

## Article

# Agentic Presence Architecture and Deployment Audit for Proactive Embodied Agents in Extended Reality

Seonghoon Ban 

RECON Labs Inc., Gangnam-gu, Seoul 06060, Republic of Korea; seonghoon.ban@gmail.com

## Abstract

Proactive embodied agents in extended reality (XR) require an explicit interface between inferred user need and embodied intervention. We present Agentic Presence, a six-layer reference architecture integrating multimodal context, intent inference, local initiative arbitration, anticipatory scheduling, and embodied actuation. Its audit separates theoretical decision regions, sensor reachability, and decisions issued under runtime constraints. A Unity implementation on tethered Meta Quest 3 ran the arbiter at a configured 10 Hz, with median execution of four microseconds and decision-to-cue latency of at most 1 ms. Across 19,929 agentic-condition ticks, the audit identified cadence and speech gates as the dominant constraints. The system issued 75 interventions, comprising 44 ACT, 31 ASK, and no OFFER actions; the recovering presence-risk proxy remained at one on 98.5% of logged ticks. This separation established which mechanisms the deployment exercised and which required different task conditions. A within-subject pilot ( $n = 13$ ) estimated the complete initiative package relative to an identical reactive companion: a co-presence difference of 0.64 points (95% CI  $[-0.11, 1.40]$ ;  $dz = 0.51$ ), with a one-sided 95% upper bound of 0.70 points for increased intrusiveness. These preliminary estimates do not isolate arbitration timing or validate the risk proxy. The contribution is an implementable architecture and a reproducible deployment audit that connects formal policy design to realized sensing and runtime behavior.

**Keywords:** human–computer interaction; virtual reality; artificial intelligence; embodied agents; multimodal interaction; mixed-initiative interaction; initiative arbitration; social co-presence; extended reality

## 1. Introduction

Embodied conversational agents in extended reality (XR) can now sustain low-latency, context-aware interaction [1–4]. Proactive assistance nevertheless raises a distinct control problem: before generating content, an embodied agent must decide whether the present moment warrants intervention and how strong that intervention should be. Recent proactive AR and embodied-agent systems already address parts of this problem through context-aware triggering, temporal constraints, and dual-process control [5–9]. The open system question is therefore not whether timing exists, but how to compose intent confidence, interruption risk, reversibility, and socially meaningful waiting in an auditable runtime policy for an agent that visibly shares the user’s space.

Mixed-initiative research formalized acting versus asking as expected-utility decisions over uncertain user goals [10], while later systems learned interruption costs from behavioral sensors [11–13] and deferred notifications to task breakpoints [14]. That lineage primarily prices attention, task time, or annoyance. XR embodiment adds a related



Academic Editor: Janos Botzheim

Received: 13 August 2026

Revised: 2 October 2026

Accepted: 5 October 2026

Published: 15 October 2026

**Copyright:** © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

but distinct design concern: an intervention can disturb the user’s experience of the virtual place, while visible inaction by an apparently attentive companion can weaken the social relation.

We address these concerns through an architecture in which inference, arbitration, and actuation can be inspected separately. The arbitration interface includes an explicit cost-bearing WAIT action, diegetic reversibility, and a recovering presence-risk proxy. The evaluation distinguishes three questions: whether a policy mode exists mathematically, whether the implemented sensors can reach it, and whether it actually influences issued decisions. Social co-presence, spatial/place presence, intrusiveness, and the internal proxy are treated as distinct constructs. The present evidence supports implementation and deployment diagnostics; the pilot provides preliminary estimates of the complete initiative package.

We address this system problem with three contributions:

- (1) An auditable reference architecture and realtime implementation (Sections 4–6). Six typed layers connect multimodal sensing and context state to intent inference, local arbitration, anticipatory scheduling, and embodied actuation. Explicit utility terms and runtime overrides expose how ACT, OFFER, ASK, and WAIT are selected. The presence-risk stock is an implemented control extension, not a validated psychological measurement.
- (2) A sensor-to-policy deployment audit (Section 7). Decision-region analysis is paired with sensor reachability, override accounting, realized mode occupancy, and frozen-input replay. In 19,929 C3 ticks, the audit identifies the dominant cadence and speech constraints and distinguishes unused or inactive mechanisms from those exercised in the deployed task. Timing and resource measurements establish the feasibility of keeping arbitration local while inference runs asynchronously.
- (3) Preliminary user-experience estimates for the complete initiative package (Sections 8 and 9). A within-subject pilot ( $n = 13$ ) contrasts no companion, a reactive companion, and the initiative-enabled system during seated visual search. The C3–C2 co-presence estimate is 0.64 points (95% CI [−0.11, 1.40]); intrusiveness is interpreted relative to an explicitly study-specific margin. The design holds embodiment constant in C2–C3 but does not isolate timing from the frequency, content, or modes of intervention.

Table 1 places this architecture among selected adjacent systems. The contribution is the explicit connection between runtime interfaces, interpretable arbitration terms, and an audit of their realized operation. The comparison is illustrative, rather than a systematic review or a claim that initiative, interruption costs, or waiting are new concepts.

**Table 1.** Illustrative design-space comparison of selected adjacent systems. The table motivates the specific composition evaluated here but is not a systematic review or a proof of firstness. Symbols: ✓ = supported; △ = partial or qualified support; ✗ = absent or not applicable to the evaluated system; n/r = not reported. Parentheses specify the relevant restriction. The ProAct row describes its physical-robot evaluation. Latency entries use the respective studies’ endpoints, not a common benchmark.

| System           | Embodied in XR | Self-Initiated | Silence Modeled | Interruption/Presence Cost | Presence Measured | Continuous Multimodal Sensing | E2E Latency Class |
|------------------|----------------|----------------|-----------------|----------------------------|-------------------|-------------------------------|-------------------|
| LLMR [3]         | ✗ (authoring)  | ✗              | ✗               | ✗                          | ✗                 | ✗                             | ~60 s             |
| DreamCodeVR [15] | ✗              | ✗              | ✗               | ✗                          | ✗                 | ✗                             | ~30 s             |
| GazePointAR [4]  | ✗              | ✗ (wake word)  | ✗               | ✗                          | ✗                 | △ (at query)                  | 7.5 s             |

Table 1. Cont.

| System                  | Embodied in XR     | Self-Initiated                    | Silence Modeled          | Interruption/Presence Cost | Presence Measured                | Continuous Multimodal Sensing | E2E Latency Class              |
|-------------------------|--------------------|-----------------------------------|--------------------------|----------------------------|----------------------------------|-------------------------------|--------------------------------|
| XR-Objects [16]         | ✗                  | ✗                                 | ✗                        | ✗                          | ✗                                | △                             | ~3 s                           |
| EmBARDiment [17]        | ✓                  | ✗<br>(push-to-talk)               | ✗                        | ✗                          | ✗                                | ✓ (gaze memory)               | n/r                            |
| ELLMA-T [18]            | ✓                  | scripted                          | ✗                        | ✗                          | qualitative                      | ✗                             | slow                           |
| Satori [5]              | ✗ (overlay)        | ✓ (task pacing)                   | ✗                        | ✗                          | ✗                                | ✓                             | 2–3 s                          |
| Sensible Agent [6]      | ✗ (overlay)        | △ (what/how only)                 | ✗                        | ✗                          | ✗                                | ✓                             | n/r                            |
| YETI [8]                | ✗                  | ✓ (offline)                       | △ (binary)               | ✗                          | ✗                                | ✓ (1 fps)                     | offline                        |
| ProAgent [7]            | ✗ (glasses)        | ✓ (threshold)                     | △                        | dedup only                 | ✗                                | ✓ (tiered)                    | 4.5 s                          |
| Inworld/Convai [1, 2]   | ✓                  | developer-triggered               | ✗                        | ✗                          | ✗                                | speech (+Convai vision)       | <1 s (reactive)                |
| NVIDIA ACE [19]         | ✓                  | △<br>(autonomous game characters) | ✗                        | ✗                          | ✗                                | speech + vision               | <1 s (reactive)                |
| ProAct (robot) [9]      | ✗ (physical robot) | ✓ (motivation)                    | △                        | ✗                          | ✓ (social)                       | ✓                             | 3.7 s<br>(annotator-estimated) |
| Agentic Presence (ours) | ✓ (head-mounted)   | ✓ (arbitrated)                    | ✓<br>(cost-bearing WAIT) | ✓ (presence-risk proxy)    | ✓ (social co-presence; pilot DV) | ✓                             | <1 s (reactive)                |

## 2. Related Work

### 2.1. Presence, Co-Presence and Breaks

Presence includes distinguishable constructs such as place illusion, plausibility, spatial presence, and social co-presence [20–26]. How strongly a virtual other is experienced as a social actor depends on its perceived agency and behavioral realism [27–30], which motivates holding embodiment constant while comparing the complete initiative package with reactive interaction. Breaks in presence (BIPs) can be elicited and counted [31], but recent evidence shows that BIP episodes differ in intensity, trajectory, and recovery rather than following a single uniform pattern [32]. A fixed damage-and-recovery stock should therefore be read as an engineering control abstraction, not as a validated psychological law. The present study measures social co-presence as its primary outcome and retrospective BIP counts only as an exploratory measure; neither is identical to the internal risk proxy used by the arbiter.

### 2.2. From Interruption Cost to Proactive Assistance

Horvitz formalized mixed initiative as an expected-utility comparison [10], and successor systems operationalized attentional state, personalized interruption cost, sensor-based interruptibility, and notification deferral [11–14]. Anticipatory mobile computing generalized this lineage to proactive service timing on personal devices [33]. More recent proactive agents learn when to suggest tasks, enter conversations, or assist with programming [34–40]. Proactive AR systems likewise use user models, multimodal context, temporal constraints, or learned proactivity scores [5–8]. These works establish that timing is an active research topic. Our contribution connects these arbitration concepts to an embodied XR implementation and evaluates their sensor reachability and runtime activation. The recovering stock

is one inspectable extension within this architecture; its inclusion does not establish an experiential benefit.

### 2.3. Embodied XR Agents and Adjacent Proactive Systems

LLM-driven systems author and modify XR worlds [3,15], resolve referents from gaze and pointing [4], and attach intelligence to physical or virtual objects [16]. Embodied XR agents use gaze context in dialog [17] and support learning in social VR [18]. Proactive AR systems select assistance content and timing [5–8], while ProAct links proactive behavior to social-presence outcomes in a physical robot [9]. ProAct [9] is included for its physical-robot evaluation of proactivity and social presence, which is relevant to embodied XR agents. Table 1 compares that evaluated deployment; the journal version also demonstrates a Blender-based virtual character, not a head-mounted XR evaluation. Agentic Presence differs in its combination of a visible XR embodiment, explicit costs for intervention and waiting, diegetic reversibility, typed runtime interfaces, and a reachability audit; Table 1 summarizes this selected comparison without claiming exhaustive coverage.

### 2.4. Timing in Language-Agent and XR Architectures

CoALA organizes language agents by memory, action space, and decision procedure [41]; Generative Agents add memory, reflection, and planning [42]; and Voyager formalizes skills-as-code [43]. These abstractions do not by themselves specify a realtime intervention policy, although adjacent proactive-agent work does model triggers and temporal constraints [5–9,34–37]. We position L4 as an implementation-level arbitration service between intent inference and embodied action, with L5 providing anticipatory scheduling. The contribution is the concrete composition and audit of this layer rather than a claim that prior agent architectures contain no notion of timing.

## 3. Agentic Presence Design Requirements

**Definition.** Agentic Presence is a design objective for XR systems in which an embodied agent exercises initiative while seeking to support social co-presence and limit agent-attributable disruption. The objective requires continuous context sensing and an explicit choice among acting, offering, asking, and waiting; its achievement must be evaluated empirically rather than assumed from the architecture.

Four engineering requirements guide the implementation: R1—continuous passive sensing: context updates without user summons; R2—arbitrated initiative: intervention modes and a cost-bearing WAIT are explicit outputs; R3—auditable risk accounting: the interface can represent interruption cost and a recovering proxy derived from intervention history; and R4—legible, reversible action: self-initiated behaviors use a stable perceptual signature and an in-world path for correction or withdrawal. The deployed build implements all four, but the evaluation separately reports which mechanisms were active. R3 does not imply that the proxy measures psychological presence or is needed in every deployment.

## 4. An Explicit Arbitration Model for the Reference Architecture

### 4.1. Setting

At each arbitration tick  $t$ , the system holds multimodal evidence  $E_t$ ; an inferred user need  $G$  with confidence  $p_t = P(G | E_t)$ ; an interruption proxy  $i_t$  in  $[0,1]$ ; a need-salience estimate  $s_t$  in  $[0,1]$ ; and a recovering presence-risk stock  $P_t$  in  $[0,1]$ .  $P_t$  is an internal bookkeeping variable, not an observed presence score. The arbiter selects a mode  $m$  in  $\{\text{ACT, OFFER, ASK, WAIT}\}$  for the current best candidate action  $a$ .

