Architecture Comparison: Training Loss SDT MoD STT Training Loss (Log Scale) 3×10^{0} 2 × 10⁰ 0.2 0.6 0.4 8.0 1.0 0.0 **Proportion of Training**