□□□□□□ `-i https://mirrors.aliyun.com/pypi/simple/`

`nvidia-smi` □□□N□□□□GPU□□□□□□□□□□□□□□□□

`lscpu` □□□cpu□□□

- `P-core` □□□□□□□□□□□□□□□□□□□□□□□□2□□□□
- `E-core` □□□□□□□□□□□□□□□□

`lscpu | grep cache` □□□□□□□□□□

`ls` □□□□□□□□□□□□

`conda env list` □□□□□□□□□□□□□

`conda deactivate` □□□□□□□□

`pwd` □□□□□□□□

`conda create -n videollava python = 3.10 -y` □□□□□□□□videollava□□□□□

`mv oldname newname` □□□□□□□□

`conda remove --name myenv --all` □□□□□□□myenv□□□□□

`rm -rf your_folder_name` □□□□□□

`conda search qwen-vl-utils -c conda-forge` □search□□□□□□□□□□□□ `-c conda-forge` □□□□□

`conda install -c conda-forge qwen-vl-utils=0.0.11 -y` □ `-y` □□□□□

`uptime` □□□□□□□□□□

`uptime -s` □□□□□□□□□□

`top` □□□□□CPU□□□□□

- □1□□□CPU□□□□
  - us□□□□□□
  - sy□□□□□□
  - id□□□□

`history` □□□□□□□□

`df -h` □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

- `tmfps` □□□□□□□□□□□□□□□RAM□□□□□
- `/dev/nvme0n1p2` □□□□□□□
- `/dev/nvme0n1p1` □□□□□□□□□□
- `efivarfs` □□□□□□□□

`free -h` □□

- `total` □□□□
- `used` □□□□□□□□□□□□□□□□□□□□
- `free` □□□□□□□□□
- `buff/cache` □□□□□□□□□□□□□□□□□□□□
- `available` □□□□□□□□□

`nvcc --verison` □□□cuda□□

`tensorboard --logdir=./logs --port=6006 --bind_all`

- `--logdir`：□□□□□□□□□□ `logging_dir` □□□
- `--port`：□□□□□□□6006□
- `--bind_all`：□□□□□□□

`ps aux` □□□□□

- `ps` □Process Status□□□□□□
- `a` □□□□□□□□□□□□□□□□□□□□□□□□□□□□
- `u` □□□□□□□□□□□□□□□□□□□□□□□□□
- `x` □□□□□□□□□□□□□□□□□□□□□□□□□

`kill -9 1234` □□□□□PID□1234□□□

`ssh -L [□□□□]:[□□□□]:[□□□□] [□□□]@[□□□□□]` □□□□□□□□□□□□□□□□□□□□□□

□□□□□

`sudo ufw status` □□□□□□□□□□□□□□□□

`sudo ufw allow/deny 8080` □□□/□□8080□□

`sudo ufw enable/disable` □□□/□□□□□

□□□□□□

`sudo systemctl status nginx` □□□□□□□

`sudo systemctl start/stop/restart redis` □□□/□□/□□□

`sudo systemctl enable/disable redis` □□□/□□□□□

`tops` □□□□□□□□□□□□□□□□□□□□□□□□□□□/□

□□□□□

- 1 GOPS = 10□□/□
- 1 POPS = 1000□□□/□

□□□□
□□

□□	□□	□□	□□□□
□□□(FP32)	torch.float32	32	□□□□□□□□□
□□□(FP16)	torch.float16	16	□□
□□□(BF16)	torch.bfloat16	16	Ampere□□GPU□□
TF(32)	□□□□(□FP32)	19	NVIDIA Ampere□□□□□□□□□
INT8	**1**□□/□□□FP32□1/4□	8-bit□□	□□□□□□
INT4	**0.5**□□/□□□FP32□1/8)	4-bit□□	□□□□□□

□□

□□□□□□□□□□□□□□□□□□□□□□□□□□

□□	□□	□□□□
fp32	□□□□□	torch.float32
fp16	□□□□□	torch.float16
bf16	□□□□□	torch.bfloat16
tf32	□□□□□(□FP32)	□□□□

```
# 8-bit□□□□
        quant_config = BitsAndBytesConfig(
            load_in_8bit=True,
            bnb_4bit_compute_dtype=torch.bfloat16
        )
```

□□□□□□□□□□bfloat16□□□□□□□□□INT8-bit

CUDA

□□□□

- **CUDA Core**□GPU □□□□□□□□□□□□□□□□□□□□□□□□□□□□
- **SM**□**Streaming Multiprocessor**□□
  - GPU □□□□□□□□□ SM □□□□□ CUDA Core□□□□□□□□□□□□□

□□□□

- □□□□□**Global Memory**□□□□□□□□□□□□□□□□□□□□□□□□□□□ CPU □ RAM□□
- □□□□□**Shared Memory**□□SM □□□□□□□□□□□□□□□ CPU □□□□
- □□□□**Registers**□□□□□□□□□□□□□□□□□□□

□□□□

- **Kernel**□□□□□□□
  - □GPU□□□□□□□□□□
- □□□□
  - **Thread**□□□□□□□□
  - **Block**□□□□□□□□□□□□SM□□□
  - **Grid**□□□Block□□□□□□□□□□□□□Kernel□□

CUDA□□□□

- **CUDA**□□□□□□□□GPU□□□□□□□□□□
- **cuDNN**□□□□□□□□□□
- **cuFFT**□□□□□□□□□
- **cuBLAS**□□□□□□□□□□□
- **NVCC**□□□□□□□.cu□□□□□□GPU□□□□□□□□□□**GPU**□□□□□