SHARE POSES, NOT PIXELS: AN INTERVENTIONAL STUDY OF SHARED WORLD MODELS FOR MULTI-AGENT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A navigation world model predicts what an agent will see after acting. It is natural to ask whether several agents could share one, each broadcasting its intent so that every agent's imagined future contains the others, and whether such a shared model could serve as a negotiation device, scoring one's own plans by the reactions they provoke. We test this with a conditional diffusion world model in the style of Navigation World Models, extended with a peer channel, inside a simulator in which interaction strength is a dial and a control world with exactly zero interaction exists for every scene. (i) The model learns to use a broadcast pose: a matched difference-in-differences against the control puts the broadcast at -0.014 LPIPS (95% CI $[-0.017, -0.012]$), five to eight times its value where nothing interacts. (ii) In closed-loop decentralised planning against neighbours of unknown disposition, reading the neighbour out of the imagined frame is no better than ignoring it (46.7% vs. 48.9% collisions) while a six-bit broadcast of its declared pose matches an oracle (2.2%), and the gap widens from 17 to 67 points as the crowd grows from 2 to 5. A ladder of oracles attributes this to field of view, the pixel readout, and the imagination, which draws a neighbour within one metre at 0.42 of its size; a readout fitted to the model's own frames does not recover it. (iii) A state head trained jointly on the same trunk does: it places the neighbour 3.2 s ahead to 0.15 m in one forward pass and learns how the neighbour reacts to the ego's action (slope 0.39 against 0.00 in the control world), a dependence the pixel objective never acquires. Yet in closed loop this head collides nearly three times as often as constant-velocity extrapolation from the same input, because what it learned is the training population's policy and the neighbour in front of it may not share it. Pixels lose the number; a learned model of others must bet on their policy; a declared intent does neither. In this setting, sharing should happen at the level of poses and intents rather than pixels.

1 INTRODUCTION

World models that predict egocentric observations conditioned on actions have become a practical tool for navigation: Navigation World Models (Bar et al., 2025, NWM) plan by imagining the frames a candidate trajectory would produce and scoring them against a goal image. When several such agents move through one space, the natural extension is a *shared* world model: each agent broadcasts a compact description of its state and intent, every agent conditions its imagination on what it hears, and, most ambitiously, each agent evaluates its own candidate plans by the reactions they would provoke in the others, turning the simulator into a negotiation device.

This paper asks whether that programme survives contact with measurement. Does a peer-conditioned world model actually *use* a broadcast pose, or does it learn to ignore the channel? Can it infer a neighbour's *reaction* to the ego's own action, which is what negotiation requires? And when the imagined frames are put to work in a closed-loop planner whose success is measured in collisions rather than in pixel loss, does a shared world model make a navigator better at its job than a broadcast of the same information consumed in closed form?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recorded multi-robot data cannot answer them, in principle: the reaction of one agent to another is a counterfactual (*what would the neighbour have done had I acted differently*), and a dataset records one outcome per scene. Worse, a standard observational measure of interaction returns the same value in a simulator with interaction switched *off* as on the real dataset, because agents sharing a room and a doorway are correlated whether or not they react to each other (Section 3). We therefore built an interventional testbed: a GPU ray-casting simulator in which reactivity is a dial, a control world with exactly zero interaction exists for every scene and seed, and any quantity can be measured as a difference-in-differences.

Our contributions are the answers, and the instruments that made them measurable.

- **The model uses the broadcast, but only for where the neighbour is.** On EgoTeam (Peng et al., 2026) the peer embedding collapses to a constant during training (its input dependence falls $39\times$ while the ego-action embedding's rises $24\times$), and the data offer nothing to learn. In the interventional world the channel comes alive: a matched difference-in-differences puts the value of the broadcast at -0.0144 LPIPS, five to eight times its value in the zero-interaction control (Section 4).
- **A pose broadcast beats a shared world model at the task, and the margin grows with the crowd.** In closed-loop planning against a neighbour of unknown disposition, the world-model conditions collide as often as a planner that ignores the neighbour, while a broadcast of the neighbour's declared future pose matches a full oracle. A ladder of oracles decomposes the loss into field of view (6.7 pp), pixel readout (10 pp) and the imagination itself (30 pp), which draws a neighbour within one metre at 0.42 of its true size; a readout fitted to the model's own frames corrects the bias but not the loss. The gap widens from 17 to 67 ppas agents are added (Section 5).
- **A state head recovers what the pixels lose, and what it recovers is a bet.** A state head trained jointly on the same trunk places the neighbour 3.2 s ahead to 0.15 m (0.40 m for the best pixel readout) in one forward pass, and learns how the neighbour reacts to the ego's action (slope 0.39; 0.00 for the same head trained in the control world), a dependence the pixel objective never acquires: the denoising loss scores a shuffled-action placebo the same as the informed arm, and a probe on the frozen pixel-trained trunk misplaces the neighbour by 0.68 m. In closed loop the head nonetheless collides 47% of the time against an assertive neighbour, nearly three times the rate of constant-velocity extrapolation from the same input, because it predicts the yielding it was trained on (Section 6).
- **A specification for the channel.** One bit per broadcast scalar retains 82% of the benefit; packet loss is worse than never having a link unless the peer channel is trained with dropout; frames must be registered to within 10° (Appendix C).

