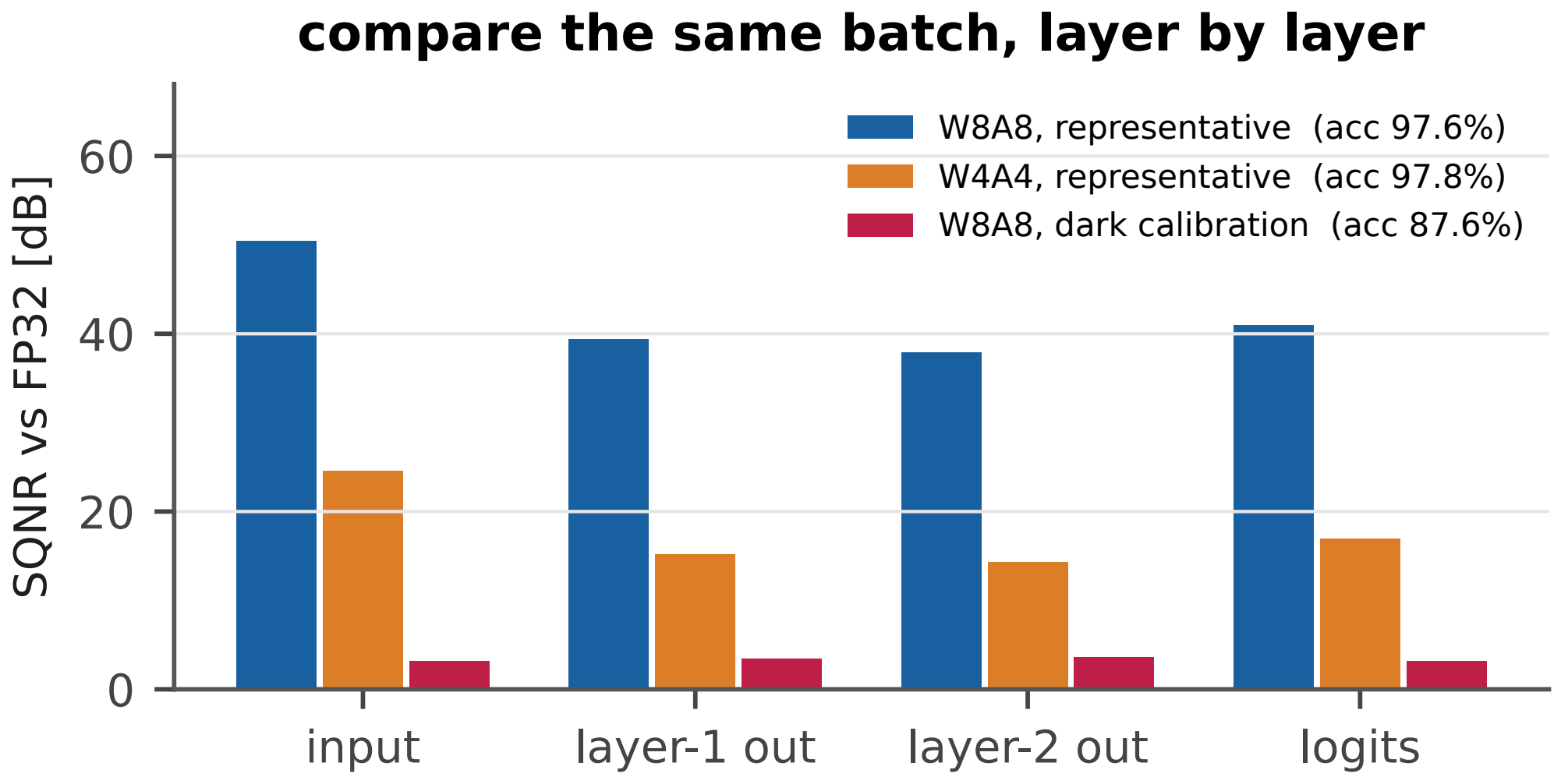

compare the same batch, layer by layer
W8A8, representative (acc 97.6%)
W4A4, representative (acc 97.8%)
W8A8, dark calibration (acc 87.6%)
SQNR vs FP32 [dB]
60
40
20
0
input
layer-1 out
layer-2 out
logits