

Setup: AWS EC2 GPU Instance (g6.xlarge)

For cost-efficiency, I used a g6.xlarge single-GPU instance.

Launch Steps:

1. **EC2 Console** - Launch new instance
2. **AMI Selection** - Use "Deep Learning Base AMI with Single CUDA (Amazon Linux 2023) 20260120"
 - Alternatively, choose any base OS and install drivers manually
 - Note: Deep learning Ubuntu 22.X variant had NVIDIA driver detection issues. So this did not work for me
3. **Instance Type** - Select g6.xlarge
 - **Important:** Request vCPU quota increase beforehand (approval may take time; good luck if you account is new)
4. **SSH Access** - Configure key pair (.pem file)
5. **Network** - Default settings; customized security group to allow SSH from my public IP
6. **Storage** - gp3, 125 GB root volume

Verification:

After connecting via SSH, run `nvidia-smi` to confirm GPU availability.

```
ec2-user@ip-172-31-20-236:~$ nvidia-smi
Mon Jan 26 21:08:26 2026

+-----+
| NVIDIA-SMI 580.126.09          Driver Version: 580.126.09    CUDA Version: 13.0     |
+-----+-----+
| GPU   Name                Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf            Pwr:Usage/Cap |      Memory-Usage | GPU-Util  Compute M. |
|               |              MIG M. |                     |
+-----+-----+
|  0  NVIDIA L4              On          | 00000000:31:00.0 Off |          0          |
| N/A   29C   P8             15W /  72W |  0MiB / 23034MiB |    0%    Default   |
+-----+-----+
|               |                     |                     |
+-----+-----+

Processes:
+-----+
| GPU   GI   CI        PID   Type   Process name                      GPU Memory |
| ID   ID                               |              Usage              |
+-----+-----+
| No running processes found          |
+-----+

[ec2-user@ip-172-31-20-236 ~]$ nvcc --version
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2025 NVIDIA Corporation
Built on Wed Aug 20 01:58:59 PM PDT 2025
Cuda compilation tools, release 13.0, V13.0.88
Build cuda_13.0.r13.0/compiler.36424714_0
[ec2-user@ip-172-31-20-236 ~]$
```

In this instance, python was already installed.