The results are not that world models are bad: the trunk learns where a broadcast neighbour will be and, given the state as a target, even how it reacts to the ego. They are that a first-person image is the wrong carrier for the quantity a collision constraint needs (closest approach, attained exactly when the neighbour has left the frame or overflows it), and that a learned model of other agents, however read out, encodes the policy of the agents it was trained on. A declared intent has neither problem. Figure 1 summarizes the prototype, the three routes to the planner, and the interventional controls.

2 Related work

Navigation world models. NWM (Bar et al., 2025) trains a conditional diffusion transformer (CDiT; building on Peebles & Xie, 2023) to predict a future egocentric frame from context frames, an ego action and a time offset, and plans by sampling trajectories and scoring imagined frames with LPIPS (Zhang et al., 2018). We keep this recipe unchanged and add a peer channel. Our planner audit (Appendix A) finds NWM's planning error on RECON (Shah et al., 2021) dominated by the LPIPS scoring channel rather than the world model ($r = 0.12$ across trajectories), which is why we measure the multi-agent question with task metrics.

Multi-agent world models. COMBO (Zhang et al., 2025) factorises joint actions across agents and plans by tree search with a vision-language model inferring others' intents; GAIA-2 (Russell et al., 2025) conditions a driving world model on other agents' states; MASS (Cai et al., 2026)

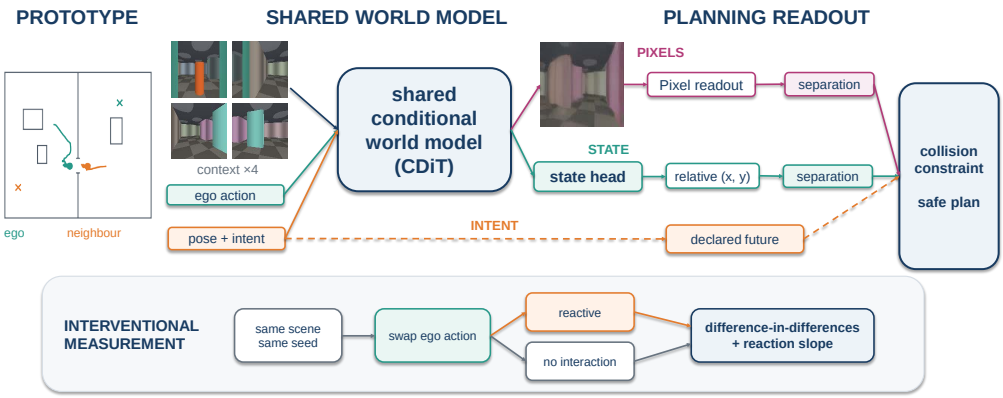

Figure 1: **Method overview: prototype and measurement in one instrument.** In the doorway prototype the ego evaluates a candidate action while a neighbour broadcasts its current pose and declared intent; a shared world model consumes the egocentric context and these inputs. Three routes reach the collision constraint: separation read from the imagined pixels, relative position read by a state head, and the declared intent used in closed form. Below, interventions replace the ego action in the same scene and seed, and a matched zero-interaction world removes the neighbour's reaction, so that difference-in-differences and reaction slope isolate what the model learned about interaction.

advances an authoritative typed state from joint actions and renders each player's view from it, for 1,024 concurrent players. These works *build* shared simulation, and MASS in particular shares state rather than pixels by design; we *measure*, with interventional controls, what an egocentric world model of the NWM kind can learn about other agents, and arrive at the same design from the other direction. Communication-constrained multi-agent RL (Sukhbaatar et al., 2016; Das et al., 2019; Foerster et al., 2016) studies what to send under bandwidth limits; our channel-degradation sweep is in that spirit, applied to a generative world model.

Anticipatory collision avoidance. Reciprocal velocity obstacles (van den Berg et al., 2011) and social forces (Helbing & Molnár, 1995) anticipate closest approach from broadcast or estimated velocities; our data-generating policy is of this family, and BROADCAST-CV is a planner given what ORCA-style methods assume.

3 SETUP: A WORLD IN WHICH INTERACTION IS A DIAL

World model and peer channel. We use CDiT-S/2 at 112×112 with a four-frame context, trained for 40k steps with the NWM recipe (SD-VAE latents, 4 channels, relative-time conditioning on ± 8 frames). A peer embedder, an MLP of the same design as the action embedder, adds a six-dimensional channel: the neighbour's current position and heading in the ego frame, and its *declared* future position and heading. Two arms differ only in this channel: NONE receives zeros, FULL receives it intact. All comparisons are paired (same context, same goal, same diffusion noise) and tested with two-sided Wilcoxon signed-rank tests.

Why not recorded data. We first trained both arms on the two-robot clips of EgoTeam (Peng et al., 2026), a multi-robot egocentric dataset. The peer embedding's input dependence, the fraction of its output energy that varies with its input, fell from 0.174 at initialisation to 0.0045 after training, while the ego-action embedding's rose from 0.023 to 0.549 (Appendix B). The channel was neither sparse nor redundant with the ego action; gradient descent switched it off. The data explain why: the ego's future action adds only 0.7% of variance to a prediction of the neighbour's displacement beyond its own momentum and geometry. But this observational number is uninformative. In our simulator with reactivity set to exactly zero, where an intervention on the ego moves the neighbour by 0.0000 m, the same regression returns 0.0073, against 0.0071 on EgoTeam. Two agents walking the same room towards the same doorway are correlated whether or not they react to each other; the question is counterfactual and needs an interventional instrument.