#### 4.2. Expected Utility with Three Design Terms

$$EU(m) = pt \cdot U(m, G) + (1 - pt) \cdot U(m, \neg G) - C_{pres}(m, it, Pt) \quad (1)$$

(i) Diegetic reversibility. Wrong autonomous action is penalized by how hard it is to undo within the fiction:

$$U(\text{ACT}, \neg G) = -\kappa_A \cdot (1 - r(a)), r(a) \in [0,1] \quad (2)$$

where  $r(a)$  is the diegetic reversibility of the candidate autonomous action: adding a removable highlight approaches one; altering shared world state mid-narrative approaches zero. Similarly  $U(\text{OFFER}, \neg G) = -\kappa_O$  (a misleading pointer to the wrong object),  $U(\text{ASK}, \neg G) = -\kappa_Q$  (a needless question, which the user's answer immediately corrects), and  $U(\text{WAIT}, \neg G) = 0$ . The ordering  $\kappa_Q < \kappa_O$  is not cosmetic: a wrong question is self-correcting, whereas a wrong suggestion sends the user to the wrong place. Section 4.3 shows that reversing it collapses the ladder. Reversibility as an action-selection penalty is not itself new—it is well developed in reinforcement learning, where reachability-based side-effect penalties and learned reversibility estimators serve the same structural role [44,45]. What  $r(a)$  measures is different, and the two come apart in a way worth stating concretely: deleting an object the agent has just spawned is fully reversible in state space, yet potentially disruptive to the fiction, because the user has seen it appear and disappear. Such a disruption is a design risk, not an automatically observed psychological break. Diegetic reversibility asks whether the undo is representable inside the fiction, not whether the prior state is reachable. Interactive-narrative systems have long weighed a closely related question when choosing between silently substituting effects and replanning around a user's action [46], though as a heuristic rather than a priced term. The system evaluated here deliberately restricts its repertoire to highly reversible behaviors ( $r \in [0.70, 0.95]$ ; the companion never manipulates objects), so the low-reversibility regime of this term is examined by numerical sensitivity analysis (Section 7.4) rather than with participants.

(ii) Silence is not free. For a co-present embodied agent, failing to respond to a salient need is itself social behavior:

$$U(\text{WAIT}, G) = -\lambda \cdot st \quad (3)$$

WAIT already has precedents in mixed-initiative utility models and in costs of notification deferral and conversational gaps [10,11,47]. Here its cost is an explicit function of inferred need confidence and need salience, allowing the architecture to represent the design concern that a visibly available companion may fail to help. Socially appropriate assistance is also studied in robotics [48]. Equation (3) makes this concern inspectable as an engineering utility term; the present pilot does not identify its weight or establish the psychological validity of a social cost of silence.

(iii) Recovering presence-risk proxy. Intervention cost is scaled by an internal stock that records recent agent-attributable risk. We write  $b_t$  in  $\{0,1\}$  for a proxy-break event, registered when the agent intervenes while the interruption estimate exceeds  $i^*$ , and  $\delta$  for the decrement applied to the stock (Table 2;  $i^* = 0.45$  and  $\delta = 0.25$ ):

$$C_{pres}(m, it, Pt) = \sigma(m) \cdot it \cdot d(Pt) \quad (4)$$

$$Pt + 1 = \text{clip}[\min(1, Pt + (\Delta t/\tau)(1 - Pt)) - b_t \delta, 0, 1] \quad (5)$$

where  $\sigma(m)$  is the mode's perceptual salience, ordered as  $\sigma(\text{WAIT}) = 0 < \sigma(\text{OFFER}) < \sigma(\text{ASK}) < \sigma(\text{ACT})$ ;  $\tau$  is a recovery time constant; and  $d(P_t) = d_0(2 - P_t)$  increases cost while the proxy is depleted. This myopic scalarized policy consults the current stock but does not

optimize the future value of  $P_{t+1}$ . Both the fixed decrement and single recovery constant are deliberately simplified engineering choices. Because empirical BIPs are heterogeneous [32],  $P_t$  should not be interpreted as a validated estimate of a participant's momentary presence.

**Table 2.** Engineering constants of the arbitration policy, calibrated on pilot traces and frozen before participant data collection. The values define the reported decision regions; they are design choices rather than psychological parameter estimates.

| Constant                                                     | Value         | Role                                                                                                        |
|--------------------------------------------------------------|---------------|-------------------------------------------------------------------------------------------------------------|
| $U(\text{ACT}, G)/U(\text{OFFER}, G)/U(\text{ASK}, G)$       | 1.0/0.6/0.45  | Value of a correct intervention, by mode                                                                    |
| $\kappa_A/\kappa_O/\kappa_Q$                                 | 1.0/0.20/0.10 | Cost of a wrong intervention; the ratio $\kappa_Q/\kappa_O = 0.5$ keeps ASK reachable (Section 7.4)         |
| $\lambda$                                                    | 0.6           | Silence-penalty scaling in Equation (3)                                                                     |
| $\sigma(\text{ACT})/\sigma(\text{ASK})/\sigma(\text{OFFER})$ | 0.7/0.4/0.3   | Perceptual salience; $\sigma(\text{WAIT}) = 0$                                                              |
| $d_0$ , with $d(P) = d_0(2 - P)$                             | 0.5           | Risk-stock cost scale in Equation (4)                                                                       |
| $i^*/\delta$                                                 | 0.45/0.25     | Proxy-break threshold and stock decrement in Equation (5); engineering constants calibrated on pilot traces |
| $\tau$                                                       | 90 s          | Risk-stock recovery time constant; an unvalidated engineering constant                                      |
| $r(a)$ for the ACT-mode candidate                            | 0.70          | Only the ACT candidate's reversibility enters Equation (2); the full repertoire spans 0.70–0.95             |
| Minimum inter-intervention gap $g$ ; per-block cap $N$       | 30 s; 8       | Overrides in Equation (6)                                                                                   |

#### 4.3. Mode Regions Under the Calibrated Utility Model

At fixed need salience, each utility is affine in  $p_t$  and  $i_t$ , so the four predefined outputs partition the decision plane into linear regions (Figure 1). Under the deployed constants, WAIT occupies the low-confidence region; increasing interruption can move the argmax from ACT toward ASK or OFFER; and active speech must still be handled by the hard WAIT override of Section 4.4. These properties belong to the stated utility ordering and parameter regime, not to the functional form in general.

All four regions are non-empty only for part of the parameter space. ASK disappears when its wrong-intervention cost approaches that of OFFER; ACT dominates when reversibility approaches one; and OFFER requires  $i_t d(P_t) > 1/6$ . A mode can therefore exist mathematically yet remain unreachable under the sensing map. Section 7.4 evaluates the intersection between theoretical regions and realized sensor values, which is the relevant property of the deployed artifact.

The stock produces model-level hysteresis only where the interruption term is non-zero. At a representative point ( $i_t = 0.34$ ,  $s_t = 0.25$ ,  $r = 0.70$ ), depletion from  $p = 1.0$  to  $p = 0.5$  shifts the first-intervention threshold from  $p = 0.24$  to  $p = 0.29$  and can soften ACT to OFFER. Because confidence is quantized and the deployed task rarely produces high interruption, Section 7.5 tests rather than assumes whether this theoretical shift altered logged decisions.

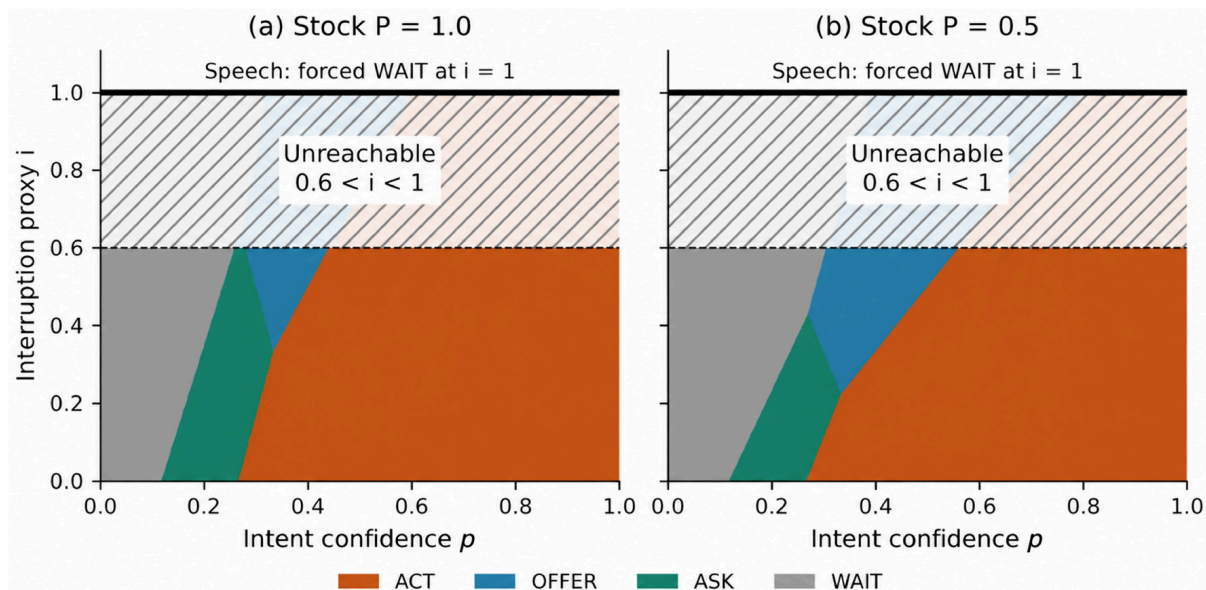


**Figure 1.** Decision regions produced by the calibrated utility model over the confidence–interruption plane at need salience  $s = 0.5$  and ACT-mode reversibility  $r = 0.70$ . Panel (a) fixes the risk stock at  $p = 1.0$ ; panel (b) fixes it at  $p = 0.5$ . Boundaries shift only where the interruption term is active. The sensing map permits  $i_t$  from 0 to 0.6 during non-speech. Values  $0.6 < i_t < 1$  are unreachable; active speech sets  $i_t = 1$  and forces WAIT independent of the utility argmax. These cases are distinct from the four utility-selected mode regions.

#### 4.4. Legibility Constraint

The utility layer of Sections 4.2 and 4.3 is not the deployed decision function. Three constraints override it, and we state them explicitly rather than folding them into the utilities, because the distinction matters for how the evaluation should be read. A fixed perceptual signature precedes every self-initiated behavior—a consistent vocalization-and-motion cue—since adaptive agents are rated worse than static ones when their variability reads as inconsistency [49], and a stable signature lets users form a partner model of an adaptive policy. A minimum interval separates interventions, and a per-block cap bounds their number. Most consequentially, active user speech forces WAIT outright:

$$m^* = \text{WAIT if speech active, or } \Delta t_{\text{since}} < g, \text{ or } n_{\text{block}} \geq N; \text{ else } \operatorname{argmax}_m EU(m) \quad (6)$$

with  $g = 30$  s and  $N = 8$ . The speech override exists because of the third property in Section 4.3: above moderate confidence the utility layer would intervene mid-utterance, and turn-taking is not a quantity we are willing to trade against expected utility. It follows that the policy evaluated with participants is Equation (6), not Equations (1)–(4) alone. We therefore report in Section 7.2 how often each override fired, so that the contribution of the utility layer can be separated from the contribution of the constraints, and the sensitivity analysis of Section 7.4, which evaluates the utilities over specified input grids, is explicitly an analysis of the utility layer rather than of the deployed policy.

#### 4.5. Practical Instantiation

Confidence  $p_t$  comes from schema-constrained LLM output, quantized to three levels, smoothed across ticks sharing a target, decayed between inferences, and invalidated when the target is resolved. The interruption and salience proxies are computed locally. Constants were calibrated in pilot traces and frozen before participant data collection.  $P_t$  is updated only from the agent’s own intervention history; the study does not establish criterion validity against an independently observed presence state.

Algorithm 1 is the deployed decision function evaluated at 10 Hz. Lines 5–10 implement the utility layer, lines 12–14 apply the speech, cadence, and cap overrides, and lines 3 and 19 update the recovering risk-stock proxy. Section 7 separates decisions produced by the utility layer from those determined by overrides.

