NVIDIA GPU Prices Are Rising: What to Do Now

January 27, 2025 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** GPU prices are spiking due to GDDR7 shortages and NVIDIA prioritizing datacenter AI over consumer cards. RTX 5090s are selling at $3,500+ (nearly double MSRP), mid-range 16GB cards face 30-40% production cuts, and the situation will likely worsen through mid-2026. If you need 24GB VRAM, buy a used RTX 3090 now ($650-800) before prices climb further.

📚 **More on this topic:** Used RTX 3090 Guide · AMD vs NVIDIA · Budget AI PC Build · Planning Tool

The GPU market is entering a new crisis—and this time it's not crypto miners or pandemic supply chains. It's AI.

NVIDIA's datacenter business now generates 12x more revenue than gaming. Memory manufacturers are prioritizing HBM for AI training over GDDR7 for consumer cards. And the company that controls 92% of the discrete GPU market has made its priorities clear: data centers first, gamers and hobbyists second.

For local AI enthusiasts, this creates a difficult situation. The cards we need most—high-VRAM models for running 32B and 70B parameter LLMs—are exactly the cards getting squeezed hardest. Here's what's happening, why, and what you can do about it.

What's Happening Right Now

The RTX 5000 Series Launch Problem

NVIDIA's RTX 5000 series launched with strong specs but immediately ran into availability issues. The RTX 5090 has a $1,999 MSRP, but good luck finding one at that price. Current retail listings show:

- **RTX 5090:** $3,500-4,000 at most retailers (nearly 2x MSRP)
- **RTX 5080:** $1,200-1,600 (20-60% over $999 MSRP)
- **RTX 5070 Ti / 5060 Ti:** Limited stock, frequently out of stock

At Newegg, you can't find an RTX 5090 below $3,500, with many models approaching $5,000. Most cards listed are shipping from China, suggesting domestic stock is exhausted. Only Microcenter appears to occasionally have cards at MSRP—if you can get there in person when stock arrives.

GDDR7 Memory Shortage

The root cause is a severe shortage of GDDR7 memory, particularly the 3GB chips needed for higher-VRAM configurations.

The 3GB GDDR7 chips are experiencing a "crazy shortage" that has effectively killed NVIDIA's plans for RTX 5000 Super variants. The Super series was expected to bump VRAM significantly—the RTX 5070 from 12GB to 18GB, the RTX 5080 from 16GB to 24GB—using these larger memory chips. That's now unlikely to happen.

The shortage isn't temporary. Memory manufacturers like Samsung and SK Hynix are asking themselves why they should dedicate fab capacity to GDDR7 when they can retool existing lines for HBM (High Bandwidth Memory) at far higher profit margins. AI datacenters pay premium prices for HBM. Consumer GPU buyers don't.

NVIDIA's Prioritization Strategy

NVIDIA has been explicit about where its focus lies. CFO Colette Kress stated during a recent earnings call: "We have evolved over the past twenty-five years from a gaming GPU company to now an AI data center infrastructure company."

The numbers back this up:

- **Data center revenue (Q3 2025):** $51.2 billion (+66% YoY)
- **Gaming revenue (Q3 2025):** $4.3 billion (+30% YoY)

In absolute terms, NVIDIA's data center business added $20.4 billion in new revenue last year. Gaming added $1 billion. When memory supply is constrained, it's obvious which segment gets priority.

NVIDIA has confirmed this directly: "Demand for GeForce RTX GPUs is strong, and memory supply is constrained. We continue to ship all GeForce SKUs and are working closely with our suppliers to maximize memory availability." Translation: we're doing what we can, but don't expect miracles.

Why Prices Are Going Up

Datacenter AI Demand Is Insatiable

The AI infrastructure buildout is consuming resources at an unprecedented scale. NVIDIA management expects to benefit from $3-4 trillion in AI infrastructure spending over the next five years. Companies like OpenAI, Microsoft, Google, and Meta are racing to build massive GPU clusters, and they'll pay whatever it takes.

This creates direct competition for the same memory chips that go into consumer GPUs. When a datacenter customer will pay premium prices for every available HBM chip, consumer GDDR7 production gets deprioritized.

Memory Supply Can't Keep Up

Manufacturing costs have surged by approximately 80% due to memory prices alone. GDDR6 memory costs rose roughly 60% between mid and late 2025, with GDDR7 increasing even more sharply.

These increases aren't driven by consumer demand—they're driven by structural reallocation of memory manufacturing toward AI workloads. The same fabs that could produce GDDR7 can be adapted to produce HBM at higher margins.

Tariffs and Trade Uncertainty

US-China trade tensions have added another layer of cost. Tariffs exceeding 100% on Chinese imports have effectively created a trade embargo on many components. NVIDIA has released China-specific variants (RTX 4090D, RTX 5090D) due to export restrictions, which has driven up global demand for unrestricted cards.