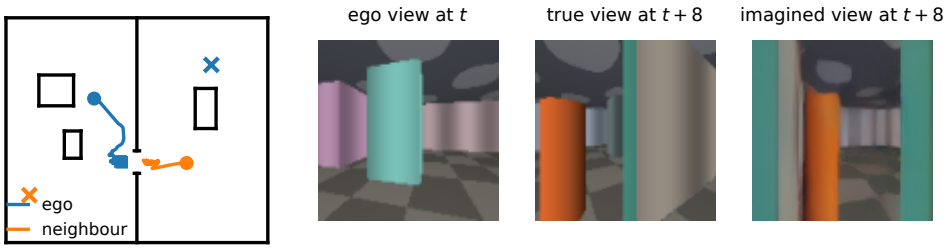

Figure 2: The interventional world. **Left:** a hall split by a wall with one gap; the two agents start on opposite sides with goals across it (crosses) and must negotiate the opening. Open and filled squares mark the ego at t and $t{+}8$. **Middle:** the ego's true view at t and at $t{+}8$, with the neighbour (orange) entering the frame. **Right:** what the FULL arm imagines it will see at $t{+}8$, given four context frames, its own future pose and the neighbour's broadcast. The neighbour is drawn where the broadcast says it will be.

Table 1: **Left:** LPIPS of NONE and FULL on the interactive world by neighbour distance and horizon; Δ = FULL − NONE, lower is better; Wilcoxon p. **Right:** matched difference-in-differences against the zero-interaction control (neighbour at 1–2 m, in view, displaced $<$ 0.3 m over the horizon); bootstrap CIs, 4000 resamples.

slice	n	NONE	FULL	Δ	p
all	1600	0.2349	0.2156	-0.0192	$3{\times}10^{-11}$
peer 1–2 m	1056	0.2679	0.2398	-0.0281	10^{-18}
peer 2–4 m	248	0.1978	0.1929	-0.0049	0.10
peer $>$ 4 m	296	0.1481	0.1484	$+0.0003$	0.30
horizon 1–2 s	571			-0.0143	0.009
horizon 2–4 s	724			-0.0291	$4{\times}10^{-12}$

matching	interactive Δ	control Δ	DiD [95% CI]
distance only	-0.0129	-0.0079	-0.0051 [-0.009, -0.001]
+ displacement, seed 0	-0.0127	-0.0018	-0.0110 [-0.015, -0.007]
+ displacement, seed 1	-0.0206	-0.0027	-0.0179 [-0.022, -0.014]
pooled			$\mathbf{-0.0144}$ [-0.017, -0.012]

The simulator. A bounded hall (12–16 m by 10–13 m) with pillars is split by a wall with a single gap of 0.95–1.25 m; two agents start on opposite sides with goals across the gap, so they must negotiate one opening (Fig. 2). Rendering is GPU ray casting at 112×112 with a $90°$ horizontal field of view; the neighbour is an upright cylinder of radius 0.25 m in a colour unique in the world, occupying 11.9% of the frame at a median 1.5 m (1–4% on EgoTeam). The data-generating policy is anticipatory: each agent accelerates toward its goal, is pushed from walls, and yields to the other in proportion to a *reactivity* dial based on time to closest approach. Reactivity zero yields a *control world*: same walls, seeds and goals, but neighbours that ignore the ego. Intervening on the ego (a left turn or a stop held for 3 s) displaces the neighbour by 0.26 m at reactivity 4 and by exactly 0 in the control. The broadcast intent of a neighbour is its own policy rolled forward with the ego held at its current velocity: what a decentralised agent can honestly transmit before the ego has chosen.

4 Does the model use the broadcast?

On 1600 paired target frames, FULL improves LPIPS over NONE by 0.0192 ($p = 3{\times}10^{-11}$), with the structure interaction predicts: -0.0281 when the neighbour is 1–2 m away, -0.0049 at 2–4 m, $+0.0003$ beyond 4 m; -0.0143 at 1–2 s horizons and -0.0291 at 2–4 s (Table 1). On EgoTeam the pattern was the reverse (largest difference at the greatest distance) and FULL was slightly *worse*, consistent with a constant-bias perturbation.

The confound, and the matched difference-in-differences. The -0.0192 cannot be read as "the model learned about interaction". A broadcast pose is useful in a world with *no* interaction: it says where a large object will be. To isolate the interaction-specific value we train all four arms (NONE/FULL $\times$ interactive/control) and take the difference-in-differences (FULL − NONE | interactive) − (FULL − NONE | control). Binned by distance alone this gave -0.0185 at 1–2 m. But almost the entire benefit sits in one cell, neighbour within 2 m *and* nearly stationary over the horizon ($\Delta = -0.0278$; other cells ≈ 0), which the two worlds populate very differently (55.1% vs. 4.1% of anchors) because a neighbour that yields stops at the doorway; composition alone predicts a distance-binned DiD of about -0.017. Matching on displacement as well as distance, the

