Quantization Robustness: ResNet-BK vs Mamba
Perplexity vs Bit Width

Mamba INT4
PPL = 211.0
(Not deployable)

Model & Thresholds
- ResNet-BK
- Mamba
- Practical deployment threshold (PPL < 100)
- ResNet-BK INT4 (< 50 PPL)
- Mamba INT4 (> 200 PPL)

211.0

Practical deployment
threshold (PPL < 100)

Deployable Region

33.0
30.2

32.7
30.2

38.7
30.5

34.4

ResNet-BK INT4
PPL = 34.4
(Deployable)

Perplexity (lower is better)

Quantization Precision

FP32   FP16   INT8   INT4