NVIDIA has moved some Blackwell chip production to TSMC's Arizona facility to bypass the worst tariff impacts, but "materials, logistics, and other costs" have still increased, pushing prices higher even for US-manufactured cards.

The Cascade Effect on Older Cards

When new cards are scarce and overpriced, demand shifts to previous-generation hardware. This is why RTX 4090 prices have climbed from the $1,600 MSRP to $2,000-3,000+ for new units. Stock is dwindling because NVIDIA halted RTX 4090 manufacturing to shift capacity to RTX 50-series chips.

The same pressure is building on used markets. Cards that would normally depreciate are holding value or even appreciating because they're the only affordable path to high-VRAM configurations.

Which Cards Are Hit Hardest

Mid-Range 16GB Cards (RTX 5060 Ti, 5070 Ti)

The 16GB tier is getting squeezed from both sides. These cards use GDDR7 memory that's in short supply, but they don't generate enough margin to justify priority allocation.

NVIDIA is reportedly planning 30-40% production cuts for H1 2026, with the RTX 5060 Ti and RTX 5070 Ti first in line for reductions. If you're planning to buy a 16GB card for local AI work, availability will likely get worse before it gets better.

Current situation:

- **RTX 5070 Ti (16GB):** Sporadic availability, ~$750-900
- **RTX 5060 Ti (16GB):** Limited stock, pricing volatile
- **RTX 4060 Ti 16GB:** Stable at ~$430-450, but a previous-gen architecture

The 24GB Tier (RTX 5090, 4090)

The flagship tier is where AI demand and consumer demand collide most directly.

Card	MSRP	Current Street Price	Change
RTX 5090	$1,999	$3,500-4,500	+75-125%
RTX 4090 (new)	$1,599	$2,000-3,000	+25-90%
RTX 4090 (used)	—	$1,800-2,200	Holding steady

The RTX 5090 situation is particularly dire. Rumors suggest prices could push toward $5,000 in Q1 2026 if supply doesn't improve. NVIDIA has stated they have no plans to officially raise MSRP, but that's irrelevant when no retailer sells at MSRP anyway.

Used Market Impact (RTX 3090, 3080)

The used market is the one bright spot—sort of. Prices have stabilized rather than crashed:

Card	Current Used Price	6-Month Trend
RTX 3090 (24GB)	$650-850	Stable/slight increase
RTX 3090 Ti (24GB)	$800-1,000	Stable
RTX 3080 12GB	$350-450	Stable
RTX 3080 10GB	$280-350	Slight decrease

The RTX 3090's 24GB of VRAM keeps its prices elevated. For local AI workloads requiring high VRAM, it remains the most practical entry point—and sellers know it.

What This Means for Local AI Hobbyists

The VRAM Squeeze Is Getting Worse

The cards local AI users need most—those with 16GB+ VRAM—are exactly the cards facing the worst supply constraints. NVIDIA is reallocating memory to higher-margin products, and consumer cards with large VRAM pools are low priority.

This is unlikely to change soon. As long as AI datacenter demand remains strong (and there's no sign of it slowing), consumer GPU availability will suffer.

Budget Builds Are Harder to Justify

Six months ago, you could plan a local AI build around an RTX 4060 Ti 16GB for ~$450 or wait for RTX 5060 Ti at a similar price point. That calculus has changed:

- RTX 5060 Ti availability is uncertain and may face production cuts
- Previous-gen 16GB cards aren't dropping in price as expected
- The gap between "budget" and "capable" keeps widening

If your budget is under $500, your best options are increasingly limited to used previous-gen cards or AMD alternatives.

Waiting May Cost More Than Buying

The typical advice during GPU transitions is "wait for prices to stabilize." That advice assumes prices will stabilize downward. Current trends suggest the opposite:

- GDDR7 shortages are structural, not temporary

- Production cuts of 30-40% are planned for H1 2026
- Datacenter demand continues to grow
- Tariffs add cost pressure regardless of supply

A card that costs $700 today may cost $850 in six months. The used RTX 3090 that's $700 now was $600 a year ago. Waiting isn't free.

Strategies to Deal With Rising Prices

Buy Now If You Need 24GB

If your use case requires 24GB VRAM (running 32B+ models, FLUX at full precision, LoRA training), buy now rather than waiting for prices to drop. They probably won't.

Best current options:

- **Used RTX 3090:** $650-800 — Best value for 24GB
- **Used RTX 4090:** $1,800-2,200 — Faster, but 2.5x the price
- **New RTX 5090:** $3,500+ — Only if money is no object

The used RTX 3090 remains the value king. At $700, you get the same 24GB VRAM as a $3,500 RTX 5090. Yes, the 5090 is significantly faster, but for most local AI workloads, VRAM capacity matters more than raw speed.

The Used Market Is Your Friend

Previous-generation cards offer the best price-to-VRAM ratio in the current market:

Card	Used Price	VRAM	Price/GB	Best For
RTX 3090	$700	24GB	$29	32B-70B models
RTX 3080 12GB	$400	12GB	$33	7B-13B models
RTX 3060 12GB	$180	12GB	$15	Entry-level, budget

Buy from eBay for buyer protection, test thoroughly within the return window, and you'll have capable hardware at a fraction of new prices.

