InternVL2

Install Conda environment:

conda create -n intern-vl2 python=3.11 -y && conda activate intern-vl2

Install required libraries:

pip install lmdeploy

pip install modelscope
pip install huggingface_hub
pip install timm

Download the Model;

❿ **download from Hugging face (https://huggingface.co/OpenGVLab/internVL2-8B)**

from huggingface_hub import snapshot_download
model_dir = snapshot_download(repo_id="OpenGVLab/InternVL2-8B", local_dir = "/home/modelscope/shan/InternVL2-8B/model-files")

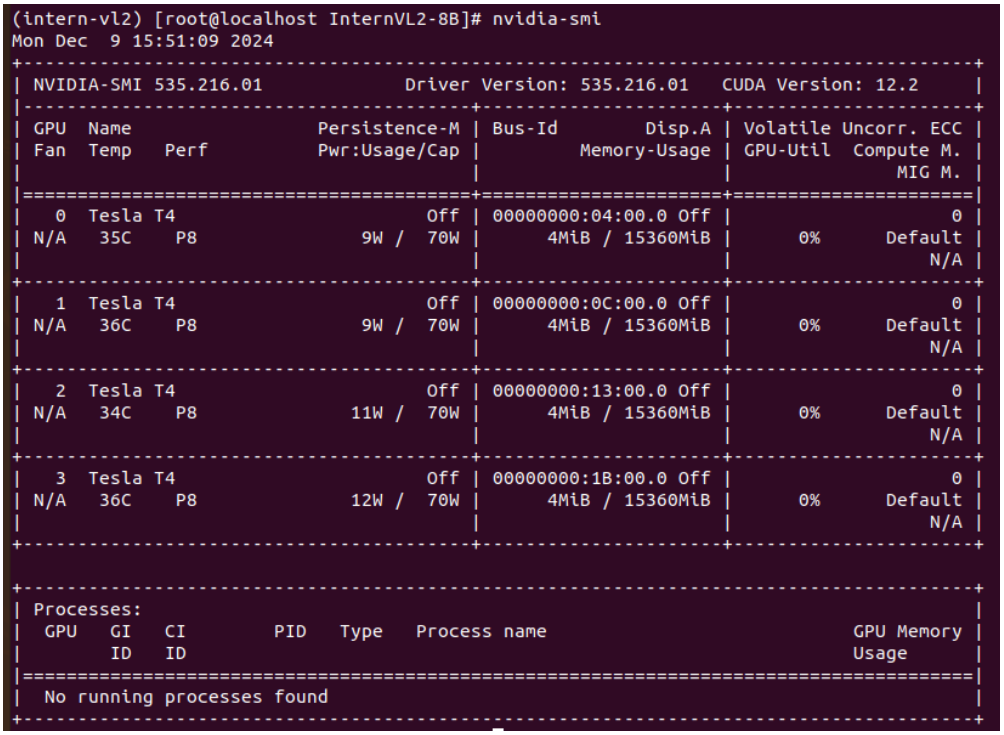
❿ **Download from Modelscope (**https://modelscope.cn/models/OpenGVLab/InternVL2-8B**)**

from modelscope import snapshot_download
model_dir = snapshot_download("OpenGVLab/InternVL2-8B", local_dir = "/home/modelscope/shan/InternVL2-8B/model-files")

❿ **download from Modelscope using command line**

create account and login
>> modelscope login --token 120ffdc5-1efb-400c-a8fb-e5789ce0985a
>> modelscope download --model OpenGVLab/InternVL2-8B --local_dir InternVL2-1B/model-files

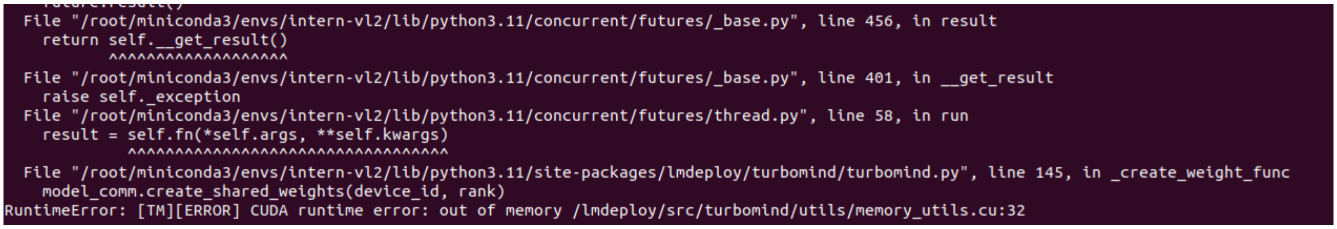
My Server GPU



```
(intern-vl2) [root@localhost InternVL2-8B]# nvidia-smi
Mon Dec  9 15:51:09 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.216.01             Driver Version: 535.216.01   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| 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  Tesla T4                       Off | 00000000:04:00.0 Off |                    0 |
| N/A   35C    P8               9W /  70W |      4MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   1  Tesla T4                       Off | 00000000:0C:00.0 Off |                    0 |
| N/A   36C    P8               9W /  70W |      4MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   2  Tesla T4                       Off | 00000000:13:00.0 Off |                    0 |
| N/A   34C    P8              11W /  70W |      4MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
|   3  Tesla T4                       Off | 00000000:1B:00.0 Off |                    0 |
| N/A   36C    P8              12W /  70W |      4MiB / 15360MiB |      0%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+


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

CUDA: Out of memory



```
    future.result()
  File "/root/miniconda3/envs/intern-vl2/lib/python3.11/concurrent/futures/_base.py", line 456, in result
    return self.__get_result()
           ^^^^^^^^^^^^^^^^^^^
  File "/root/miniconda3/envs/intern-vl2/lib/python3.11/concurrent/futures/_base.py", line 401, in __get_result
    raise self._exception
  File "/root/miniconda3/envs/intern-vl2/lib/python3.11/concurrent/futures/thread.py", line 58, in run
    result = self.fn(*self.args, **self.kwargs)
             ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/root/miniconda3/envs/intern-vl2/lib/python3.11/site-packages/lmdeploy/turbomind/turbomind.py", line 145, in _create_weight_func
    model_comm.create_shared_weights(device_id, rank)
RuntimeError: [TM][ERROR] CUDA runtime error: out of memory /lmdeploy/src/turbomind/utils/memory_utils.cu:32
```

Reference:
https://internvl.readthedocs.io/en/latest/internvl2.0/quick_start.html
https://modelscope.cn/models/OpenGVLab/InternVL2-8B/summary
https://github.com/OpenGVLab/InternVL
https://huggingface.co/OpenGVLab/internVL2-8B