GPU Training Speed Tiers (2026)

A quick, practical ranking for deep learning training throughput

PSA: Mind you, this list was **generated by Gemini 3 Pro** as a quick guide. Although the few comparisons I have checked so far hold when looking at serious testing sites, do check actual speeds yourselves before making any relevant decision. You can check sites like trooper.ai/benchmarks

Be aware of how large the VRAM of each GPU is, as this could break your code if you pretend to load a model + batch_size bigger than the VRAM available for the GPU that you choose.

Tier 1: The Hyperspeed Class *(100% – 90% Speed)*

The fastest silicon in existence. These cards chew through gradients.

1. **NVIDIA RTX 5090 (32GB)**

   Speed Score: 100/100

   Why: The Blackwell architecture + GDDR7 memory makes this significantly faster than the A100 in raw floating-point operations (FLOPS). If the batch fits, nothing beats this card in 2026.

2. **NVIDIA RTX 4090 (24GB) / NVIDIA A100 (40GB/80GB)**

   Speed Score: ~85–90/100

   The Comparison:

   - **RTX 4090:** Actually faster than the A100 in raw compute (TFLOPS) for FP16/BF16 training on smaller models.
   - **A100 (Colab Pro):** Wins on memory bandwidth (HBM2e). It feeds data to the chip faster. In practice, they are often neck-and-neck for training speed, with the A100 being more consistent for massive batches.

Tier 2: The High-Performance Class *(75% – 60% Speed)*

Excellent for serious Deep Learning work.

3. **NVIDIA RTX 5080 (16GB)**

   Speed Score: ~75/100

   Why: Faster cores than the 3090/4080, but the memory bus is narrower than the 4090. It has the raw horsepower of a Ferrari but the gas tank (VRAM) of a sedan.

4. **NVIDIA RTX 4080 Super / RTX 4080 (16GB)**

   Speed Score: ~65/100

   Why: A very fast card, significantly ahead of the 3090 in pure compute, though it lacks the memory bandwidth of the top tier.

5. **NVIDIA RTX 3090 Ti / RTX 3090 (24GB)**

   Speed Score: ~60/100

   Why: The old king. It relies on brute force and massive memory bandwidth (936 GB/s). It is roughly equal to the RTX 5070 Ti in speed but holds its ground due to that wide memory bus.

6. **NVIDIA V100 (16GB/32GB)**

   Speed Score: ~58/100

   Why: Found in older Colab Pro instances. It uses HBM2 memory (super fast), so for certain matrix multiplications, it can still beat modern consumer cards like the 4070 Ti.

Tier 3: The Mid-Range Workhorses *(50% – 35% Speed)*

Great for fine-tuning standard models (BERT, ResNet, Llama-7B).

7. **NVIDIA RTX 5070 Ti / RTX 4070 Ti Super**
   Speed Score: ~50/100
   Why: Fast modern cores, but narrow memory buses limit them compared to the 3090/4080 class.
8. **NVIDIA RTX 3080 Ti / RTX 3080 (10GB/12GB)**
   Speed Score: ~45/100
   Why: Still a beast. Its memory bandwidth (760 GB/s) is actually higher than the RTX 4070 series, making it surprisingly fast for training even in 2026.
9. **NVIDIA L4 (24GB)**
   Speed Score: ~42/100
   Why: The standard ”modern” Colab GPU. It is roughly equivalent to an RTX 3070 or 4070 in speed, but optimized for enterprise stability.
10. **Apple M3 Ultra (Mac Studio)**
    Speed Score: ~40/100
    Why: Warning: Benchmarking Apple vs. NVIDIA is tricky. In raw training (using MPS), the M3 Ultra behaves roughly like an RTX 3080 or 4070. It is much slower than a 4090, but it has the memory capacity of a server farm.
11. **NVIDIA RTX 5070 / RTX 4070**
    Speed Score: ~38/100
    Why: Very efficient, but strictly middle-of-the-road for training speeds.

Tier 4: The Entry Level *(30% – 15% Speed)*

Functional, but you will be waiting a while.

12. **Apple M4 Max / M3 Max**
    Speed Score: ~30/100
    Why: Roughly comparable to an RTX 3070 / 4060 Ti in training speed. Incredible for a laptop, but not a dedicated training monster.
13. **NVIDIA RTX 4060 Ti (16GB) / RTX 3070 Ti / RTX 3070**
    Speed Score: ~25/100
    Why: The 4060 Ti is popular for its 16GB VRAM, but its raw compute speed is quite low due to a tiny 128-bit memory bus. It is slower than the 3070 in many tasks.
14. **Apple M2 Ultra**
    Speed Score: ~25/100
    Why: Slower cores than the M3/M4, but massive bandwidth. Comparable to a 3070 Ti.
15. **NVIDIA RTX 3060 Ti / RTX 4060 / RTX 3060 (12GB)**
    Speed Score: ~20/100
    Why: The baseline. The 3060 (12GB) is beloved not for speed, but for fitting models. It is slow.

Tier 5: The "Slow Lane" *(<15% Speed)*

Patience is required.

16. **NVIDIA Tesla T4 (Colab Free Tier)**
    Speed Score: ~15/100
    Why: It uses the Turing architecture (2018). It is significantly slower than even a basic RTX 3060. It takes about 3–4x longer to train a model on a T4 than on a 3080.
17. **Apple M4 Pro / M3 Pro / M2 Max**
    Speed Score: ~12/100
    Why: Good for inference, weak for training.
18. **Apple M1 Ultra**
    Speed Score: ~10/100
    Why: First-gen silicon. It shows its age now.
19. **Apple M4 / M3 / M2 / M1 (Base Chips)**
    Speed Score: <5/100
    Why: These have no active cooling (MacBook Air) or very few GPU cores. Do not train on these.