---

**Algorithm 1** The deployed decision function evaluated at 10 Hz

---

```

1 input p_t, i_t, s_t, P_t, candidate action a, clock t
2 # Equation (5), recovery half
3 P_t <- min (1, P_t + (dt/tau) × (1 − P_t))
4 # Equations (1)–(4), utility layer
5 x <- i_t × d0 × (2 − P_t)
6 EU [ACT] <- p_t × U_ACT − (1 − p_t) × kappa_A × (1 − r(a)) − sigma_ACT × x
7 EU [OFFER] <- p_t × U_OFFER − (1 − p_t) × kappa_O − sigma_OFFER × x
8 EU [ASK] <- p_t × U_ASK − (1 − p_t) × kappa_Q − sigma_ASK × x
9 EU [WAIT] <- −lambda × p_t × s_t
10 m <- argmax_m EU [m]
11 # Equation (6), legibility overrides
12 if speech_active: m <- WAIT; reason <- SPEECH
13 elif t − t_last < g: m <- WAIT; reason <- CADENCE
14 elif n_block ≥ N: m <- WAIT; reason <- CAP
15 # actuation and Equation (5), damage half
16 if m ≠ WAIT:
17 emit fixed legibility cue; execute m with a
18 t_last <- t; n_block <- n_block + 1
19 if i_t > i_star: P_t <- max (0, P_t − delta) # break event b_t = 1
20 log (t, p_t, i_t, s_t, P_t, x, EU [·], m, reason, latencies)

```

---

#### 4.6. Relation to Prior Formalisms

Table 3 attributes the closest formal antecedents. The contribution is the runtime composition and its audit, not novelty of any individual term.

**Table 3.** Closest formal antecedents of the three design terms. The claimed contribution is their composition in an embodied-XR initiative arbiter and the accompanying reachability audit.

| Term                                         | Closest Prior Formalism                                                                                                                                                                            | What It Already Does                                                                                                                                                      | Instantiation Here                                                                                                                                                                    |
|----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Non-zero utility of silence, Equation (3)    | Horvitz's $u(\neg A, G)$ [10]; expected cost of deferral in attention-sensitive alerting [11]; cost of remaining in a gap, $C_G(\tau) = C_G^p \cdot \tau$ , in decision-theoretic turn-taking [47] | The no-action branch already carries non-zero, context- or time-varying cost; in turn-taking it grows linearly with elapsed silence                                       | WAIT cost is parameterized by inferred need confidence and salience in a visible XR companion. The social interpretation is a design rationale, not a validated psychological effect. |
| Recovering presence-risk proxy, Equation (5) | Habituation-and-recovery state dynamics in intervention scheduling [50]; presence as a state inferable from break events [31]                                                                      | A state decremented per intervention, recovering geometrically between them, scaling what the next intervention is worth, and, separately, presence as a measurable state | The recovering state scales intervention cost inside an embodied-XR arbiter; it is an engineering risk proxy rather than an observed presence measure.                                |

Table 3. Cont.

| Term                                    | Closest Prior Formalism                                                                                                                              | What It Already Does                                                                                                                           | Instantiation Here                                                                                                                                                                      |
|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Diegetic reversibility,<br>Equation (2) | Reachability-based side-effect penalties and learned reversibility estimators [44,45]; intervention versus accommodation in narrative mediation [46] | Reversibility as an action-selection penalty is well developed; undoing without the user noticing is the core heuristic of narrative mediation | The penalty encodes whether an action can be undone within the fiction, rather than only whether the prior world state is reachable. Its perceptual interpretation remains unvalidated. |

The WAIT and intervention terms exert opposing pressures, so calibration must consider the utility constants together with the sensor range. A policy carrying only an inaction cost may over-intervene; a policy carrying only intervention cost may default to silence. The deployment audit measures which pressure actually affected decisions.

## 5. Reference Architecture

Six layers carry the loop from sensing to actuation (Figure 2); their interfaces and latency tiers are summarized in Table 4.



**Figure 2.** Agentic Presence’s reference architecture. Six typed layers connect sensing to embodied actuation. L4 evaluates the initiative policy at 10 Hz without a network call; L5 schedules prepared output. The feedback path updates the internal presence-risk stock after an intervention issued at high estimated interruption.

**Table 4.** Architecture layers, interfaces, and latency tiers.

| Layer                     | Function                                                                                                                                                   | Key Interface (Typed Events)                   | Latency Tier            |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|-------------------------|
| L1 Multimodal Sensing     | View-direction dwell, visual-search kinematics, time-since-last-find, speech activity (VAD), task state                                                    | DwellEvent, IdleEvent, SpeechState, SceneState | ms                      |
| L2 Context State          | Rolling window; interruption state; need salience                                                                                                          | ContextFrame                                   | ms                      |
| L3 Intent Inference       | LLM structured output {intent, confidence, target, urgency}; consistency smoothing                                                                         | IntentEstimate                                 | $10^2$ – $10^3$ ms      |
| L4 Initiative Arbiter     | Section 4 policy; mode selection; signature and cadence enforcement                                                                                        | Decision {mode, action, rationale}             | ms                      |
| L5 Anticipatory Scheduler | Drafts the candidate utterance ahead of need; emits a fixed local cue on decision, masking speech-synthesis latency; anytime fallback to preset utterances | PreparedAction                                 | masks $10^0$ – $10^2$ s |
| L6 Embodied Actuation     | Locomotion, pointing, highlighting, speech—each with a fixed signature                                                                                     | —                                              | ms                      |

Latency tiers. The architecture separates fast local motion and arbitration from network-dependent conversational output and slower generative preparation [51–53]. In the evaluated build, L5 has a limited role: an utterance is drafted during intent inference and handed to speech synthesis after a fixed local cue. The measured quantities are arbitration execution, cue onset, and synthesis hand-off (Section 7); audio onset was not instrumented. The evaluated timeout path substitutes a preset utterance. Generative-content prefetching and anytime partial-result refinement [54] are possible extensions of the same scheduling interface, not capabilities validated in this study.

## 6. Implementation

The reference implementation runs in Unity (Universal Render Pipeline), presented stereoscopically on a Meta Quest 3 head-mounted display driven from a workstation over a tethered PC-VR link, with rendering and sensing on the workstation, external model services accessed over the network, and head tracking supplied by the headset. The evaluated environment is a polygonal twilight campground; the Gaussian-splatting capture environments that motivated the tethered configuration, a continuation of earlier volumetric-capture work [55], were not part of the evaluated build. No layer of the architecture depends on the presentation device, but we report the configuration that was actually evaluated. Quest 3 provides no eye tracking, so we term the attention signal view-direction dwell rather than gaze: it is a forward raycast from the tracked head pose, not a gaze vector. Interruption state is accordingly derived from real head kinematics rather than from a virtual camera under mouse control, which is what makes the angular-speed term of the interruption estimate a measurement rather than an input-device artifact.

The participant is seated at a campfire-side table (Figure 3); the companion is embodied as a small non-humanoid creature perched across the table at a 32-degree offset, keeping it within peripheral view for the whole block. Low-fidelity, non-humanoid embodiment

is a deliberate choice that limits potential uncanny-valley effects—avatar realism alters how human-like virtual others are perceived [56]—and holds embodiment constant for the comparison of the complete initiative package with reactive interaction. The companion never manipulates objects.

(a) in-headset view



(b) ACT: torch-beam cue



**Figure 3.** The implemented system: (a) the in-headset polygonal twilight campground, with the creature companion perched across the table inside peripheral view; (b) the companion mid-ACT, directing its torch beam toward an unreported anomaly while the fixed legibility cue plays. In-headset UI text is shown in English translation for presentation; the deployed system ran in Korean.

The arbiter (L4) is plain arithmetic evaluated at 10 Hz; it makes no network call, and the language model sits off the critical path supplying only  $p$ . Speech runs over a streaming realtime model (gpt-realtime-2.1; OpenAI, San Francisco, CA, USA; over a persistent WebSocket, the microphone left continuously open, the model's semantic turn detection used for endpointing, and its automatic response generation disabled so that self-initiated speech is issued only on an L4 decision, with reactive replies requested at the detected end of the participant's turn); intent inference is served by a cloud LLM (google/gemini-3.5-flash-lite; Google LLC, Mountain View, CA, USA; accessed via OpenRouter, OpenRouter, Inc., New York, NY, USA; constrained to a JSON schema, temperature 0) triggered by sensing events plus a 5 s heartbeat, with preset utterances substituted when a round trip exceeds 2.5 s. Sensing thresholds include: view-direction dwell > 2 s on an unreported anomaly, and no-find interval > 25 s.

Context variables. Interruption state is  $i_t = 1$  when the participant is speaking and  $i_t = c_v \times \nu$  otherwise, where  $\nu$  normalizes head angular speed between 30 and 120 degrees per second and is clipped to [0,1]. Need salience combines seconds since the last restoration, the fraction of difficult remaining anomalies, and repeated scanning of one region. Both proxies were computed locally every frame and were calibrated on pilot traces rather than independently validated against interruptibility or need judgments. Active speech also

forces WAIT. Synthesized agent speech is excluded from the detector by a hold-off so that the risk stock does not register the agent's own output as participant interruption.

Co-calibrating the sensing map with the policy. Because Equation (4) multiplies mode salience by  $i_t$ , the sensing range determines which decision regions can be reached. An initial gain  $c_v = 0.3$  made OFFER and the proxy-break condition unreachable. The deployed gain  $c_v = 0.6$  and threshold  $i^* = 0.45$  were selected from pilot traces so that the sensing image intersected every theoretical region and the break condition. This calibration guarantees only geometric reachability; it does not guarantee that participant sessions will occupy those regions. The retained materials do not document the number of pilot users or traces, overlap with the main-study participants, or the full parameter-selection procedure. The calibration process therefore cannot be independently reconstructed, and task-specific tuning cannot be excluded. The released constants and Section 7.4 reproduce the behavior of the final configuration through grid analysis and deployment accounting.

Mode-to-behavior mapping. Each mode is realized by a fixed repertoire, and only the ACT-mode candidate's reversibility enters Equation (2). ACT is a diegetic guiding action taken without asking: the companion leans toward the region holding an unreported anomaly and directs a torch beam at it ( $r = 0.70$ —the beam is withdrawn on request and leaves world state unchanged, but it spends the participant's discovery, which is what makes it the least reversible thing the companion does). OFFER is a low-salience remark that leaves the decision with the participant, e.g., "something over by the tent looks off to me" ( $r = 0.95$ ). ASK is a question requesting permission or information, e.g., "want me to take the left half?" ( $r = 0.90$ ). WAIT is an attentive idle: the companion stays animated and view-tracking but silent, which is what makes Equation (3) meaningful rather than vacuous. The companion never manipulates objects and never resolves an anomaly itself, so nothing in the repertoire falls below  $r = 0.70$ .

Block procedure and anomaly repertoire. Five anomalies are placed per set from a fixed repertoire of five kinds (Figure 4), one of each per set: an object floating 0.07 m above its support, tilted  $10^\circ$  from upright, inverted  $180^\circ$  about its longest horizontal axis, enlarged to  $1.40\times$ , or rotating slowly about the world vertical at  $0.9^\circ/\text{s}$ . Every kind is geometric or kinematic, and this is deliberate: the scene is a twilight campground, and a pilot showed that an appearance-based manipulation—we had used a 40% desaturation—is not reliably visible at that light level, so it would have contributed misses that index the lighting rather than the participant's attention. Scale changes are enlargements only, for the same reason in a different register: at equal log-ratio a shrunken object is markedly harder to detect than an enlarged one, and with one scale anomaly per set that asymmetry would fall entirely on whichever set carried it. Magnitudes are identical across the three sets, so set difficulty varies only through which object carries which kind and where that object sits. Rotationally symmetric objects are excluded from the two rotation kinds, since a rotation no observer can see would be logged as a miss rather than as an undetectable manipulation. Fourteen of the fifteen table objects carry an anomaly in at least one of the three sets—thirteen in exactly one set and one in two—so an object's manipulation in an earlier block is no reliable guide to the current one. Each block opens with the scene in its normal state for 7 s—without this reference a participant cannot know that an object's size or pose is wrong, only that it is unfamiliar—followed by a one-second fade to black under a musical sting, during which the anomalies are applied. Every object on the table, not only the five carrying anomalies, is replaced within its eccentricity zone at that moment; if only the manipulated objects moved, the preview would function as an answer key and the task would become change detection rather than a judgment about how objects should look. The block timer, the arbiter, and the intent service all start after the fade returns, so the preview contributes to neither time-to-find nor the interruption estimate.



**Figure 4.** The anomaly mechanism, which is identical in all three conditions. Each row shows one anomaly type in its altered state (**left**) and after it has been restored (**right**); the two rows illustrate a scale anomaly and an inversion anomaly, cropped from practice-set captures for legibility. Five anomalies are placed per set, three sets are matched for difficulty, and the final restoration in a set triggers a scene-wide payoff. The task supplies the need the arbiter reasons about: an anomaly the participant has dwelt on but not reported is the target of an intervention.

