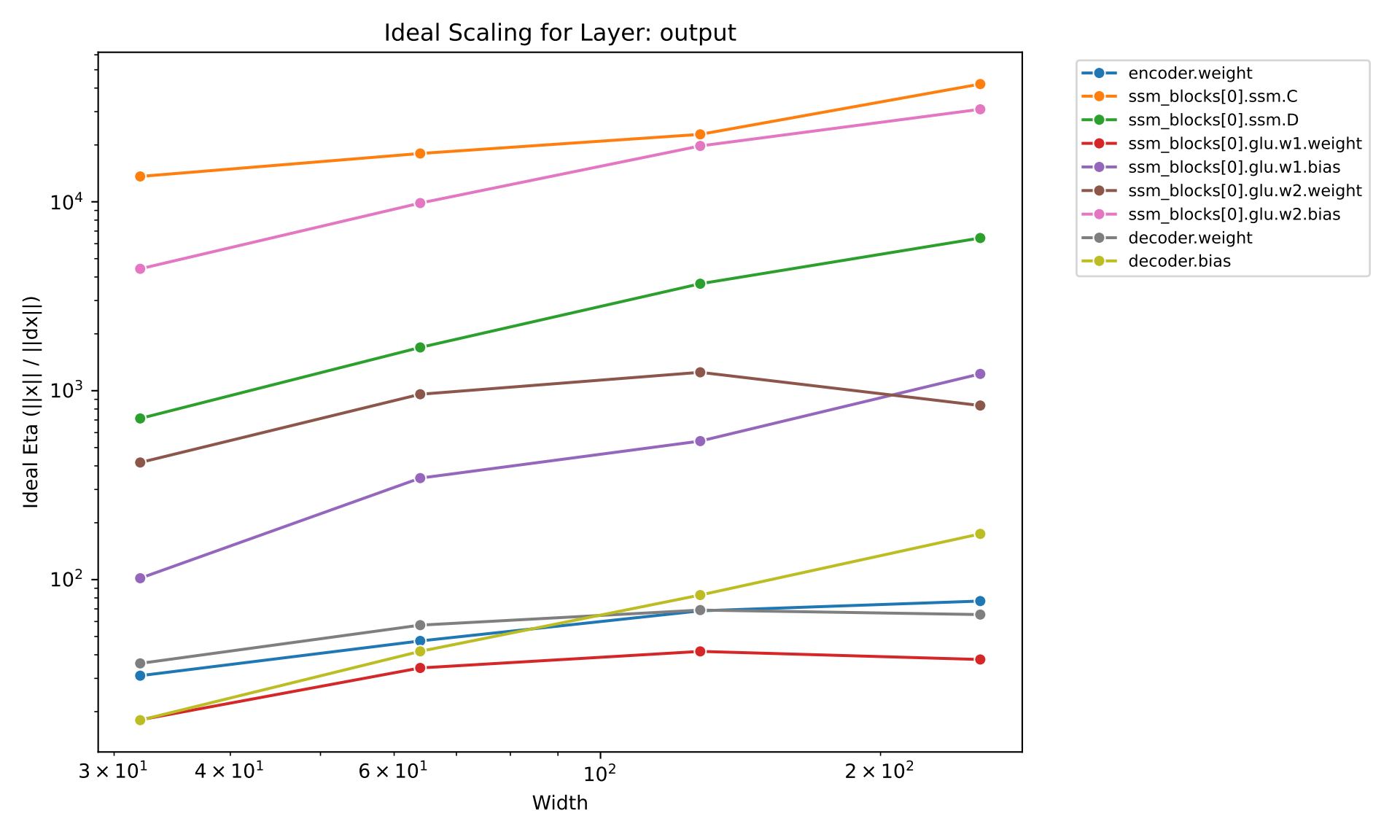
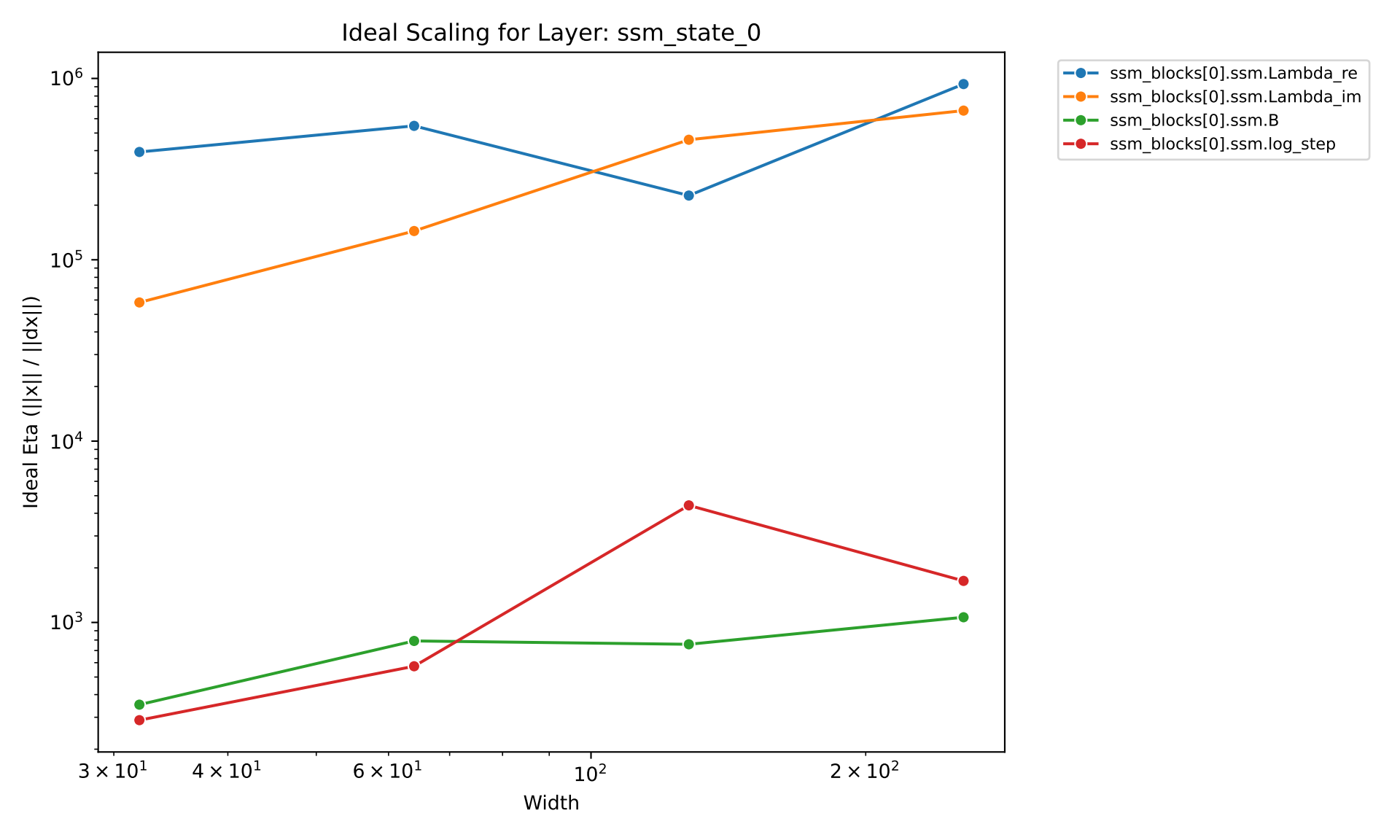

Ideal Scaling for Layer: output
Ideal Eta (||x|| / ||dx||)
Width
10^4
10^3
10^2
3×10^1
4×10^1
6×10^1
10^2
2×10^2
encoder.weight
ssm_blocks[0].ssm.C
ssm_blocks[0].ssm.D
ssm_blocks[0].glu.w1.weight
ssm_blocks[0].glu.w1.bias
ssm_blocks[0].glu.w2.weight
ssm_blocks[0].glu.w2.bias
decoder.weight
decoder.bias


Ideal Scaling for Layer: ssm_state_0
Ideal Eta (||x|| / ||dx||)
Width
10^6
10^5
10^4
10^3
3×10^1
4×10^1
6×10^1
10^2
2×10^2
ssm_blocks[0].ssm.Lambda_re
ssm_blocks[0].ssm.Lambda_im
ssm_blocks[0].ssm.B
ssm_blocks[0].ssm.log_step