	RTX 5090	RTX 4090	Intel Arc Pro B70	RTX 3090
VRAM	32GB GDDR7	24GB GDDR6X	32GB GDDR6	24GB GDDR6X
Bandwidth	1,792 GB/s	1,008 GB/s	608 GB/s	936 GB/s
Token Speed (8B)	~186 t/s	~128 t/s	~40-50 t/s (est.)	~87 t/s
Power Draw	575W	450W	230W	350W
Price	\$2,000 MSRP (~\$3,800 street)	\$1,400-1,800	\$949 new	\$800-900 used

### RTX 5090 — \$1,999 MSRP (~\$3,800 street)

The fastest single GPU for local AI by a wide margin. 32GB GDDR7 at 1,792 GB/s delivers 186 tok/s on 8B models, 124 on 14B, and 234 on MoE architectures like Qwen3 30B-A3B. The extra 8GB over the 4090 means Qwen 3.5 27B at Q4 fits with room for 32K+ context. See our [full RTX 5090 benchmark breakdown](#) for detailed numbers.

The problem: street prices are nearly double MSRP. At \$3,800, you could buy four used RTX 3090s (96GB total VRAM). At \$2,000 MSRP, it's the clear 2026 pick. At scalper prices, it's not.

### Intel Arc Pro B70 — \$949 new

Intel's wildcard entry, launched March 25, 2026. 32GB GDDR6 at 608 GB/s. That's the same VRAM as the RTX 5090 at half the price. The bandwidth is the bottleneck — 608 GB/s puts single-stream generation below the RTX 3090's level. Early llama.cpp testing on SYCL shows it trailing a 3090 on 27B Q4 models.

The real risk is software. Intel archived the IPEX-LLM repo, SYCL still has flash attention bugs on Battlemage, and Ollama doesn't support it. This is a card for people who want 32GB of VRAM for under \$1,000 and accept the software tradeoffs. Multi-GPU vLLM setups are where it might shine — Level1Techs tested 4x B70s hitting 369 tok/s output throughput on Qwen 3.5 27B. Read our [full Intel 32GB GPU analysis](#).

Is the 4090 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. And if the 5090 ever becomes available at MSRP, it replaces the 4090 as the obvious high-end pick.

## 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 (March 2026):

Card	Fair Auction Price	Fair BIN Price
RTX 3060 12GB	\$150-180	\$180-220
RTX 3090	\$750-850	\$800-1,050
RTX 3090 Ti	\$850-950	\$950-1,100
RTX 4090	\$1,300-1,500	\$1,500-1,800
RTX 5090	\$3,000-3,500	\$3,500-3,800

## Before You Buy: System Requirements

### Power Supply Considerations

GPU	Minimum PSU	Recommended PSU
RTX 3060 12GB	500W	550W
RTX 3090	750W	850W
RTX 4090	850W	1000W

GPU	Minimum PSU	Recommended PSU
RTX 5090	1000W	1200W
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-8B models great, 14B with quantization
\$250-300	Intel Arc B580 12GB (new)	7B-8B at 62 tok/s (Vulkan only, no Ollama)
\$500	RTX 3060 12GB + save for 3090	Same as above — save the difference
\$900	RTX 3090 (used)	30B models, 70B with Q4, all image gen
\$950	Intel Arc Pro B70 (new)	32GB VRAM for 27B-34B models (immature software)
\$1,500	RTX 4090 or dual 3090	70B usable, fast everything
\$2,000	RTX 5090 (if available at MSRP)	32GB, fastest single GPU, 70B Q4, 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 in 2026:

1. **Entry point:** RTX 3060 12GB (\$180-220 used) or Intel Arc B580 (\$250 if you're OK with Vulkan)
2. **Sweet spot:** RTX 3090 24GB (\$800-900 used) – still the VRAM-per-dollar king
3. **Mid-range new:** RTX 5060 Ti 16GB (\$429 MSRP, wait for street to normalize)
4. **32GB budget:** Intel Arc Pro B70 (\$949) – if you accept the software risk
5. **Premium:** RTX 4090 24GB (\$1,400-1,800) or RTX 5090 32GB (\$2,000 MSRP if you can find one)
6. **Multi-GPU:** Dual RTX 3090 (~\$1,800) for 48GB total

Intel is finally a real option in 2026, but CUDA still wins on software maturity. The used NVIDIA market is your friend – buy used, stress test thoroughly, and don't overpay for VRAM you can get cheaper on older cards.

---

## Related Guides

---

- [RTX 5090 vs DGX Spark vs AMD: Full Benchmark Shootout](#)
- [Intel's 32GB GPU for Local AI](#)
- [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](#)
- [ROCm vs CUDA for Local AI in 2026](#)

Get notified when we publish new guides.

[Subscribe](#) – free, no spam

---

Source: <https://insiderllm.com/guides/gpu-buying-guide-local-ai/>

Free guides for running AI locally