**Report adjudication.** A verbal report is converted into a restoration by a language model that receives the transcript, the participant's current view-direction target, and the object the right-hand controller is pointing at, and returns whether the utterance is a report, which object it refers to, and whether that object is in fact anomalous; controller pointing takes precedence over view direction when both are available, since participants do not know the objects' names and point at what they mean. A report counts only if its description does not contradict the manipulation actually applied—a vague report ("something about that one looks off") passes, while calling an enlarged object shrunk is referred to the experimenter rather than scored, so that naming an object and guessing a description cannot inflate the find count. The adjudicator runs in all three conditions and the scoring path is identical in each: the companion's conversational replies never draw on the anomaly list, and in this reactive role it never volunteers an object the participant has not already named; the self-initiated interventions of Section 4, which do target sensed unreported anomalies, are a separate pathway and are logged as interventions, not as scoring feedback. It does, however, respond to what the participant has just claimed—acknowledging a correct report, saying that a named object looks fine to it when the report is a false alarm, and asking what the participant noticed when the report is ambiguous. Two consequences follow and we state them rather than leave them implicit. First, verbal feedback is by construction absent in C1, which would make the behavioral measures of Section 8.3—finds per block and no-find intervals—incomparable across the embodiment step, because a participant who cannot tell a rejected report from an unheard one repeats herself and spends search time doing it. We therefore carry the same information on a non-verbal channel in all three conditions: a short tick acknowledges that an utterance was received and is being adjudicated, and a two-tone descending cue marks a report of a normal object as rejected; the restoration payoff that follows a correct report is

likewise identical everywhere. What remains particular to C2 and C3 is that the companion also says so, and names the object it is talking about. Second, the feedback is identical in C2 and C3, since it is a response to the participant rather than an act of initiative, so it does not enter the contrast the study is built on. Pointing at an object also prompts a non-committal reply from the companion in the two agent conditions, phrased identically whether or not the object is anomalous, so that pointing serves reference resolution without becoming an oracle that would convert visual search into exhaustive probing. Find time is stamped at the end of the participant's utterance rather than at adjudication, so model latency does not enter the measure, and an experimenter key overrides the adjudicator in either direction with the override recorded. The released source routes adjudication through OpenRouter using `cfg.intentModel`, whose released asset specifies `google/gemini-3.5-flash-lite`. It preserves the Korean prompt, JSON decision schema, confidence thresholds, and timeout setting. This identifies the configuration in the artifact, not the exact hosted version used for every request: the session logs do not preserve an independently verifiable adjudication-model version. Across the 39 completed study blocks, the event audit found one `wizard_adjudicate` record, marked `accepted = 0` with an `already-restored-or-missing` note, and no `accepted wizard-adjudication` records. These are counts of logged manual events, not a verified rate of reversals of model decisions (Supplementary Software S2).

**Participant-facing interface.** A world-anchored board beyond the far edge of the table carries the current instruction and the number of anomalies restored—remaining time is deliberately not shown, since visible time pressure would feed directly into search kinematics and intervention acceptance—and it never displays the condition, the set identifier, or any arbiter quantity. A single control carries everything the participant does: the right-hand trigger points at objects during a block, selects answers on the questionnaire, and advances the session at the three transitions between those phases—entering the practice block, leaving it, and after each questionnaire. The three uses cannot collide because the phases they belong to are disjoint, and the advance prompt appears only while that use is live. A cursor marks where the ray meets the questionnaire panel, since a ray aimed head-on projects to a point and gives the participant no cue as to where it lands. Restorations use the shared adjudication pathway, with experimenter keys available for manual adjudication; the tutorial ends automatically when both practice anomalies are found. Questionnaires are answered inside the headset on a world-space panel rather than by removing the display, so that instruments asking about presence are not completed after presence has been broken by doffing; the panel occludes the scene while it is up, so that residual anomalies cannot be searched for during a questionnaire, and the seated rig is re-centered with its eye height recalibrated at the start of every block. The soundscape—a low wind bed and two cricket beds placed to the sides—is spatialized and identical in all conditions.

**Configuration and logging.** The build used Unity 6000.0.63f1 (Unity Technologies, San Francisco, CA, USA) with URP and OpenXR (Khronos Group, Beaverton, OR, USA), an NVIDIA RTX 3080 workstation (NVIDIA Corporation, Santa Clara, CA, USA), and a tethered Meta Quest 3 (Meta Platforms, Inc., Menlo Park, CA, USA) at 72 Hz. The release includes configuration assets, source prompts, constants, and task definitions; session logs record the configuration fields available at runtime. These records are not immutable versions of the external hosted models. Each arbitration tick recorded all policy inputs, four utility values, overrides, selected mode, latencies, and subsequent participant response. The deposited package contains the frozen configuration represented in the logs, task definitions, pseudonymized interaction records, questionnaires, and analysis scripts (Data Availability Statement).

## 7. System Evaluation and Deployment Audit

The system evaluation uses decision logs from the pilot to assess implementation feasibility and realized policy operation. It distinguishes the utility calculation from hard runtime gates, numerical reachability from observed occupancy, and issued interventions from subsequently logged outcomes. Frozen-input replay supports local decision diagnostics; it is not a controlled test of alternative user experiences.

### 7.1. Latency Decomposition

We report per-stage latency over all logged events on the self-initiated path—sensing trigger → context frame → LLM inference → arbitration decision → actuation onset—and report the reactive-reply path as a single latency class. Table 5 reports the self-initiated path over all 75 issued interventions. Context variables are computed in the same frame as their sensing inputs; intent inference is the only network hop (median of 1014 ms over 2109 calls) and sits off the critical path: once the arbiter selects a mode, the fixed legibility cue begins within a millisecond, and the pre-drafted utterance is handed to the streaming synthesizer within a millisecond of the cue. The audio onset of streamed synthesis is governed by the realtime speech service and is not separately instrumented; the 160 ms cue masks its leading edge. The reactive-reply path closes entirely inside the streaming speech model’s own turn loop (subsecond class) and is not decomposed per stage.

**Table 5.** Per-stage latency of the self-initiated path (75 interventions, 13 sessions). The reactive-reply path closes inside the streaming speech model and is reported as a latency class, not per stage.

| Stage                                 | Latency                            | Notes                                       |
|---------------------------------------|------------------------------------|---------------------------------------------|
| Sensing trigger → context frame       | same frame (13.9 ms as mean)       | $i_t, s_t$ computed every frame             |
| Context frame → intent estimate $p_t$ | median of 1014 ms; p95 1416 ms     | 2109 calls; timeout: 0.33%; 5 s heartbeat   |
| Arbitration tick (Algorithm 1)        | median of 4 $\mu$ s; p95 5 $\mu$ s | 10 Hz nominal; 111 ms realized period       |
| Decision → legibility cue onset       | $0.1 \pm 0.2$ ms (max 1.0)         | fixed 160 ms cue                            |
| Cue onset → synthesis hand-off        | $0.3 \pm 0.4$ ms                   | audio onset masked by cue; not instrumented |
| Age of $p_t$ at intervention          | median of 1.3 s                    | freshness decay of Section 4.5              |
| Preset fallback in issued speech      | 0 of 75                            | —                                           |

### 7.2. Runtime Decisions and Annotation Scope

Across 19,929 C3 ticks, the system issued 75 interventions (2.05 per minute): ACT with 44, ASK with 31, and OFFER with zero. Forty-two interventions (56.0%) were followed by inspection or report of the referenced target within 15 s. The cadence gate applied on 18,957 ticks (95.1%) and the speech gate on 694 (3.5%); the per-block cap never fired. There were 255 unoverridden ticks (1.3%) and 23 no-candidate records (0.1%). This accounting identifies the hard gates as the dominant restrictions on issued initiative in the deployed regime, irrespective of the utility mode that would otherwise have been selected.

A post hoc annotation exercise is retained only as an internal engineering diagnostic. Policy labels were masked and moments shuffled, but the primary annotator was the system author; a second annotator coded 45 moments, with Cohen’s kappa = 0.20 and a raw agreement of 58%. Nineteen disagreements were adjudicated in favor of the primary code. These annotations do not provide independent ground truth for intervention quality,

and their precision estimates are excluded from the main policy comparison. Recall cannot be estimated because all need episodes were not independently segmented. The original annotation records remain available for transparency.

### 7.3. Counterfactual Policy Comparison on Frozen Inputs

Counterfactual replay compares the deployed schedule with (i) a confidence-only rule under the same hard overrides and (ii) schedules sampled uniformly from each participant's feasible candidate ticks subject to the same intervention count and minimum 30 s gap. Replay evaluates selection on frozen traces; it does not model how alternative interventions would change later participant behavior.

The deployed and confidence-only rules selected 75 and 76 intervention moments, respectively (Table 6). These totals count all non-WAIT modes, not OFFER actions; the deployed total comprises 44 ACT, 31 ASK, and zero OFFER actions. The random null preserved each participant's deployed intervention count and enforced a minimum 30 s gap. The sensor-context comparison uses no subjective annotation. Intervention quality is not inferred from the author-led coding exercise.

**Table 6.** Annotation-free policy comparison on frozen inputs. Intervention moments count all non-WAIT modes, not OFFER actions. Tail probabilities are medians of participant-level directional Monte Carlo probabilities from 100,000 schedules sampled under the minimum 30 s cadence constraint; Fisher-combined  $p$ -values are reported in Section 7.3. Author-led precision estimates are excluded because the annotations are not independent ground truth.

| Policy                     | Intervention Moments | Mean $i_t/s_t$ | Median Tail Probability vs. Random Null    |
|----------------------------|----------------------|----------------|--------------------------------------------|
| Deployed (Equation (6))    | 75                   | 0.018/0.585    | i: 0.63; s: 0.83 (median tail probability) |
| Confidence-only            | 76                   | 0.031/0.580    | —                                          |
| Cadence-constrained random | 75 (matched)         | null reference | —                                          |

For each participant, 100,000 schedules were sampled exactly from the feasible cadence-constrained set. Every draw was asserted to satisfy the 30 s gap. Directional tail probabilities tested whether deployed moments had lower mean interruption, higher mean need salience, or, post hoc, higher mean confidence than feasible random schedules. Participant  $p$ -values used the Monte Carlo +1 correction and were combined with Fisher's method; a sensitivity analysis excluded ticks with no-candidate action.

The released legacy random-schedule annotations do not enforce the 30 s cadence constraint used by the corrected null. They are therefore not used to estimate precision for the cadence-constrained random comparator. The log-based context analysis did not distinguish the deployed schedule from feasible random schedules on the two specified statistics: median participant tail probability was 0.63 for lower interruption (Fisher chi-square (26) = 14.54,  $p = 0.965$ ) and 0.83 for higher salience (chi-square (26) = 33.87,  $p = 0.138$ ). A post hoc comparison found higher intervention confidence than the constrained null (median tail probability = 0.28; chi-square (26) = 54.48,  $p < 0.001$ ). Excluding no-candidate ticks produced the same interpretation ( $p = 0.965$ , 0.154, and 0.001, respectively). Together with the similar intervention counts under the confidence-only rule and the low observed interruption range, these results are consistent with selection driven mainly by confidence

behind cadence and speech constraints. They do not establish a causal timing advantage or equivalence of user experience across policies.

Counterfactual replay assumes that participant behavior will remain unchanged after an alternative intervention. The assumption is most defensible for a single decision and deteriorates over longer horizons; the replay therefore evaluates moment-level policy selection, not a simulated alternative user experience.

#### 7.4. Utility Sensitivity and Sensor Reachability

We evaluate utility-region occupancy on a deterministic grid rather than simulate new participant sessions. Each fixed-stock setting uses 241 confidence values in  $[0,1]$ , 61 interruption values in  $[0, c_v]$ , and 21 salience values in  $[0,1]$ , including endpoints. Figure 5a sweeps ACT reversibility over  $[0,1]$  in steps of 0.02; Figure 5b sweeps  $\kappa_Q/\kappa_O$  over  $[0,1.2]$  in steps of 0.05, holding the other Table 2 constants fixed. The declared ACT–OFFER–ASK–WAIT order breaks exact ties. Grid shares quantify model geometry, not probabilities of user states. This analysis covers the low-reversibility range absent from the action repertoire, but it is not an empirical validation of that range or a comprehensive sweep of all constants. The calculations and figure-generation code are included in Supplementary Software S2.

