Weak-IV inference runtime scaling