Table 2: Closed-loop planning, 2 agents, doorway layout. Model-free conditions: 90 episodes (seeds 0–89); world-model conditions: the first 30 of the same seeds. Collision rate against the assertive neighbour is the discriminating metric. World-model conditions use a pixel readout with margin 1.25 m (tuned in their favour, Fig. 3); Fisher exact tests against BROADCAST-INTENT and BLIND.

condition	assertive: coll.	yielding: coll.	both succeed	p vs. intent / blind
REACTIVE (no planner)	0.0%	0.0%	38.9% (deadlocks)	
BLIND	48.9%	0.0%	47.8%	
BROADCAST-CV	16.7%	0.0%	78.9%	
BROADCAST-INTENT	**2.2%**	0.0%	**93.3%**	
NONE (world model, no broadcast)	46.7%	0.0%	53.3%	2×10^{-8} / 1.0
FULL (world model + broadcast)	46.7%	0.0%	53.3%	2×10^{-8} / 1.0
ORACLE	0.0%	0.0%	92.2%	

interaction-specific value is -0.0144 LPIPS (pooled over two training seeds, each excluding zero); the control world's own broadcast gain is -0.002, five to eight times smaller. The same confound produced the spurious observational R^2 above.

5 DOES IT HELP THE TASK?

LPIPS on predicted pixels is a weak proxy for planning quality (Appendix A), so we turn to a task metric. Same world and agents; the ego replans every 4 steps over a lookahead of 8 by choosing among eleven candidate velocities (five headings $\times$ two speeds, plus waiting); every condition shares an identical geometric goal term (distance-to-go routed through the doorway) and differs only in its estimate of the neighbour's minimum separation over the lookahead, used as a hard constraint (≥ 0.75 m). Success is reaching the goal within 250 steps without ever coming within 0.5 m of a neighbour.

Heterogeneous neighbours. The data-generating neighbour always yields; barging through it succeeds 95% of the time with zero collisions. We therefore evaluate every policy against two neighbour types it cannot distinguish a priori: *yielding* (the training policy) and *assertive* (holds its line to its own goal, avoiding walls and other neighbours but not the ego). The discriminating metric is the collision rate against the assertive neighbour; we also report success against *both* types.

Conditions. A ladder from knowing nothing about the neighbour to knowing everything: REACTIVE (the data-generating policy, no planner); BLIND (planner with goal term only); BROADCAST-CV (the neighbour's broadcast pose carried forward at constant velocity, no model); BROADCAST-INTENT (the neighbour's declared future pose used directly: the same six numbers the FULL arm receives, consumed in closed form); NONE and FULL (the neighbour's position read out of the frame the world model imagines for each candidate); ORACLE (the neighbour's true future by forking the simulator). BROADCAST-INTENT is the control that decides whether the world model earns its place.

Result. Table 2. REACTIVE never collides but deadlocks at the door 61% of the time; the task is "decide who goes first". A planner that ignores the neighbour resolves the deadlock and collides 49% of the time against an assertive one; extrapolating the broadcast cuts that to a third; the declared intent brings it to 2.2%, indistinguishable from the oracle. The two world-model conditions (FULL given the same declared intent as BROADCAST-INTENT, NONE given nothing) collide 46.7% of the time, statistically indistinguishable from BLIND ($p = 1.0$) and far from BROADCAST-INTENT ($p = 2\times10^{-8}$). The broadcast changes what the model draws, not what the planner does.

5.1 ATTRIBUTION: FIELD OF VIEW, READOUT, OR IMAGINATION?

"The model does not help" has three causes, and a ladder of oracles separates them (Table 3). Grant perfect knowledge of the neighbour's future (ORACLE); then require only that the neighbour be somewhere in the ego's 90° view during the lookahead (ORACLE-FOV); then additionally require its distance to be recovered from the pixels of the *true* rendered frame by the same detector (ORACLE-RENDER); then replace the true frame with the imagined one (FULL). Each rung adds one defect.

Table 3: Attribution ladder, assertive neighbour, 2 agents, the same 30 episodes (seeds 0–29) in every cell. Each rung adds one defect to the one above it. All pixel readouts use the tuned margin of 1.25 m; at the naïve margin (0.75 m) ORACLE-RENDER and FULL both read 40.0% and cannot be told apart. The learned readouts R1 and R2 are described in Section 5.2.

condition	defect added	hand readout	learned readout
ORACLE	(none)	0.0%	
ORACLE-FOV	neighbour must be in view	6.7%	
ORACLE-RENDER	distance read from pixels of the *true* frame	16.7%	23.3% (R1)
FULL	true frame replaced by imagined frame	46.7%	40.0% (R1) / 56.7% (R2)
NONE	and no broadcast	46.7%	40.0% (R1) / 43.3% (R2)
BLIND	(floor)	40.0%	

Table 4: What an imagined frame carries. **Top:** on frames where the neighbour is truly in view, how often the FULL arm draws it and the median ratio of drawn to true size. **Bottom:** readouts on true and imagined frames with the neighbour truly in view within 1.5 m: recall, median position error, and the fraction of truly close (< 1.25 m) neighbours read as close, the event the constraint depends on.

true distance	0–1 m	1–2 m	2–4 m	> 4 m
model draws the neighbour	75%	79%	70%	66%
drawn size / true size	0.42	0.64	0.79	0.88

frames	readout	recall < 1.5 m	position error	close read as close
true	R1	99.7%	0.12 m	88%
imagined (FULL)	R1	78.2%	0.56 m	65%
imagined (FULL)	R2	79.1%	0.40 m	68%
imagined (NONE)	R2	76.1%	0.44 m	67%

Field of view costs 6.7 pp. Not more: a working avoidance policy steers early, while the neighbour is still ahead.