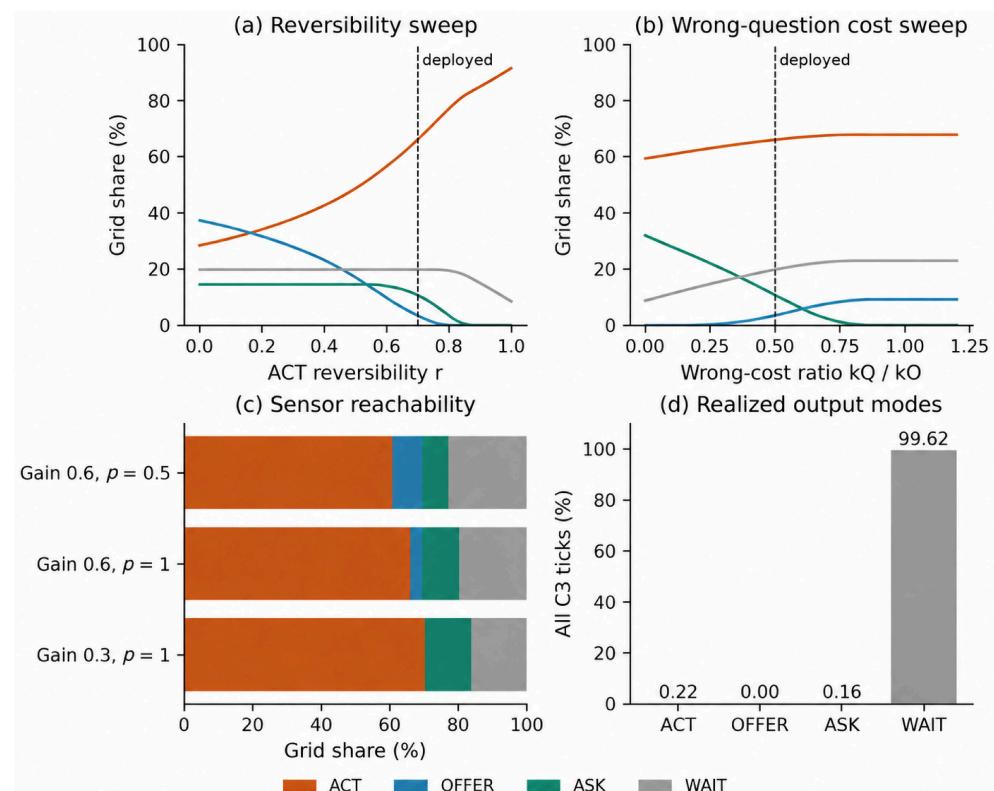


**Figure 5.** Utility sensitivity and deployment coverage. (a) ACT reversibility and (b) relative wrong-question cost change mode-region occupancy at  $p = 1.0$  and  $c_v = 0.6$ . (c) Reachable grid shares compare sensing gains and two fixed-stock values. (d) Issued modes use all 19,929 C3 ticks as the denominator; WAIT includes runtime overrides. Grid shares in (a–c) and observed tick shares in (d) are different quantities and are not estimates of the same distribution.

Figure 5c separates sensor reachability from the observed tick-level mode frequencies in Figure 5d. Under the initial sensing gain  $c_v = 0.3$ , the reachable-space split at  $p = 1.0$  was ACT with 70.2%, ASK with 13.6%, WAIT with 16.2%, and OFFER with 0.0%. With  $c_{nu} = 0.6$ , the theoretical reachable split was ACT with 66.1%, ASK with 10.7%, OFFER with 3.4%, and WAIT with 19.8% at  $p = 1.0$ . Participant traces occupied a much narrower

region: outside active speech, interruption had a median of zero and a 95th percentile of 0.141; with speech-saturated ticks included, interruption exceeded 1/3 on 5.3% of ticks; and no OFFER was issued. Thus calibration made OFFER geometrically reachable, but the deployed sessions still realized a three-mode policy.

### 7.5. Empirical Influence of the Presence-Risk Stock

The stock audit must distinguish the state used to make a decision from the state written after acting. The released implementation computes the utilities and `mode_if_P1` before applying the break decrement, then logs the updated  $p$ . Only one proxy break occurred (Figure 6); logged  $p$  was below one on 294 ticks (1.5%), including 19 with non-zero interruption. For replay, the decision-time stock is recovered by adding  $\delta = 0.25$  on the break tick, where  $p$  changed from 1.0 to 0.75. With this correction, fixing  $p$  at one produces no utility-argmax changes in the rounded-input replay (0/19,929). The earlier one-change result arose from using the postaction stock at that break tick. Only 18 ticks combine an already depleted decision-time stock with non-zero interruption; on all 18, the argmax of the logged utilities agrees with the online `mode_if_P1` field, with a minimum top-two utility gap of 0.023. Supplementary Software S2 reports rounding diagnostics and the comparison procedure. The deployment therefore provides no observed decision-level benefit or criterion-validity evidence for the stock; the audit identifies the conditions under which this extension still requires testing.

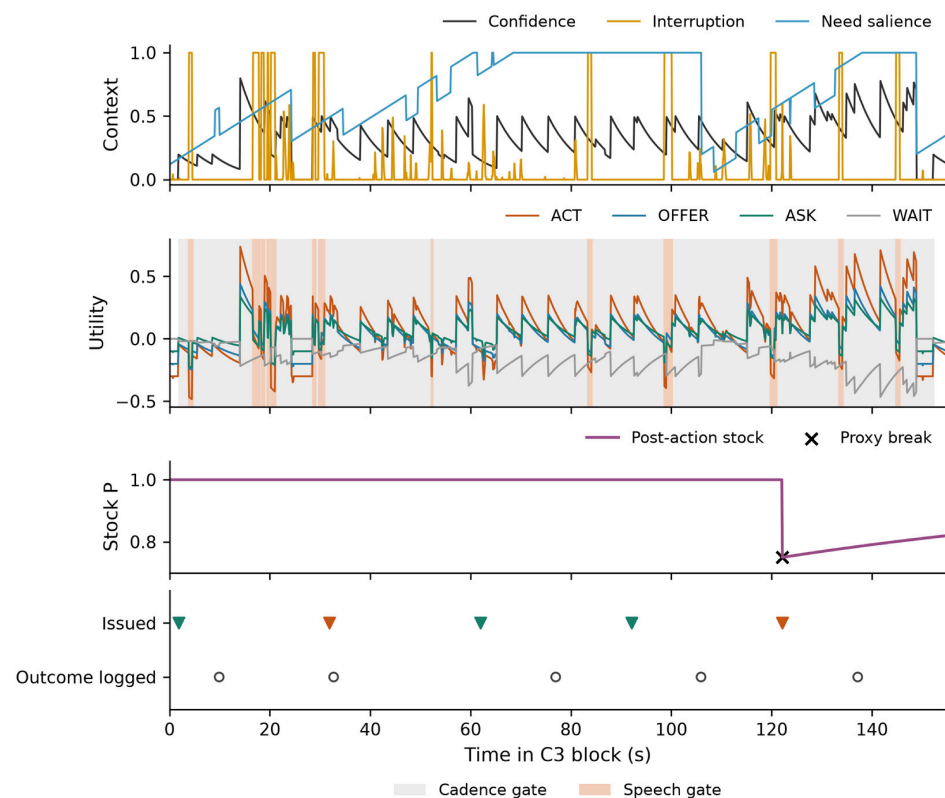


**Figure 6.** C3 block from participant 8, selected to show the only proxy-break update. Context inputs, utilities with gate shading, and postaction stock are followed by a separate event panel. Colored triangles mark issued interventions from non-WAIT arbiter ticks (colors follow the utility panel: orange, ACT; green, ASK); hollow circles mark later outcome-record emission, not intervention onset. The stock decrement follows the decision at the break tick. The proxy break is defined by the system and is not an independently observed psychological break in presence.

### 7.6. Runtime Resource Characterization

Three measurements bound the runtime cost (Table 7): arbiter execution time, inference latency and fallback frequency, and rendering/frame performance. The arbiter itself is fixed-cost arithmetic; the reported measurements establish implementation feasibility but do not establish the validity of the presence-risk proxy.

**Table 7.** Runtime resource measurements. The nominal 72 Hz frame budget is 1000/72 ms; the 4  $\mu$ s median arbiter time occupies approximately 0.03% of that budget. Measured frame time is reported separately. Cue precedence refers to synthesis hand-off, not measured audio onset.

| Quantity                                       | Value                                                               |
|------------------------------------------------|---------------------------------------------------------------------|
| Arbiter execution per tick                     | Median: 4 $\mu$ s; p95 5 $\mu$ s; max 3300 $\mu$ s (first-tick JIT) |
| Median arbiter time/nominal 72 Hz frame budget | $\approx$ 0.03%                                                     |
| Realized arbitration period                    | 111 ms as median (nominal: 100 ms)                                  |
| Measured frame time                            | 13.9 ms as mean; p95 15.8 ms (72 Hz held)                           |
| Intent-inference round trip                    | Median: 1014 ms; p95 1416 ms (2109 calls)                           |
| Inference timeouts (>2.5 s)                    | 0.33% (7 of 2109)                                                   |
| Preset fallback in issued speech               | 0 of 75                                                             |
| Cue preceding synthesis hand-off               | 75 of 75 (fixed 160 ms cue)                                         |

These are reported separately from the latency decomposition of Section 7.1 because they answer a different question. Section 7.1 asks what the user waits for; this subsection asks what the architecture costs, which is the quantity a reader deciding whether to adopt it needs.

## 8. Materials and Methods for the Pilot Package Comparison

The within-subject pilot manipulates whether the otherwise identical embodied companion initiates behavior (Table 8). C1–C2 contrasts companion availability, including embodiment and reactive interaction; the prospectively specified primary C2–C3 contrast estimates the marginal effect of adding the complete initiative package.

**Table 8.** Pilot-study conditions. The anomaly-restoration mechanism is identical in all three conditions; embodiment, seating, idle behavior, voice, and response-latency masking are identical in C2 and C3. C2 disables self-initiated behavior, whereas C3 enables the complete initiative package through L4.

| Code | Condition | System Configuration                                                                                                                                                                                        |
|------|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C1   | No agent  | L6 disabled entirely; anomalies are reported verbally and restore identically in all conditions.                                                                                                            |
| C2   | Reactive  | Full stack; L4 forced to WAIT on every self-initiated path. The companion remains embodied and reactive but never initiates. C2–C3 therefore manipulates the complete initiative package, not timing alone. |
| C3   | Agentic   | Full stack; L4 active per Section 4.                                                                                                                                                                        |

The C2–C3 contrast evaluates the complete initiative package: enabling L4 can change intervention frequency, content, modes, and timing together. It therefore does not identify the independent effect of the proposed timing policy. That effect requires a proactive comparator matched on intervention number, content, modes, and embodiment. Section 7 provides a deployment and offline decision audit, not an alternative human-subject control.

### 8.1. Study Questions and Directional Hypotheses

The following directional hypotheses and primary contrast were fixed in writing before data collection but were not publicly preregistered. They are retained as the prospectively specified study questions; given the achieved sample and measurement limitations, their results are interpreted as preliminary estimates rather than confirmatory tests of the architecture's individual mechanisms.

**H1 (primary, directional).** *Social co-presence is higher in C3 than in C2. This tests the initiative package; it does not estimate the silence weight  $\lambda$  or establish that the model's WAIT cost is psychologically valid.*

**H2 (non-inferiority).** *The mean C3–C2 increase in intrusiveness is below a prospectively specified 1-point margin on the 7-point scale. The margin was a study-specific threshold chosen by the authors and corresponds to one response-scale step. The spacing of individual response categories alone does not establish that a 1-point increase in a group mean is acceptable; the non-inferiority interpretation is therefore conditional on this threshold. This tests whether initiative produces a large average increase relative to the reactive floor; it is not an equivalence test and does not isolate the interruption or stock terms.*

**H3 (exploratory accompaniedness gradient).** *Accompaniedness increases across C1, C2, and C3. Pairwise contrasts distinguish the embodiment step from the additional initiative step.*

An exploratory probe asks whether the count of high-interruption interventions is associated with C3's co-presence, BIP, and intrusiveness. Because  $P_t$  is an internal proxy and not an observed presence measure, any association would be criterion-related evidence rather than direct validation.

