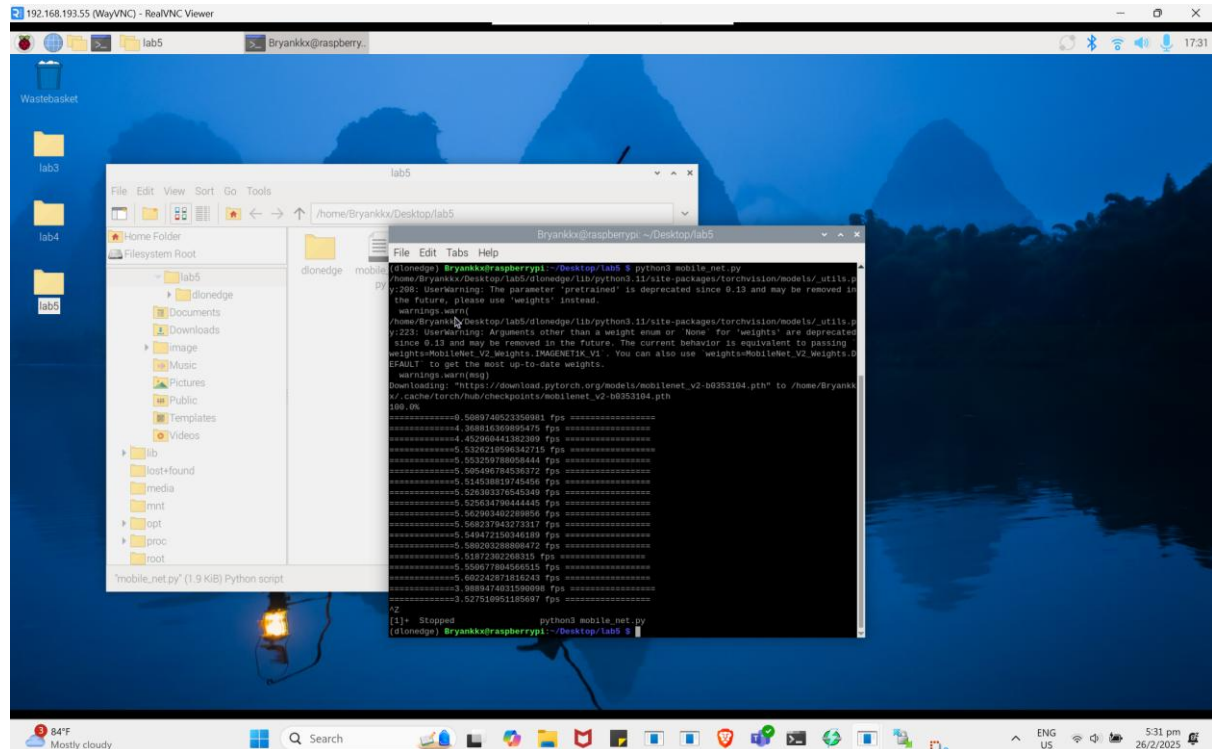
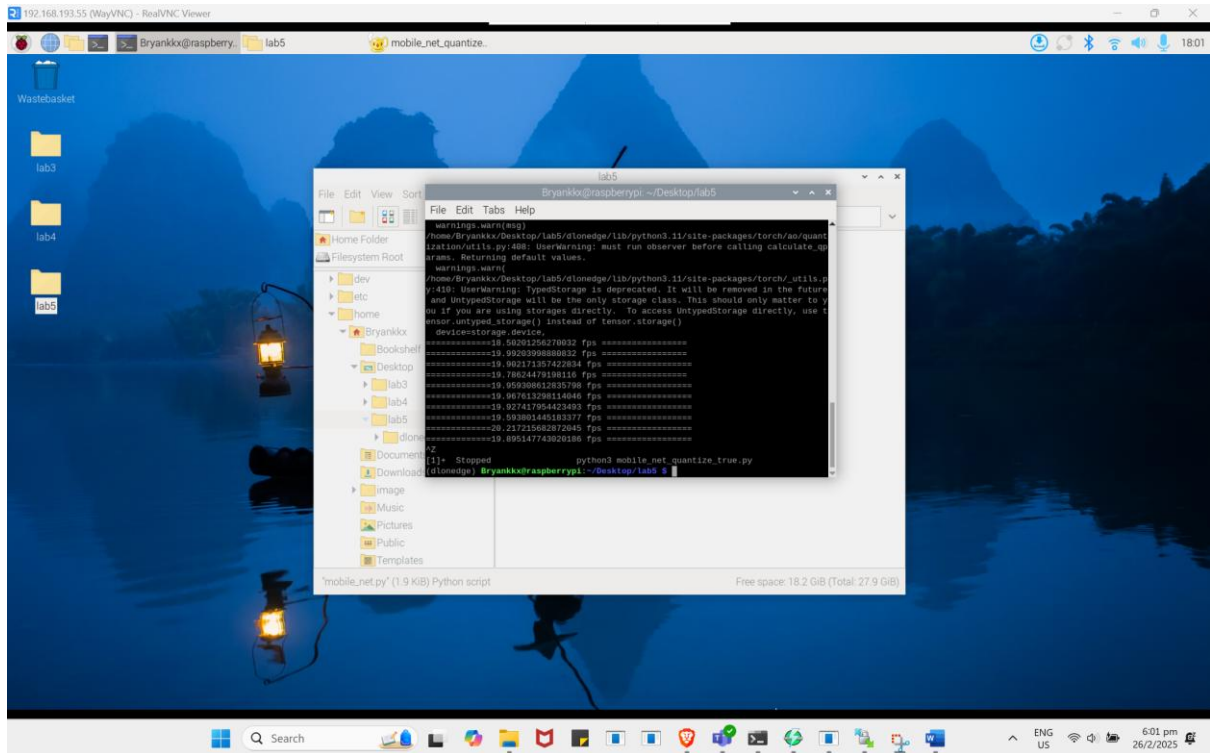