The pixel readout costs 10 pp, for a structural reason. We instrumented the geometrically unsafe candidates (true closest approach < 0.75 m; n=160). The neighbour's pixels are in the frame at the sampled instants in 45.6% of them and the detector returns a reading in 44.4%; yet the reading is small enough to trip the constraint in **0.0%**. The readings are biased optimistic by a median $+0.51$ m (true closest approach 0.41 m, read 0.87 m), more than the whole safety margin. Two structural sources: at closest approach the neighbour is still in frame in only 30.6% of cases (it has passed to the side), and when it is, the reading is $+0.42$ m too far, because a neighbour close enough to matter overflows the frame and a clipped silhouette bounds distance only from above. Raising the margin compensates partially and non-monotonically (40→17→30→30% at 0.75–2.5 m; Fig. 3): past a point every candidate is infeasible and the fallback is blind again.

The imagination costs 30 pp. With the same detector and the same tuned margin, true frames give 16.7% and imagined frames 46.7%. A methodological point: at the naïve margin both rungs read 40.0%, indistinguishable from BLIND, and the imagination appeared to cost nothing. A readout that never fires flattens every difference above it.

5.2 Is it the hand-written detector? Learned readouts

The colour-and-geometry detector is the natural suspect, so we replace it with two learned readouts that differ only in what they were fitted to (Appendix F). **R1** is a small CNN trained on 80k true renders with geometric labels; within 1.5 m it localises the neighbour to 0.12 m with 99.7% recall, where the hand detector failed outright. **R2** is R1 fine-tuned on 7,380 frames *imagined by each arm* from held-out episodes, each labelled with where the neighbour really was at the imagined time, so it may learn the model's habits (blur, shrinkage, misplacement). Table 4 reports what the imagined frames contain and what the readouts recover; Table 3 reports the closed-loop consequence.

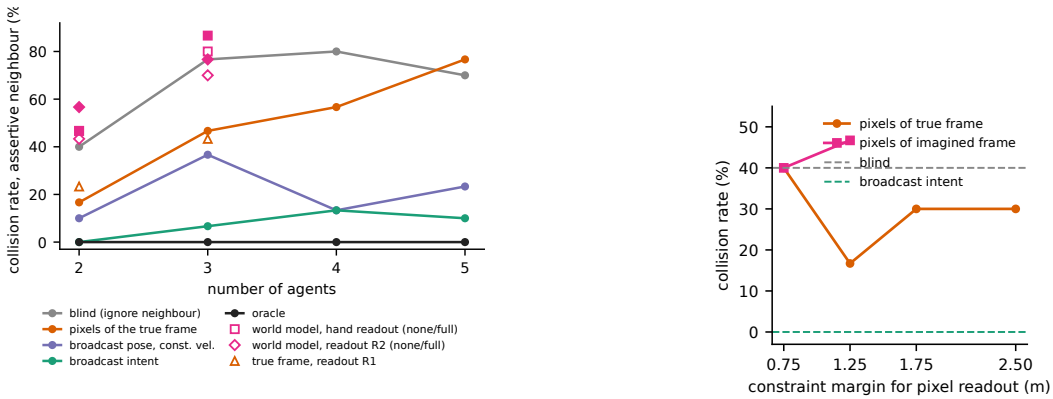

Figure 3: **Left:** collision rate against the assertive neighbour as agents are added (30 episodes per cell). The pixel readout uses *true* frames, the ceiling for any world model; its gap to a pose broadcast widens monotonically (p = 0.052, $9{\times}10^{-4}$, $9{\times}10^{-4}$, $2{\times}10^{-7}$) and at five agents reading pixels is no better than ignoring the neighbour. ORACLE stays at zero: the task remains solvable; the channel does not. **Right:** raising the constraint margin for pixel readouts helps once and then reverses.

The imagination shrinks close neighbours. When the neighbour is truly within 1 m, the model draws it three times in four, at less than half its true size; the ratio recovers with distance. A neighbour drawn at 0.42 of its size is a neighbour read at roughly twice its distance, by any readout that has not been taught otherwise. Out of view, the model draws one anyway 10% of the time.

Fine-tuning removes the bias but not the noise. R1 on imagined frames reads close neighbours $+0.11$ m too far; R2 removes this (-0.05 m) and brings the error from 0.56 to 0.40 m, still 3.3× its error on true frames, while close-read-as-close moves from 65% to 68%, against 88% on true frames.

Closed loop. R1 on *true* frames collides 23.3% of the time against the assertive neighbour, no better than the tuned hand detector (16.7%, $p = 0.75$) and still far from the pose broadcast (2.2%, $p = 8{\times}10^{-4}$). A readout nearly perfect on what is in the frame does not close the gap, because closest approach happens when the neighbour is not in the frame. On imagined frames, R1 gives 40.0% for both arms and R2 gives 43.3% (NONE) and 56.7% (FULL), none distinguishable from BLIND ($p \geq 0.5$) and all far from BROADCAST-INTENT ($p \leq 10^{-6}$). Every readout we could build, including one fitted to the model's own output, leaves the world-model conditions at the floor.