### 8.2. Study Design, Environment, and Participants

The within-subject design scheduled all six condition orders across three matched seated visual-search sets. Thirteen adults completed all blocks, below the planned target of 16–18; five additional starts ended before the first block because of technical setup and contributed no outcome data. The original rationale for the target and an a priori power or precision calculation could not be reconstructed from the available documentation. The revision-stage sensitivity analysis in Section 8.4 characterizes the achieved sample; it does not supply a retrospective planning justification. The realized counts for C1–C2–C3, C1–C3–C2, C2–C1–C3, C2–C3–C1, C3–C1–C2, and C3–C2–C1 were 2, 1, 3, 2, 3, and 2, respectively. Tables S1 and S2 in the Supplementary Materials Document give the counts and each completed participant's condition, period, and task-set mapping, verified against block-entry logs. Recruitment used professional networks in design and technology. Participants were unpaid adult volunteers. None were employees, contractors, direct reports, or collaborators of RECON Labs or the author. The minimal-data protocol recorded confirmation of adult status but no demographic profile, limiting assessment of sample diversity. Each block ended after five restorations or 240 s. The task used no virtual locomotion; discomfort was checked after every block and no participant withdrew. The study was conducted as internal research at RECON Labs Inc. No independent ethics committee or Institutional Review Board reviewed this study. Before data collection, the company assessed the protocol as low risk and determined internally that external ethical review was not required; it does not maintain an Institutional Review Board. All participants gave written informed consent, including separate consent for audio capture, and could stop without consequence. This was a company-internal determination, not

approval or exemption by an independent ethics committee; the process is described in the Institutional Review Board Statement.

### 8.3. Measures

The primary outcome was the six-item co-presence subscale of the Networked Minds Social Presence Measure [57], administered after C2 and C3 on a 7-point response format. This social construct is distinct from spatial/place presence and from the internal  $P_t$  proxy. Intrusiveness, the outcome of the prospectively specified non-inferiority test (H2), used the same 7-point format. Secondary exploratory outcomes included the IPQ general item, spatial presence and involvement [23] on a unified 7-point format, accompaniedness, retrospective BIP count, partner-model stability, perceived initiative, and ratings of the companion's timing and manner. Behavioral measures included participant utterances, intervention acceptance within 15 s, companion dwell, social-referencing glances, no-find intervals, false reports, and finds. The complete Korean items and scoring rules are included in the repository. ChatGPT (OpenAI; GPT-5.6 sol) was used to review the Korean translation. The retained materials do not document independent back-translation or formal cross-cultural validation. The unified IPQ response format is a modification of the original instrument; neither the translation review nor within-sample reliability establishes equivalence to the original version. Questionnaire-based presence measurement itself has documented reliability and sensitivity limits [58,59], a caveat that applies to all self-report outcomes reported here.

### 8.4. Procedure and Statistical Analysis

The hypotheses, primary contrast, analysis plan, and policy constants were prospectively documented but not publicly preregistered. Directional tests are therefore reported together with two-sided confidence intervals and robustness analyses rather than being treated as definitive confirmation.

After consent and a separate practice set, participants completed three blocks of up to 4 min, each followed by an in-headset questionnaire. Conditions were only labeled Version A/B/C. H1 used the prospectively specified one-sided paired t-test after checking paired-difference normality; we also report  $d_z$ , the corresponding two-sided 95% t interval, the two-sided  $p$ -value, an exact sign-flip sensitivity, and a signed-rank sensitivity. H2 was a one-sided upper-margin non-inferiority test of C3-minus-C2 intrusiveness against  $\delta = 1.0$  point; sensitivity analyses used margins of 0.50 and 0.75. H3 and other outcomes were exploratory, using Friedman tests and Holm-adjusted pairwise contrasts where applicable. We inspected period, C2/C3 temporal order, and task-set sensitivity because technical restarts left sequence cells uneven. With  $n = 13$  these checks are descriptive and low-powered. Tests used  $\alpha = 0.05$ . A revision-stage sensitivity analysis using the noncentral t distribution estimates the standardized paired difference detectable with 80% power. At  $n = 13$ , the threshold is  $d_z = 0.73$  for a one-sided test and 0.85 for a two-sided test; at  $n = 16$ –18 it is 0.65–0.61 and 0.75–0.70, respectively (Table S4 in the Supplementary Materials Document). These are design-sensitivity calculations, not observed power, an a priori justification, or evidence that null findings establish no effect. With the observed paired-difference SD of 1.25, the co-presence 95% interval has a half-width of 0.76 points. Analyses were performed in Python 3.12 with SciPy 1.18.0 and NumPy 2.5.2. The corrected cadence-constrained sampler and machine-readable results are provided as Supplementary Software S1; remaining frozen materials and environment details are deposited [60]. Revision-stage diagnostics and replacement Figures 1, 5 and 6 were generated with Python 3.12.14, NumPy 2.3.5, SciPy 1.18.1, and Matplotlib 3.11.2 (Supplementary Software S2).

## 9. Results

Thirteen participants completed all three blocks. Condition-specific internal consistency was co-presence  $\alpha = 0.88$  in C2 and  $0.64$  in C3, intrusiveness  $\alpha = 0.79$  and  $0.94$ , and partner-model  $\alpha = 0.83$  and  $0.88$ , respectively. Item-level diagnostics (Table S3 in the Supplementary Materials Document) show C3 corrected item–total correlations of  $0.24$ – $0.55$  and alpha-if-deleted values of  $0.51$ – $0.64$ ; deleting an item does not materially resolve the low consistency. On C3 items, 4–6 of 13 participants chose the scored maximum, consistent with restricted upper-end variation. Five item means increased in C3, whereas the reverse-scored item about the companion forgetting the participant (cop6) decreased from  $5.54$  to  $5.38$ . These diagnostics indicate heterogeneous item behavior rather than a uniform scale shift. The original six-item score is retained; no post hoc item deletion is used to strengthen H1. Unless noted, descriptives are mean (SD) on seven-point scales.

**Manipulation check.** Perceived initiative was higher in C3 ( $5.54$  ( $1.66$ )) than C2 ( $4.23$  ( $1.92$ )), where  $t(12) = 2.34$ , one-sided  $p = 0.019$ , and  $d_z = 0.65$ . Change in perceived initiative was strongly associated with change in co-presence (Spearman  $\rho = 0.87$ ,  $p < 0.001$ ), consistent with possible construct overlap or common-method variance. This association does not independently validate the co-presence effect.

**H1 (co-presence, primary).** C3 was higher than C2 ( $5.94$  ( $0.78$ ) vs.  $5.29$  ( $1.21$ ); Figure 7), where mean paired difference =  $0.64$ ,  $SD = 1.25$ ,  $t(12) = 1.85$ , prospectively specified one-sided  $p = 0.045$ , and  $d_z = 0.51$ . The two-sided  $p$ -value was  $0.089$  and the  $t$ -based two-sided 95% confidence interval was  $[-0.11, 1.40]$ . The exact one-sided sign-flip  $p$ -value was  $0.044$ , whereas the signed-rank sensitivity was approximately  $0.055$ . Only five of 13 leave-one-participant-out analyses retained  $p < 0.05$ . With  $n = 13$ , C3's co-presence  $\alpha = 0.64$ , and a two-sided interval that includes zero, this is a preliminary, statistically fragile estimate of the initiative-package contrast rather than confirmatory evidence of a co-presence effect.

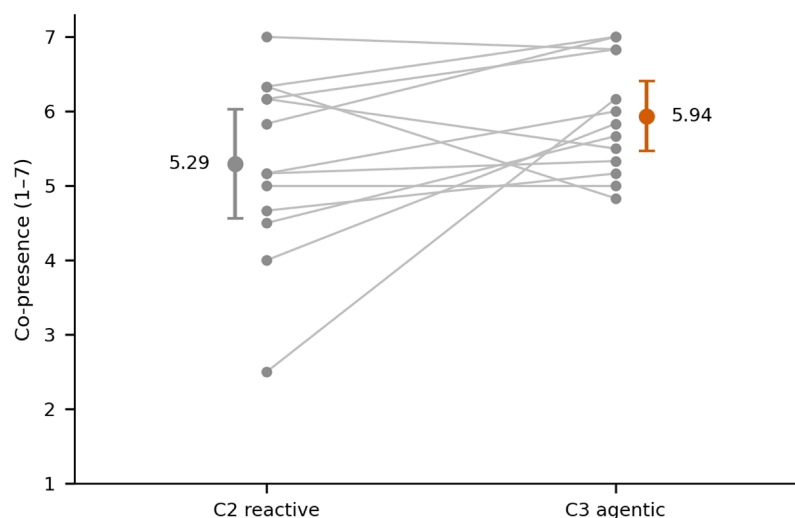


**Figure 7.** Co-presence by condition (C2 reactive vs. C3 agentic). Gray lines show individual participants; the gray (C2) and orange (C3) points show condition means with 95% CIs.

**H2 (intrusiveness non-inferiority).** C3 and C2 means were  $1.64$  ( $1.24$ ) and  $1.54$  ( $0.42$ ), respectively, where mean paired difference =  $0.10$  and  $SD = 1.21$ . The one-sided upper 95% confidence bound was  $0.70$ , below the one-point margin ( $p = 0.010$ ). Sensitivity  $p$ -values were  $0.038$  for the  $0.75$ -point margin and  $0.129$  for the  $0.50$ -point margin. Table S3 in the Supplementary Materials Document displays every paired difference, the group-mean upper confidence bound, and all three margins. The conclusion meets the one- and  $0.75$ -point criteria, but not the stricter  $0.50$ -point criterion. This margin-dependent result does not establish equivalence or the absence of smaller increases.

Order sensitivity. Because the realized sequences were uneven, we examined whether the C2-to-C3 co-presence difference varied by temporal order. Mean differences were 0.86 when C2 preceded C3 and 0.45 when C3 preceded C2 (Welch  $p = 0.600$ ). A paired-difference regression adjusted for period and task set estimated a C3–C2 difference of 0.65 (two-sided 95% CI  $[-0.27, 1.58]$ ,  $p = 0.142$ ). With 13 participants and five fitted coefficients, this is a low-powered sensitivity analysis and does not rule out order, period, task-set, or carryover effects.

*H3 (accompaniedness).* The omnibus condition effect was significant (Friedman chi-square (2) = 10.09,  $p = 0.006$ , Kendall  $W = 0.39$ ; Figure 8). Holm-corrected contrasts supported C2 > C1 ( $p = 0.004$ ) and C3 > C1 ( $p = 0.021$ ), but not the initiative-specific C3 > C2 step ( $p = 0.18$ ). This pattern is consistent with greater accompaniedness when a companion is available; it does not establish an additional accompaniedness effect of initiative.

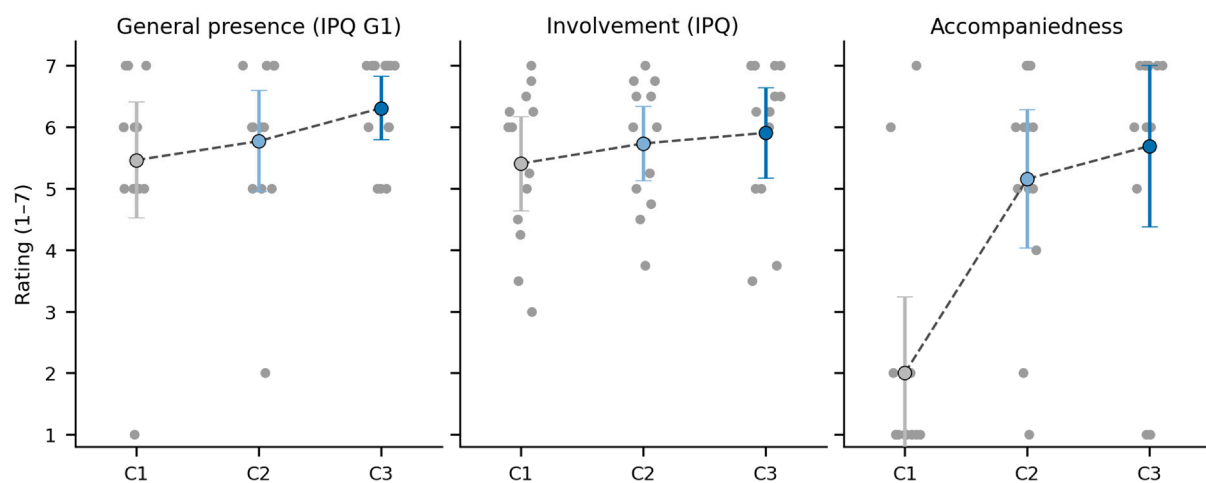


**Figure 8.** General presence, involvement, and accompaniedness across conditions. Error bars show 95% CIs.

*Exploratory measures.* Retrospective BIP counts differed across the three conditions (C1 at 1.69, C2 at 0.85, C3 at 0.62; Friedman  $p = 0.011$ ), but the initiative-specific C3–C2 signed-rank contrast was not significant (one-sided  $p = 0.227$ ). Spatial presence showed a non-significant trend (4.82, 5.37, 5.82;  $p = 0.052$ ); the IPQ general item and involvement did not differ reliably. Partner-model stability was numerically lower in C3 than C2 (3.87 vs. 4.36, n.s.). Ratings of the appropriateness of the companion’s timing and manner were high and statistically indistinguishable between conditions (timing of 5.2 in C2 vs. 5.4 in C3; manner of 5.1 vs. 5.5; n.s.). Accordingly, neither the BIP result nor the spatial-presence trend validates  $P_t$ .

