```
Running on CUDA
Vocabulary Size: 30000
Train tokens: 500000 (after batchify)
--- ResNet-BK Ultra v2: O(N) + Hybrid Analytic Grad + Sparse MoE ---
Model Parameters: 4.15M
Total Steps (approx): 585
BKCore GRAD_BLEND = 0.5
  [Step 50] Epoch 1 | Loss: 7.4817 | LR: 0.000984
  [Step 100] Epoch 1 | Loss: 7.1682 | LR: 0.000937
  [Step 150] Epoch 1 | Loss: 7.2618 | LR: 0.000862
=========================================================
===
Epoch 1/3 | Time: 28.82s | Avg Loss: 7.6057 | Perplexity: 2009.60
=========================================================
===
  [Step 200] Epoch 2 | Loss: 7.0199 | LR: 0.000764
  [Step 250] Epoch 2 | Loss: 7.0463 | LR: 0.000652
  [Step 300] Epoch 2 | Loss: 7.0798 | LR: 0.000532
  [Step 350] Epoch 2 | Loss: 7.1368 | LR: 0.000413
=========================================================
===
Epoch 2/3 | Time: 24.11s | Avg Loss: 7.0517 | Perplexity: 1154.78
=========================================================
===
  [Step 400] Epoch 3 | Loss: 7.0109 | LR: 0.000304
  [Step 450] Epoch 3 | Loss: 6.9486 | LR: 0.000213
  [Step 500] Epoch 3 | Loss: 7.0623 | LR: 0.000146
  [Step 550] Epoch 3 | Loss: 6.9950 | LR: 0.000108
=========================================================
===
Epoch 3/3 | Time: 24.25s | Avg Loss: 7.0229 | Perplexity: 1122.06
=========================================================
===
```