5.3 DOES THE CONCLUSION TRAVEL?

We extended the simulator to any number of agents (every neighbour avoids all others; the two-agent case reproduces bit-for-bit) and reran the ladder. Fig. 3: from 2 to 5 agents, BLIND goes 40→77→80→70%, the true-frame pixel readout 17→47→57→77%, BROADCAST-INTENT 0→7→13→10%, ORACLE stays at 0%. At three agents the world-model conditions collide 80–87% of the time with the hand readout and 70–77% with R2, against BLIND's 76.7% ($p \geq 0.5$) and BROADCAST-INTENT's 6.7% ($p < 10^{-6}$); R1 on true frames gives 43.3% ($p = 0.002$ against the broadcast). Extra agents are extra occluders. In an open hall without a doorway all conditions improve (BLIND 20%) but the order is unchanged (pixels 13.3%, poses 3.3%).

6 READING THE STATE OFF THE TRUNK

Every failure in Section 5 is a failure to get one number out of an image, so we stop going through the image. We add a *state head*: a two-layer MLP on the mean of the target-frame tokens after the last transformer block, trained jointly with the diffusion loss to output the neighbour's position at the target time in the ego frame, the two numbers the FULL arm receives as input (Appendix F). At planning time it is queried once per candidate, at the final diffusion timestep with a noise target, so everything it knows comes from the context frames and the conditioning, which is all an imagined

Table 5: The state head against model-free estimators and the broadcast, two agents. *Error*: median error of the neighbour's predicted position 3.2 s ahead (319 held-out anchors). *Slope / sign*: slope of the predicted on the true change in the neighbour's future when the ego's action is changed, and the fraction of pairs with a true reaction above 0.1 m in which the prediction moves the right way (1276 pairs, one noise target shared across alternatives; ignoring the ego gives slope 0 and sign 50%; Fig. 4). Parentheses: a second training seed. *Bias*, *unsafe→safe*: on the planner's decisions against the assertive neighbour (first 30 episodes), median signed error of predicted separation within 2 m, and the fraction of truly unsafe candidates (< 0.75 m; n=696) read as safe. *Collisions*: assertive neighbour, 90 episodes, same seeds in every row; no row ever hits the yielding neighbour.

estimator	input	error	slope / sign	bias (assertive)	unsafe→safe	collisions
BLIND	–	–	–	–	–	48.9%
current pose, held	pose	0.22 m	–	–	–	–
BROADCAST-CV	pose	0.32 m	–	+0.11 m	22.7%	16.7%
probe on the frozen trunk, NONE	frames	1.55 m	0.48 / 62%	–	–	–
probe on the frozen trunk, OBSERVED	pose + frames	0.68 m	0.16 / 56%	–	–	–
head, NONE	frames	0.51 m	0.22 / 62%	+0.79 m	92.2%	47.8%
head, OBSERVED	pose + frames	**0.15 m** (0.16)	**0.39** / 74% (0.34 / 71%)	+0.29 m	56.0%	46.7%
trained with interventional branches	pose + frames	0.16 m	0.33 / 73%	+0.36 m	60.2%	42.2%
trained in the zero-interaction control	pose + frames	1.16 m (1.08)	0.00 / 51% (0.01 / 49%)	+0.11 m	20.5%	14.4%
worst case of the two heads above	pose + frames	–	–	–	–	10.0%
BROADCAST-INTENT	declared future	–	–	–	–	**2.2%**

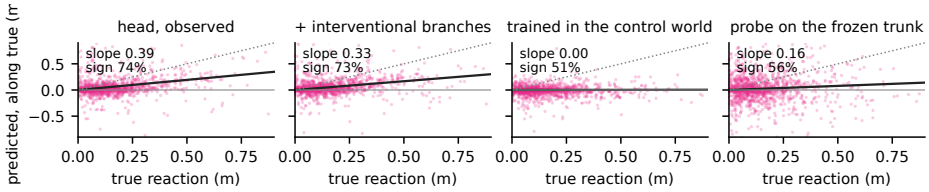

Figure 4: Does the predicted neighbour move with the ego's action? One point per (anchor, alternative action) pair: true change in the neighbour's future against the change in the prediction, projected on the true direction. Solid: slope through the origin; dotted: identity.

frame could have carried. We train it on the NONE arm (frames only) and the OBSERVED arm (frames plus the neighbour's current pose); on FULL the target is the input. Trained jointly at unit weight the head is not free (paired LPIPS on 1600 held-out targets rises from 0.227 to 0.275 for OBSERVED and from 0.235 to 0.473 for NONE), so we also fit a *probe*, a head on the frozen headless trunk of Section 4, which settles whether the pixel-trained representation already holds the number or the auxiliary loss puts it there.