Consider AMD (Finally)

For years, "just buy NVIDIA" was the only reasonable advice for local AI work. In 2025, that's finally changing.

AMD's RX 7900 XTX offers:

- **24GB VRAM** at $900-1,000 new
- **ROCm support** that has matured significantly
- **Compatibility** with Ollama, LM Studio, KoboldCpp, and other popular tools

The caveats are real: ROCm works best on Linux, setup is more technical than CUDA, and you may hit compatibility issues with cutting-edge models. But for users comfortable with Linux and willing to troubleshoot occasionally, the RX 7900 XTX is a legitimate alternative. Read our AMD vs NVIDIA comparison for the full picture.

Performance reaches approximately 80% of RTX 4090 speeds for LLM inference with ROCm 5.6+. That's a meaningful gap, but at potentially half the price for equivalent VRAM, it's worth considering.

Wait Strategically (When It Makes Sense)

Waiting makes sense in specific situations:

- **You only need 8-12GB VRAM:** Lower-tier cards aren't as supply-constrained
- **AMD's next generation interests you:** RDNA 4 may offer better AI performance
- **You can wait 12+ months:** Supply may normalize by late 2026
- **Your current hardware is adequate:** Don't upgrade out of FOMO

Waiting doesn't make sense if:

- You need 24GB VRAM now
- You're planning a build around mid-range 16GB cards (supply getting worse)
- You expect prices to drop significantly in 6 months (unlikely)

Decision Matrix

Your Situation	Recommended Action
Need 24GB, budget <$1,000	Buy used RTX 3090 now
Need 24GB, budget $1,500+	Consider used RTX 4090

Your Situation	Recommended Action
Need 16GB, can use Linux	Consider AMD RX 7900 XT ($700-800)
Need 12GB, tight budget	Buy used RTX 3060 12GB ($180)
Current setup works fine	Wait and monitor prices
Planning build for late 2026	Wait for supply to normalize

Price Predictions and What to Watch

Q1-Q2 2026 Outlook

Based on current trends, expect:

- **RTX 5090:** Prices likely to remain $3,000+ through mid-2026; rumors of $5,000 possible
- **RTX 5080:** May stabilize around $1,200-1,400 if supply improves slightly
- **RTX 5070 Ti / 5060 Ti:** Production cuts mean continued scarcity; prices volatile
- **RTX 4090 (used):** Likely to increase 10-20% as new stock depletes
- **RTX 3090 (used):** Expect gradual increase to $800-900 range
- **AMD RX 7900 XTX:** Stable or slight decrease; best value for 24GB if ROCm works for you

A DRAM shortage is expected to further impact consumer GPUs, with 10-20% retail price increases predicted for high-VRAM cards.

Signals That Prices Might Drop

Watch for:

- Memory manufacturers announcing GDDR7 capacity expansion
- NVIDIA announcing increased GeForce allocation
- Datacenter GPU demand cooling (unlikely near-term)
- AMD gaining significant market share (forces competitive response)
- New tariff negotiations reducing trade barriers

Signals That Prices Will Keep Rising

Watch for:

- Further production cut announcements (already rumored for 5060 Ti/5070 Ti)
- RTX 5000 Super series official cancellation
- Memory prices continuing upward
- Major AI companies announcing infrastructure expansion
- Trade tensions escalating further

Currently, more signals point toward continued price pressure than relief.

The Bottom Line

The GPU market is entering a structural shift that disadvantages local AI enthusiasts. NVIDIA has evolved into a datacenter company that happens to sell gaming cards, and when resources are constrained, consumer products take the hit.

For local AI users, the practical advice is:

1. **If you need 24GB VRAM:** Buy a used RTX 3090 now ($650-800). Don't wait for prices to drop—they're more likely to rise.
2. **If you need 16GB VRAM:** Act quickly on RTX 4060 Ti 16GB or consider AMD RX 7900 XT. Mid-range supply is tightening.
3. **If you're flexible on ecosystem:** AMD's RX 7900 XTX with ROCm is finally a viable alternative for Linux users willing to do some setup work.
4. **If you can wait 12+ months:** Supply may normalize by late 2026, but there's no guarantee prices will return to historical norms.
5. **If you're eyeing an RTX 5090 at MSRP:** Good luck. Realistically, budget $3,500+ or look elsewhere.

The era of cheap high-VRAM GPUs may be over. Plan accordingly.

Related Guides

- Used RTX 3090 Buying Guide for Local AI
- AMD vs NVIDIA for Local AI: Is ROCm Finally Ready?
- Build a Local AI PC for Under $500
- Local AI Planning Tool — VRAM Calculator

Source: https://insiderllm.com/guides/nvidia-gpu-prices-rising-2025/