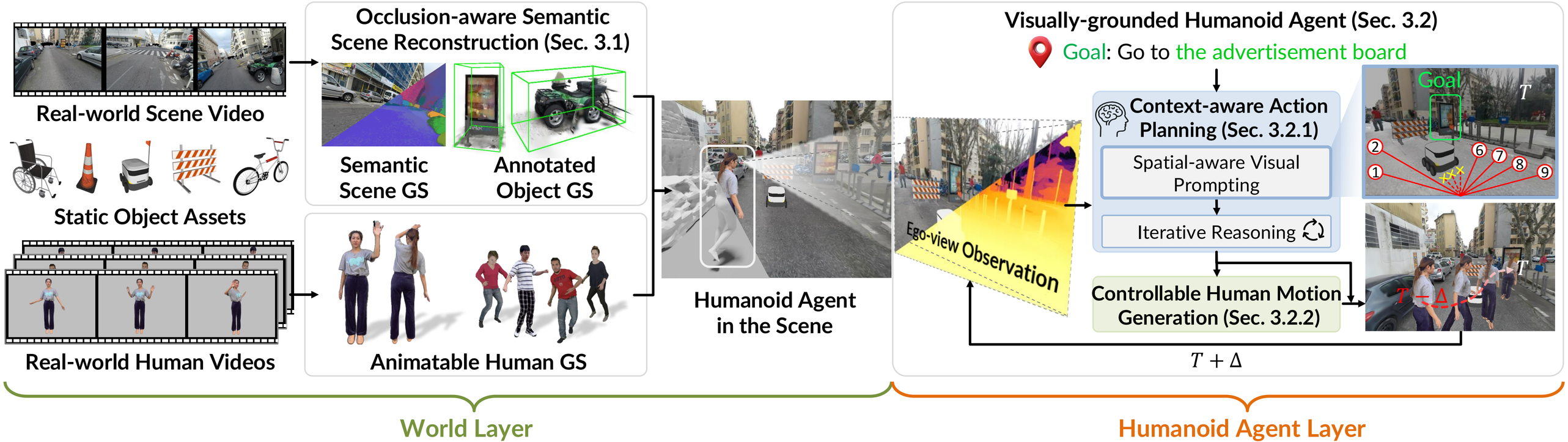
Real-world Scene Video
Static Object Assets
Real-world Human Videos
Occlusion-aware Semantic Scene Reconstruction (Sec. 3.1)
Semantic Scene GS
Annotated Object GS
Animatable Human GS
Humanoid Agent in the Scene
Visually-grounded Humanoid Agent (Sec. 3.2)
Goal: Go to the advertisement board
Ego-view Observation
Context-aware Action Planning (Sec. 3.2.1)
Spatial-aware Visual Prompting
Iterative Reasoning
Controllable Human Motion Generation (Sec. 3.2.2)
Goal
T
1
2
6
7
8
9
T + Δ
World Layer
Humanoid Agent Layer