*Behavioral traces.* Finds per block did not differ ( $p = 0.33$ ). False-report rate was lower in C3 than C2 (0.88 vs. 1.88 per minute; two-sided Wilcoxon  $p = 0.009$ ), and the maximum no-find interval was shorter (70.9 vs. 93.7 s; two-sided  $p = 0.048$ ). Participants spent little time looking at the companion (mean of 6.7 s per block in C2, 4.9 s in C3), social-referencing glances around restoration payoffs were rare (nine of 56 payoffs in C2, seven of 59 in C3), and utterance counts did not differ reliably across conditions. These exploratory traces are compatible with useful initiative, but they do not identify which policy term caused the difference. Forty-two of 75 interventions (56.0%) were accepted within 15 s, with median response latency of 2.0 s.

*Presence-risk-stock probe.* The planned association was not estimable because only one proxy-break event occurred across all participants. No participant-level inference about the stock mechanism is therefore warranted.

## 10. Discussion

The main contribution is a reference architecture whose runtime decisions can be inspected against its formal policy. Typed interfaces separate asynchronous inference from local arbitration and embodied actuation, while logs expose utilities, hard gates, and issued modes. The median arbiter execution time of four microseconds establishes feasibility for this implementation. The reachability and deployment audit makes an additional distinction that latency alone cannot provide: an implemented mechanism may be mathematically reachable yet contribute little or nothing to the decisions made in the evaluated task.

This distinction leads to a practical configuration recommendation. For deployments limited to the observed quiet, seated regime, a simpler candidate profile would retain the cadence and speech constraints, confidence-based selection, and the exercised ACT, ASK, and WAIT behaviors, with the stock fixed at one. The four-mode model and adaptive stock can remain explicit extensions for tasks that reach their operating conditions. This recommendation follows from the diagnostic evidence; it does not retrospectively change the deployed policy or claim that a simplified profile has been validated with users. A useful stock test requires interventions at non-zero interruption, depletion that persists to later eligible decisions, and inputs near a mode boundary. Increasing interruption alone is insufficient when the speech or cadence gate still forces WAIT. Event-level interruption or spatial-presence judgments would additionally be needed to establish criterion validity.

The user study contributes a complementary, preliminary estimate of adding the complete initiative package. Its co-presence interval spans both a small negative difference and a substantial positive difference, and the intrusiveness conclusion depends on the chosen margin. Low C3 scale consistency and the association between perceived initiative and co-presence constrain interpretation. No matched proactive control was collected, so these outcomes cannot establish that the proposed timing policy is better than another proactive schedule. A future matched comparison would hold intervention number, content, modalities, and embodiment constant while varying timing. The same separation of implementation from evidence applies to L5: generative-content prefetching [53,61–65] is an architectural extension, whereas this study evaluated only prepared utterances.

Privacy and reproducibility also constrain deployment. An XR agent can expose sensitive visual context, speech, and behavioral traces when data are sent to external services. Selective image encryption, including the semantic-region approach of Liu et al. [66], addresses protection of sensitive image regions and is relevant to future visual-stream extensions. It does not validate initiative arbitration, and no such encryption algorithm was evaluated here. Deployments should define what data may leave the device, minimize retained content, and specify access and retention controls. For reproducibility, the released local utilities, configuration, and archived-input analyses can be rerun; live proprietary-service outputs cannot be reproduced at model-version level from the retained records.

## 11. Limitations

The evidence covers one short seated visual-search task, one non-humanoid companion, Korean interaction, tethered PC-VR, and 13 professional-network volunteers with no demographic profile. Uneven condition orders and the small sample limit sensitivity to period, learning, and carryover effects. The C3–C2 manipulation changes the initiative package rather than timing alone. The social co-presence outcome is distinct from spatial/place presence and the internal risk stock, and neither the retrospective BIP counts nor the questionnaire results provide criterion validation of that stock.

The stock is an intervention-history control proxy, with simplified damage and recovery constants [32]. The interruption, salience, reversibility, and confidence inputs lack independent criterion validation. Incomplete calibration records limit assessment of task-specific tuning and pilot-participant overlap; the questionnaire translation and response-format adaptation lack documented cross-cultural validation. OFFER was reachable in the model but not issued, and low reversibility was examined numerically rather than with participants. The post hoc author-led annotation is unsuitable as independent evidence of intervention quality. The one-point intrusiveness margin is a study-specific judgment, not an externally validated minimum important difference. Audio onset was not instrumented, hosted-model version records are incomplete, and rounded logs limit bit-exact replay at decision boundaries. Short exposure also leaves long-term partner-model stability untested.

## 12. Conclusions

Agentic Presence provides an auditable six-layer architecture for proactive embodied XR agents, demonstrated in a realtime Unity implementation. Its central methodological contribution is to connect formal arbitration regions to sensor reachability and then to actual decisions under runtime constraints. The deployment audit identified cadence and speech gating as dominant and separated exercised ACT, ASK, and WAIT behaviors from an unused OFFER mode and an inactive adaptive-stock mechanism. This evidence supports concrete decisions about which parts of the architecture to retain or simplify for similar deployments.

The small pilot complements the system evidence with preliminary estimates of the complete initiative package, not proof of policy-specific timing benefits. The architecture, decision logs, and reproducible analyses provide a basis for implementations whose claimed mechanisms can be checked against what their sensors and runtime actually exercise.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/electronics15204736/s1>, Software S1: Corrected cadence-constrained null analysis (corrected\_cadence\_null.py and corrected\_cadence\_null\_results.json); Software S2: Revision-stage analysis and figure generation (revision\_analysis.py, machine-readable results, and source data and generation code for Figures 1, 5 and 6); Table S1: Realized condition-order counts; Table S2: Condition, period, and task-set mapping for each completed participant; Table S3: Item-level diagnostics of the co-presence scale; Table S4: Design-sensitivity analysis; Figure S1: Paired intrusiveness differences with the group-mean upper confidence bound and non-inferiority margins.

**Funding:** This work was supported by the Industrial Technology Innovation Program (RS-2025-02317327, Development of an AI Design Generation Pipeline Based on the Design Workflow), funded by the Ministry of Trade, Industry & Energy (MOTIE), Republic of Korea, and by the Ministry of Culture, Sports and Tourism and the Korea Creative Content Agency through the Culture Technology Research and Development Program (RS-2024-00393915).

**Institutional Review Board Statement:** No independent ethics committee or Institutional Review Board reviewed this study. Before data collection, RECON Labs Inc., which does not maintain an Institutional Review Board, assessed the short, seated XR protocol as low risk and determined internally that external ethical review was not required. This was a company-level determination, not approval or exemption by an independent ethics committee. Participants were adult volunteers; none were employees, contractors, direct reports, or collaborators of the company or the author. All participants provided written informed consent and could stop without consequence. Separate written consent was obtained for audio capture, and audio and transcribed material were stored under participant codes. The study involved no medical intervention, and no identifiable participant images were included in the manuscript.

**Informed Consent Statement:** Written informed consent for participation was obtained from all participants. Separate written consent was obtained for in-session audio capture. No identifiable participant image is included in the manuscript.

**Data Availability Statement:** The frozen configuration represented in the session logs, task definitions, analysis scripts, the Unity arbiter implementation, questionnaire items, and pseudonymized questionnaire responses and interaction logs are available on Zenodo [60]. The repository’s legacy cadence-null script is superseded for Section 7.3 by Supplementary Software S1. Tables S1–S4 and Figure S1 are presented in the accompanying Supplementary Materials Document. Supplementary Software S2 provides their machine-readable source files, revision-stage diagnostics, corrected stock replay, and revised figures without replacing the original deposited records. The polygonal environment and third-party assets are not redistributed. The deposited package should be cited and used with attention to the consent scope and re-identification risk of free-speech transcripts.

**Acknowledgments:** During manuscript revision, the author used OpenAI GPT-5.3-Codex (accessed on 2 October 2026) and Anthropic Claude 5.5 (accessed on 13 August 2026) to assist with manuscript restructuring, language revision, consistency checking, reference verification, and analysis-code audit. ChatGPT (OpenAI; GPT-5.6 sol) was also used for Korean questionnaire translation review, as described in Section 8.3. The author reviewed and edited all outputs and takes full responsibility for the content of this publication.

**Conflicts of Interest:** The author is the founder and CEO of RECON Labs Inc., which develops commercial generative-AI production tools and provided employment, laboratory space, hardware, software infrastructure, and administrative support for this research. The study did not evaluate a commercial RECON Labs product. The author was solely responsible for the study design, data collection, analysis, interpretation, and manuscript preparation.

## References

1. Inworld AI. Realtime API: Expressive Speech-to-Speech with Realtime TTS-2. Available online: <https://inworld.ai/realtime-api> (accessed on 8 August 2026).
2. Convai. Convai—Conversational AI for Virtual Worlds. Available online: <https://www.convai.com/> (accessed on 8 August 2026).
3. De La Torre, F.; Fang, C.M.; Huang, H.; Banburski-Fahey, A.; Amores Fernandez, J.; Lanier, J. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), Honolulu, HI, USA, 11–16 May 2024; Article 600, pp. 1–22. [CrossRef]
4. Lee, J.; Wang, J.; Brown, E.; Chu, L.; Rodriguez, S.S.; Froehlich, J.E. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’24), Honolulu, HI, USA, 11–16 May 2024; Article 408, pp. 1–20. [CrossRef]
5. Li, C.; Wu, G.; Chan, G.Y.-Y.; Turakhia, D.G.; Castelo Quispe, S.; Li, D.; Welch, L.; Silva, C.T.; Qian, J. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25), Yokohama, Japan, 26 April–1 May 2025; Article 1229, pp. 1–24. [CrossRef]
6. Lee, G.; Xia, M.; Numan, N.; Qian, X.; Li, D.; Chen, Y.; Kulshrestha, A.; Chatterjee, I.; Zhang, Y.; Manocha, D.; et al. Sensible Agent: A Framework for Unobtrusive Interaction with Proactive AR Agents. In Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST ’25), Busan, Republic of Korea, 28 September–1 October 2025; Article 119, pp. 1–22. [CrossRef]
7. Yang, B.; Xu, L.; Zeng, L.; Guo, Y.; Jiang, S.; Lu, W.; Liu, K.; Li, Y.; Jiang, X.; Xing, G.; et al. ProAgent: Harnessing On-Demand Sensory Contexts for Proactive LLM Agent Systems in the Wild. *arXiv* **2025**, arXiv:2512.06721.
8. Bandyopadhyay, S.; Bahirwani, V.; Aggarwal, L.; Guda, B.; Li, L.; Colaco, A. YETI (YET to Intervene) Proactive Interventions by Multimodal AI Agents in Augmented Reality Tasks. *arXiv* **2025**, arXiv:2501.09355.
9. Zhang, Z.; Kang, Z.; Zhao, R.; Feng, Y.; Jiang, B.; Ji, H.; Liu, L. ProAct: Harnessing Streaming Motion Generation and Agentic Reasoning for Real-Time Embodied Social Interaction. *ACM Trans. Graph.* **2026**, in press. [CrossRef]
10. Horvitz, E. Principles of Mixed-Initiative User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’99), Pittsburgh, PA, USA, 15–20 May 1999; pp. 159–166. [CrossRef]
11. Horvitz, E.; Jacobs, A.; Hovel, D. Attention-Sensitive Alerting. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI ’99), Stockholm, Sweden, 30 July–1 August 1999; pp. 305–313.

