(a) ALFWorld | Object and Light

Turn on a lamp while holding the requested object.



obs. 1



obs. 3



obs. 6



obs. 8

Claude Opus 5.5 Reported success | 8 code rounds | 80.9 s

(b) DiscoveryWorld | Archaeology

Date the artifacts and flag the site containing the oldest one.



obs. 1



obs. 13



obs. 31

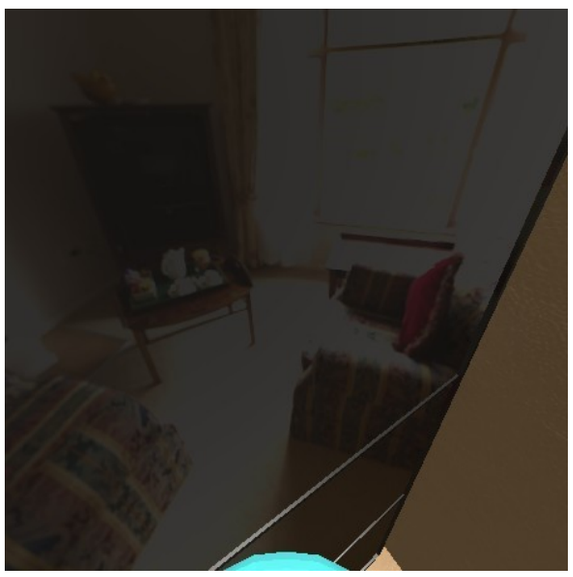


obs. 44

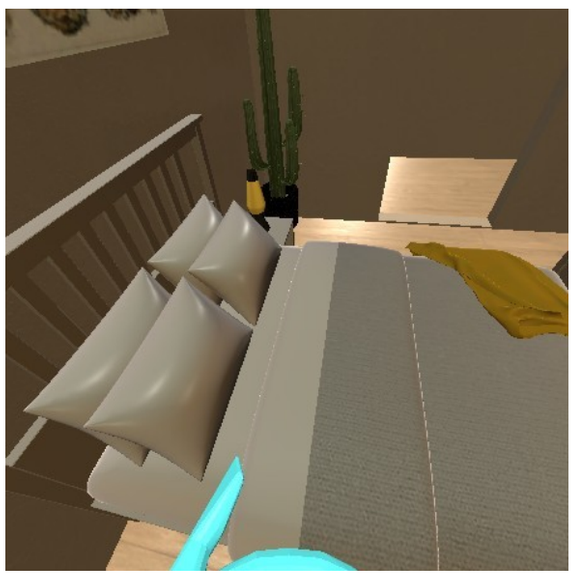
Claude Opus 5.5 Reported success | 7 code rounds | 167.9 s

(c) HumanCLAW | Sitting

Find the couch, approach until touching it, and sit.



obs. 1



obs. 20



obs. 49



obs. 69

Claude Opus 5.5 Task failure | 70 code rounds | 933.6 s