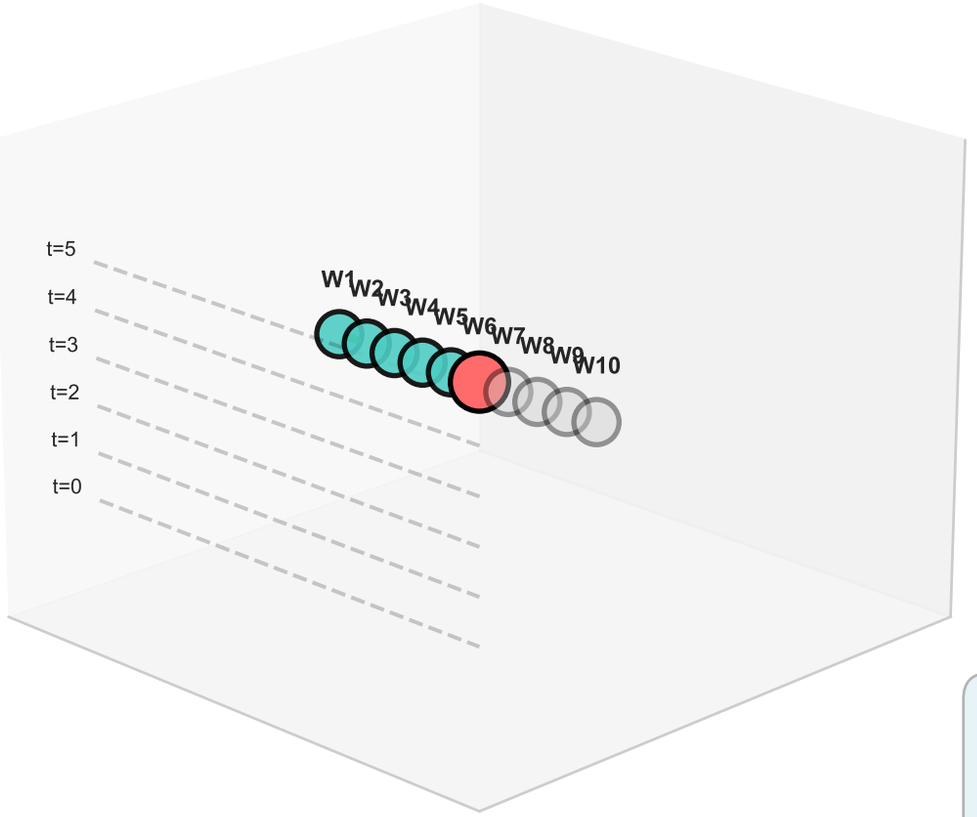
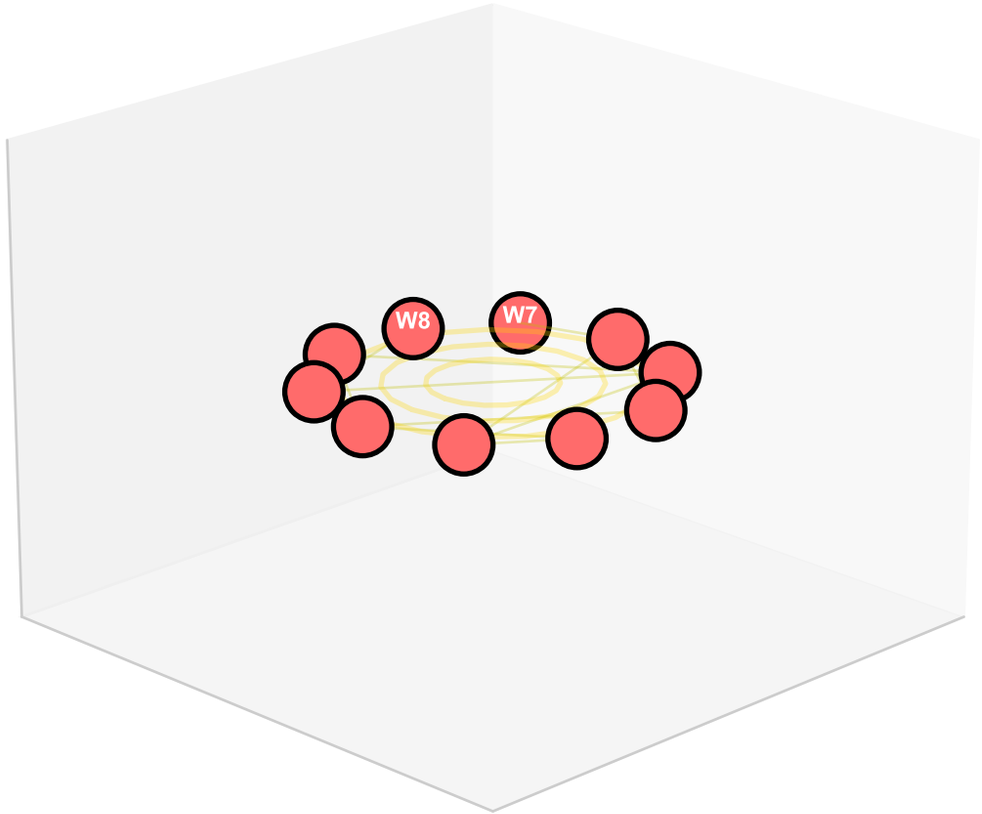
Processing Speed: Sequential vs Parallel

Sequential (RNN): One Word at a Time
Processing word 6 of 10 (Time step 6)



Parallel (Transformer): All Words at Once
Processing all 10 words simultaneously (Time step 1)



Sequential (RNN):
• 10 words = 10 time steps
• 100 words = 100 time steps
• GPU Utilization: ~5%
• Training: 90 days

Parallel (Transformer):
• 10 words = 1 time step
• 100 words = 1 time step
• GPU Utilization: ~95%
• Training: 1 day