12. Horvitz, E.; Koch, P.; Apacible, J. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04), Chicago, IL, USA, 6–10 November 2004; pp. 507–510. [\[CrossRef\]](#)
13. Fogarty, J.; Hudson, S.E.; Atkeson, C.G.; Avrahami, D.; Forlizzi, J.; Kiesler, S.; Lee, J.C.; Yang, J. Predicting Human Interruptibility with Sensors. *ACM Trans. Comput.-Hum. Interact.* **2005**, *12*, 119–146. [\[CrossRef\]](#)
14. Iqbal, S.T.; Bailey, B.P. Effects of Intelligent Notification Management on Users and Their Tasks. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08), Florence, Italy, 5–10 April 2008; pp. 93–102. [\[CrossRef\]](#)
15. Giunchi, D.; Numan, N.; Gatti, E.; Steed, A. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. In Proceedings of the 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Orlando, FL, USA, 16–21 March 2024; pp. 579–589. [\[CrossRef\]](#)
16. Dogan, M.D.; Gonzalez, E.J.; Ahuja, K.; Du, R.; Colaço, A.; Lee, J.; Gonzalez-Franco, M.; Kim, D. Augmented Object Intelligence with XR-Objects. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), Pittsburgh, PA, USA, 13–16 October 2024; Article 19, pp. 1–15. [\[CrossRef\]](#)
17. Bovo, R.; Abreu, S.; Ahuja, K.; Gonzalez, E.J.; Cheng, L.-T.; Gonzalez-Franco, M. EmBARDiment: An Embodied AI Agent for Productivity in XR. In Proceedings of the 2025 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Saint-Malo, France, 8–12 March 2025; pp. 708–717. [\[CrossRef\]](#)
18. Pan, M.; Kitson, A.; Wan, H.; Prpa, M. ELLMA-T: An Embodied LLM-Agent for Supporting English Language Learning in Social VR. In Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25), Funchal, Portugal, 5–9 July 2025; pp. 576–594. [\[CrossRef\]](#)
19. NVIDIA. NVIDIA ACE—ACE Documentation. Available online: <https://docs.nvidia.com/ace/latest/index.html> (accessed on 8 August 2026).
20. Slater, M. Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments. *Philos. Trans. R. Soc. B* **2009**, *364*, 3549–3557. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Slater, M.; Wilbur, S. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence Teleoper. Virtual Environ.* **1997**, *6*, 603–616. [\[CrossRef\]](#)
22. Witmer, B.G.; Singer, M.J. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence Teleoper. Virtual Environ.* **1998**, *7*, 225–240. [\[CrossRef\]](#)
23. Schubert, T.; Friedmann, F.; Regenbrecht, H. The Experience of Presence: Factor Analytic Insights. *Presence Teleoper. Virtual Environ.* **2001**, *10*, 266–281. [\[CrossRef\]](#)
24. Skarbez, R.; Brooks, F.P., Jr.; Whitton, M.C. A Survey of Presence and Related Concepts. *ACM Comput. Surv.* **2017**, *50*, 96. [\[CrossRef\]](#)
25. Biocca, F.; Harms, C.; Burgoon, J.K. Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria. *Presence Teleoper. Virtual Environ.* **2003**, *12*, 456–480. [\[CrossRef\]](#)
26. Oh, C.S.; Bailenson, J.N.; Welch, G.F. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Front. Robot. AI* **2018**, *5*, 114. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Blascovich, J.; Loomis, J.; Beall, A.C.; Swinth, K.R.; Hoyt, C.L.; Bailenson, J.N. Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychol. Inq.* **2002**, *13*, 103–124. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Fox, J.; Ahn, S.J.; Janssen, J.H.; Yeykelis, L.; Segovia, K.Y.; Bailenson, J.N. Avatars Versus Agents: A Meta-Analysis Quantifying the Effect of Agency on Social Influence. *Hum.-Comput. Interact.* **2015**, *30*, 401–432. [\[CrossRef\]](#)
29. Nowak, K.L.; Biocca, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence Teleoper. Virtual Environ.* **2003**, *12*, 481–494. [\[CrossRef\]](#)
30. von der Pütten, A.M.; Krämer, N.C.; Gratch, J.; Kang, S.-H. "It Doesn't Matter What You Are!" Explaining Social Effects of Agents and Avatars. *Comput. Hum. Behav.* **2010**, *26*, 1641–1650. [\[CrossRef\]](#)
31. Slater, M.; Steed, A. A Virtual Presence Counter. *Presence Teleoper. Virtual Environ.* **2000**, *9*, 413–434. [\[CrossRef\]](#)
32. Rivi re, J.-P.; Malo, R.; Varlin Grassi, S.; Pri , Y. What's in a BIP? Exploring the Lived Experiences of Breaks in Presence. *Virtual Real.* **2026**, *30*, 126. [\[CrossRef\]](#)
33. Pejovic, V.; Musolesi, M. Anticipatory Mobile Computing: A Survey of the State of the Art and Research Challenges. *ACM Comput. Surv.* **2015**, *47*, 47:1–47:29. [\[CrossRef\]](#)
34. Lu, Y.; Yang, S.; Qian, C.; Chen, G.; Luo, Q.; Wu, Y.; Wang, H.; Cong, X.; Zhang, Z.; Lin, Y.; et al. Proactive Agent: Shifting LLM Agents from Reactive Responses to Active Assistance. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
35. Chen, T.; Batchelder, N.S.; Liu, A.; Smith, N.A.; Gollakota, S. LlamaPIE: Proactive In-Ear Conversation Assistants. In Proceedings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025; pp. 13801–13824. [\[CrossRef\]](#)
36. Patel, D.A.; Melvin, I.; Malon, C.; Min, M.R. DiscussLLM: Teaching Large Language Models When to Speak. *arXiv* **2025**, arXiv:2508.18167.

37. Pu, K.; Lazaro, D.; Arawjo, I.; Xia, H.; Xiao, Z.; Grossman, T.; Chen, Y. Assistance or Disruption? Exploring and Evaluating the Design and Trade-offs of Proactive AI Programming Support. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan, 26 April–1 May 2025; pp. 1–21. [\[CrossRef\]](#)
38. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fournery, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Glasgow, UK, 4–9 May 2019; Paper 3, pp. 1–13. [\[CrossRef\]](#)
39. Meurisch, C.; Mihale-Wilson, C.A.; Hawlitschek, A.; Giger, F.; Müller, F.; Hinz, O.; Mühlhäuser, M. Exploring User Expectations of Proactive AI Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 146. [\[CrossRef\]](#)
40. Kraus, M.; Wagner, N.; Minker, W. Effects of Proactive Dialogue Strategies on Human-Computer Trust. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), Genoa, Italy, 12–18 July 2020; pp. 107–116. [\[CrossRef\]](#)
41. Summers, T.R.; Yao, S.; Narasimhan, K.; Griffiths, T.L. Cognitive Architectures for Language Agents. *arXiv* **2024**, arXiv:2309.02427.
42. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), San Francisco, CA, USA, 29 October–1 November 2023; Article 2, pp. 1–22. [\[CrossRef\]](#)
43. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv* **2024**, arXiv:2305.16291.
44. Krakovna, V.; Orseau, L.; Martic, M.; Legg, S. Penalizing Side Effects Using Stepwise Relative Reachability. In Proceedings of the Workshop on Artificial Intelligence Safety (AISafety 2019), Co-Located with IJCAI-19 CEUR Workshop Proceedings, Macao, China, 11–12 August 2019. CEUR Workshop Proceedings Vol. 2419.
45. Grinsztajn, N.; Ferret, J.; Pietquin, O.; Preux, P.; Geist, M. There Is No Turning Back: A Self-Supervised Approach for Reversibility-Aware Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Virtual, 6–14 December 2021; pp. 1898–1911.
46. Riedl, M.; Saretto, C.J.; Young, R.M. Managing Interaction Between Users and Agents in a Multi-Agent Storytelling Environment. In Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS '03), Melbourne, Australia, 14–18 July 2003; pp. 741–748. [\[CrossRef\]](#)
47. Raux, A.; Eskenazi, M. A Finite-State Turn-Taking Model for Spoken Dialog Systems. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2009), Boulder, CO, USA, 31 May–5 June 2009; pp. 629–637.
48. Ramnauth, R.; Brščić, D.; Scassellati, B. Should I Help?: A Skill-Based Framework for Deciding Socially Appropriate Assistance in Human-Robot Interactions. In Proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2024), Pasadena, CA, USA, 26–30 August 2024; pp. 2051–2058. [\[CrossRef\]](#)
49. Edwards, J.; Doyle, P.R.; Branigan, H.P.; Cowan, B.R. Comparing Perceptions of Static and Adaptive Proactive Speech Agents. In Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24), Luxembourg, 8–10 July 2024; pp. 1–12. [\[CrossRef\]](#)
50. Mintz, Y.; Aswani, A.; Kaminsky, P.; Flowers, E.; Fukuoka, Y. Nonstationary Bandits with Habituation and Recovery Dynamics. *Oper. Res.* **2020**, *68*, 1493–1516. [\[CrossRef\]](#)
51. Lee, K.; Chu, D.; Cuervo, E.; Kopf, J.; Degtyarev, Y.; Grizan, S.; Wolman, A.; Flinn, J. Outatime: Using Speculation to Enable Low-Latency Continuous Interaction for Mobile Cloud Gaming. In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '15), Florence, Italy, 19–22 May 2015; pp. 151–165. [\[CrossRef\]](#)
52. van Waveren, J.M.P. The Asynchronous Time Warp for Virtual Reality on Consumer Hardware. In Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology (VRST '16), Munich, Germany, 2–4 November 2016; pp. 37–46. [\[CrossRef\]](#)
53. Hua, W.; Wan, M.; Vadrevu, S.; Nadel, R.; Zhang, Y.; Wang, C. Interactive Speculative Planning: Enhance Agent Efficiency through Co-design of System and User Interface. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, 24–28 April 2025.
54. Zilberstein, S. Using Anytime Algorithms in Intelligent Systems. *AI Mag.* **1996**, *17*, 73–83. [\[CrossRef\]](#)
55. Ban, S.; Yi, T.; Hyun, K.H. Intersection of Seeing: New Ways of Experiencing Reality Using Autonomous Volumetric Capture System. *Proc. ACM Comput. Graph. Interact. Tech.* **2023**, *6*, 19:1–19:9. [\[CrossRef\]](#)
56. Latoschik, M.E.; Roth, D.; Gall, D.; Achenbach, J.; Waltemate, T.; Botsch, M. The Effect of Avatar Realism in Immersive Social Virtual Realities. In Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST '17), Gothenburg, Sweden, 8–10 November 2017; Article 39, pp. 1–10. [\[CrossRef\]](#)
57. Harms, C.; Biocca, F. Internal Consistency and Reliability of the Networked Minds Social Presence Measure. In Proceedings of the PRESENCE 2004: 7th Annual International Workshop on Presence, Valencia, Spain, 13–15 October 2004; pp. 246–251.
58. Kukshinov, E.; Tu, J.; Szita, K.; Senthil Nathan, K.; Nacke, L.E. Widespread yet Unreliable: A Systematic Analysis of the Use of Presence Questionnaires. *Interact. Comput.* **2026**, *38*, 163–178. [\[CrossRef\]](#)

59. Degenhard, A.; Askari, A.; Rietzler, M.; Rukzio, E. When Do We Feel Present in a Virtual Reality? Towards Sensitivity and User Acceptance of Presence Questionnaires. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), Yokohama, Japan, 26 April–1 May 2025; pp. 1–29. [\[CrossRef\]](#)
60. Ban, S. *Agentic Presence—Study Data, Arbiter Implementation, and Analysis Package*, Version 2; Zenodo: Geneva, Switzerland, 2026. [\[CrossRef\]](#)
61. Sun, C.; Han, J.; Deng, W.; Wang, X.; Qin, Z.; Gould, S. 3D-GPT: Procedural 3D Modeling with Large Language Models. In Proceedings of the 2025 International Conference on 3D Vision (3DV), Singapore, 25–28 March 2025; pp. 1253–1263. [\[CrossRef\]](#)
62. Hu, Z.; Iscen, A.; Jain, A.; Kipf, T.; Yue, Y.; Ross, D.A.; Schmid, C.; Fathi, A. SceneCraft: An LLM Agent for Synthesizing 3D Scenes as Blender Code. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024; PMLR 235, pp. 19252–19282.
63. Engstler, P.; Shtedritski, A.; Laina, I.; Rupprecht, C.; Vedaldi, A. SynCity: Training-Free Generation of 3D Worlds. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2025), Honolulu, HI, USA, 19–23 October 2025; pp. 27585–27595.
64. Yao, K.; Zhang, L.; Yan, X.; Zeng, Y.; Zhang, Q.; Xu, L.; Yang, W.; Gu, J.; Yu, J. CAST: Component-Aligned 3D Scene Reconstruction from an RGB Image. *ACM Trans. Graph.* **2025**, *44*, 1–19. [\[CrossRef\]](#)
65. Lee, S.W.; Jo, T.H.; Jin, S.; Choi, J.; Yun, K.; Bromberg, S.; Ban, S.; Hyun, K.H. The Impact of Sketch-Guided vs. Prompt-Guided 3D Generative AIs on the Design Exploration Process. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, 11–16 May 2024; Article 1057, pp. 1–18. [\[CrossRef\]](#)
66. Liu, B.; Song, W.; Zheng, M.; Fu, C.; Chen, J.; Wang, X. Semantically enhanced selective image encryption scheme with parallel computing. *Expert Syst. Appl.* **2025**, *279*, 127404. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.