Part 1. [sample code](#) is used to directly load pre-trained MobileNetV2 model, doing model inference and finally, Observe the fps as shown in screenshot below when run on RaspberryPi 4B. As shown, with no optimization of model, we could only achieve of 5-6 fps much below our desired target.



```
(dionedge) Bryankkx@raspberrypi: ~/Desktop/lab5
/home/Bryankkx/Desktop/lab5/dionedge/lib/python3.11/site-packages/torchvision/models/_utils.py:208: UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'weights' instead.
  warnings.warn(
/home/Bryankkx/Desktop/lab5/dionedge/lib/python3.11/site-packages/torchvision/models/_utils.py:223: UserWarning: Arguments other than a weight enum or 'None' for 'weights' are deprecated since 0.13 and may be removed in the future. The current behavior is equivalent to passing weights=MobileNet_V2_Weights.IMAGENET1K_V1. You can also use 'weights=MobileNet_V2_Weights.DEFAULT' to get the most up-to-date weights.
  warnings.warn(msg)
Downloading: "https://download.pytorch.org/models/mobilenet_v2-0b353104.pth" to /home/Bryankkx/.cache/torch/hub/checkpoints/mobilenet_v2-0b353104.pth
100.0%
=====9.5089740523350981 fps=====
=====4.208815090909415 fps=====
=====4.452960441382309 fps=====
=====5.5326219596342715 fps=====
=====5.532879780904444 fps=====
=====5.595496784536372 fps=====
=====5.514538819745456 fps=====
=====5.526303216545349 fps=====
=====5.525834790444445 fps=====
=====5.562903402209856 fps=====
=====5.508227943273217 fps=====
=====5.549472150546189 fps=====
=====5.586203238808472 fps=====
=====5.51872362260315 fps=====
=====5.550677804586515 fps=====
=====5.602242871816243 fps=====
=====3.3089474031190008 fps=====
=====3.527510951185697 fps=====
^Z
[1] Stopped python3 mobile_net.py
(dionedge) Bryankkx@raspberrypi:~/Desktop/lab5 $
```

Part 2. Edit line number 11 as shown below to enable quantization in [sample code](#) to use quantized version of MobileNetV2 model.



- **Part 3.** Uncomment lines 57-61 in [sample code](#) to print the top 10 predictions in real-time as shown in below video

