Model	Params	MACs	FLOPs (2×MAC)	TorchScript size	Avg latency (ms, b=1 GPU)	p50 (ms)	p90 (ms)	Peak GPU mem (bytes)	Val acc
ResNet18	11,180,103	1,824,804,359	3.6496e9	44,887,426 B (~42.8 MiB)	**2.20 ms**	2.08	2.57	65,116,160 (~62.1 MiB)	**0.8233**
MobileNetV2	2,232,839	319,027,975	6.3806e8	9,450,712 B (~9.0 MiB)	**4.03 ms**	3.77	3.94	29,756,416 (~28.4 MiB)	0.7758
EfficientNet-B0	4,016,515	408,924,863	8.1785e8	16,803,222 B (~16.0 MiB)	**5.62 ms**	5.66	5.85	36,921,344 (~35.2 MiB)	0.7968

Quick interpretation (trade-offs)

- **ResNet18** — *best accuracy* (82.3%) and fastest inference by a comfortable margin (2.2 ms). Heavy-ish params and traced size, highest peak memory. Good if you can afford ~44MB model file and ~62MB GPU memory for models + activations.
- **MobileNetV2** — *smallest artifact* (≈9 MB traced), lowest memory footprint (~28 MB), slightly slower (~4.0 ms) than ResNet18 here (implementation/kernel differences make MobileNet slower on desktop GPU sometimes). Best candidate if you need small on-disk size and low memory.
- **EfficientNet-B0** — *middle ground* in params/size and best MACs-to-accuracy ratio for some classes — but slowest (~5.6 ms) in your runs. Good compromise if you want reduced params vs ResNet18 with similar accuracy trend.

Important: on small GPUs / edge devices, **params and MACs do not always equal lower latency** — kernel shapes, depthwise convs, and memory/launch overhead matter. Here ResNet18 ended up fastest on the RTX 3050.