Where the neighbour will be: solved. Table 5, column *error*. Given the neighbour's current pose, the head places it 3.2 s later to a median 0.15 m, better than holding it still (0.22 m) or extrapolating it (0.32 m), and nearly three times better than the best pixel readout on imagined frames (0.40 m, Table 4, on the easier subset of frames where the neighbour is in view). From frames alone the head manages 0.51 m with a heavy tail (mean 1.09 m) where the neighbour is out of view. The probe on the frozen trunk does not get there: 0.68 m given the pose and 1.55 m from frames, worse than holding the neighbour still, although its readout has $60\times$ the parameters of the joint head (Appendix F). The number is not lying in the pixel-trained representation waiting to be read; the auxiliary loss puts it there, at the cost to the pixels above. The head shows not what the world model knew but what the same trunk, data and conditioning can be made to know when the target is the state rather than the image. On EgoTeam's recorded robots the same head beats holding the neighbour still (0.35 vs. 0.38 m) but not extrapolating its velocity (0.23 m; Appendix B): what it learned here is the doorway, not motion in general.

Whether it depends on the ego: yes, and the pixels hide it. Negotiation requires knowing how the neighbour's future depends on the ego's action. From an anchor, the ego's true action and mild alternatives (veer $\pm 30°$, slow, stop) give, by re-simulation, the neighbour's true reaction (median 0.19 m) and, from the control world, a frozen neighbour that does not react; the model scores both target frames by its denoising loss. The OBSERVED arm prefers the reacting frame 55.7% of the time (n=1281), but so does NONE (55.8%) and so does a placebo whose actions are shuffled against

the frames (56.4%; paired Wilcoxon $p = 0.41$); whether the reacting neighbour ends nearer or farther explains all of it (62% vs. 44%, identical across arms). The loss prefers frames with a closer neighbour and carries no trace of the reaction. Interventional training data, 1,740 re-simulated branches in which the ego's action is replaced and the neighbour reacts (Appendix F), leave the pixel score at the placebo ($p = 0.13$), and a $4\times$ larger model gains at most four points (Appendix E).

Then we asked the head, with the same anchors and alternatives (Table 5, column *slope / sign*; Fig. 4). Its prediction of the neighbour moves with the ego's action: slope 0.39 on the true change (bootstrap 95% CI $[0.30, 0.47]$ over anchors), with the right sign in 74% of pairs where the reaction exceeds 0.1 m. That this is knowledge of the reaction and not of the room is settled by the control: the same head trained in the zero-interaction world (same halls, same doorways, same co-location, neighbours that ignore the ego) has slope 0.00 ($[-0.04, 0.04]$) and sign agreement 51%. A second training seed of each reproduces both (0.34 $[0.25, 0.42]$ and 71%; 0.01 $[-0.03, 0.06]$ and 49%). The interventional branches add nothing (0.33): in a simulator without hidden confounders, the observational conditional is already the interventional one. So the dependence of the neighbour's future on the ego's action is learnable, from watching, by this trunk on this data. The pixel objective does not learn it: a 0.19 m reaction is a few pixels and the denoising loss is dominated by where the neighbour is; the probes on the frozen trunks, which see the ego's action through the conditioning, recover at most a trace of it (sign 56% and 62%, against 74% for the joint head) while misplacing the neighbour by 0.7–1.6 m, as a readout can learn the average response to a veer or a stop without knowing where the neighbour is.

In closed loop: a model of the neighbour is a bet on its policy. The head that predicts the yielding neighbour to 0.15 m collides with the assertive one 46.7% of the time (n=90), no better than ignoring it (48.9%, Fisher $p = 0.88$) and nearly three times worse than constant-velocity extrapolation from the very same input (16.7%, $p < 10^{-4}$). Columns *bias* and *unsafe→safe* say why. Against the yielding neighbour every estimator is nearly unbiased (head $+0.04$ m, extrapolation $+0.07$ m). Against the assertive neighbour the head reads separations $+0.29$ m too large and calls 56% of truly unsafe candidates safe: it predicts the yielding it was trained on, and the neighbour in front of it does not yield. The head trained in the control world, where no neighbour ever gives way, is the mirror image: poorly calibrated on yielders (1.16 m) and as good as extrapolation on the assertive neighbour ($+0.11$ m, 20.5%). A learned model of others encodes the population it learned from; its estimate is a bet on which population it is facing. Hedging is possible: the control-world head alone gives 14.4%, the extrapolation's rate ($p = 0.84$), and taking the smaller separation of the two heads gives 10.0%, the best any model-side estimate achieves, though not separable from extrapolation at this n ($p = 0.27$); that is what carrying one model per imaginable disposition buys. The declared intent, which needs no bet, gives 2.2%. This sharpens Section 5: pixels fail because they lose the number; a state head keeps the number and fails because it must guess the policy. Sharing the intent removes both problems.

7 Discussion

What we take from this. A shared world model in the NWM style learns where a broadcast neighbour will be and draws it; trained to output the neighbour's state as well, the same trunk also learns how the neighbour responds to the ego, which the pixel objective alone does not put into the representation. Neither reaches the planner through the image. Read directly, the model's estimate is accurate for the neighbours it was trained among and optimistic for the others, and a planner cannot tell which it faces. A pose broadcast has no blind spot and no prior, degrades gracefully with crowd size, and needs six bits (Appendix C). On this evidence, if a fleet is to share a world model of this kind, the sharing should happen in the space of poses and intents, and the world model's job is to render that shared state into each agent's view. Whether another carrier (depth, a bird's-eye view, a wider field of view) moves the balance, this instrument can now ask.

