Architecture Comparison: Training Perplexity 10⁵ Dynamic (gamma=0.5) MoD (gamma=0.5) Training Perplexity (Log Scale) 10 200 400 100 300 500 600 700 **Training Step**