Limitations. The stage is synthetic: ray-cast walls and a coloured cylinder. It does what recorded data cannot (exact counterfactuals, a zero-interaction control), but the effect sizes are the simulator's. The model is small (CDiT-S, 112 px, 40k steps; CDiT-B in Appendix E). Separation is Euclidean and counts proximity through the wall, a small conservative bias equal across conditions. Neighbours follow the data-generating policy, never a planner.

Reproducibility statement

All simulator, training, evaluation and planning code is included as supplementary material with the exact commands and seeds used for every table and figure; the simulator is deterministic given a seed and needs no external assets. Appendix F lists architectures, training schedules, the matching procedure for the difference-in-differences, and the planner's candidate set and constraint.

AI use statement

Generative AI tools were used throughout this work: to draft and refactor the simulator, training, evaluation and planning code; to design and run experiments under the authors' direction; to perform statistical analyses; and to draft and edit the text. All code was tested by the authors (including bit-exact regression tests where existing results were reused), all numbers were regenerated from the authors' own runs, and all claims were checked by the authors against the underlying results. The authors take responsibility for the final content, including text, claims and artifacts produced with the aid of generative AI.

References

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ziqi Cai, Siqi Yang, Yimu Wang, Zixian Gao, Yunheng Liu, Shuchen Weng, Erwin Wu, Kaipeng Zhang, and Boxin Shi. MASS: Multiplayer world models with authoritative shared state. *arXiv preprint arXiv:2608.06257*, 2026.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. In *International Conference on Machine Learning (ICML)*, 2019.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51 (5):4282–4286, 1995.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Kunyu Peng, Zhikun Zhou, Kailun Yang, Di Wen, Ruiping Liu, Yufan Chen, Junwei Zheng, Hao Shi, Yi Zhou, M. Saquib Sarfraz, Danda Pani Paudel, and Luc Van Gool. Seeing together: Multi-robot cooperative egocentric spatial reasoning with multimodal large language models. *arXiv preprint arXiv:2605.18431*, 2026.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025.

Dhruv Shah, Benjamin Eysenbach, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. In *Conference on Robot Learning (CoRL)*, 2021.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics Research: The 14th International Symposium ISRR*, pp. 3–19. Springer, 2011.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Yilun Du, and Chuang Gan. COMBO: Compositional world models for embodied multi-agent cooperation. In *International Conference on Learning Representations (ICLR)*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A PLANNER AUDIT ON RECON

We reproduced NWM planning on RECON (ATE 1.21 m vs. 1.13 reported) and decomposed the error with $n = 11$ paired trajectories. Random sampling of 120 trajectories gives ATE 2.19; LPIPS selection 1.09; an oracle choosing the best of the same 120 candidates 0.53; the best constant-curvature arc 0.43. LPIPS captures 63% of the available within-pool gain and ranks candidates within a trajectory at $\rho = 0.61$ (best candidate's median rank 18.5/121) but does not correlate with ATE across trajectories ($r = +0.117$, $p = 0.25$). Spending the same 120 rollouts as multi-round CEM (24×5 or 12×10) does not help (Wilcoxon $p = 0.12, 0.90$). The action parameterisation is not the bottleneck; the scoring channel is.

B EGOTEAM: THE CHANNEL THAT SWITCHED ITSELF OFF

Data. EgoTeam (Peng et al., 2026) provides synchronised egocentric RGB-D video from teams of 2–4 robots in Habitat and iGibson scenes, with per-robot poses. We use the two-robot clips (107,010 anchor frames) and construct, for each anchor, the ego's context frames, the ego's future action, and the neighbour's current and future pose in the ego frame from the recorded trajectories, the same six numbers the simulator later broadcasts.

Precondition: does the model attend to its inputs at all? At 224 px and 16k steps the ego action itself had almost no effect on the prediction (perturbing it changed the output by 0.04–0.07$\times$ the change caused by resampling the diffusion noise); a NONE/FULL comparison at that stage would have been a guaranteed false negative. At 112 px, ±4 s horizons and 40k steps the ego-translation sensitivity reached 0.41 and short-horizon prediction first beat copying the last context frame. We treat a sensitivity of 0.4 as the threshold for running a channel comparison.

Sensitivity probe. Before comparing arms we verify the model attends to its inputs: holding context and noise fixed and perturbing one input, we measure output change relative to that caused by resampling the noise. On EgoTeam the ego translation scores 0.62 and the broadcast intent 0.01; in the simulator, 1.16 and 0.68. A model that does not react to a channel cannot be used to test that channel.

Filtered retraining. Suspecting dilution, we retrained both arms only on anchors where the neighbour was visible and within 4 m (22,377 anchors, 20.9%), to 40k steps each (final loss 0.1085 / 0.1082). On 1600 paired target frames FULL was significantly *worse* overall ($\Delta = +0.0130$, $p = 5\times10^{-9}$), with the largest degradation at 2–4 m ($+0.0161$) and none at 0–1 m ($+0.0041$, $p = 0.54$), the opposite of the gradient interaction predicts, and consistent with a constant-bias perturbation plus run-to-run variance.

Input dependence. For an embedding $e(\cdot)$ evaluated on 512 test inputs we report $\mathrm{Var}_x[e(x)]/\mathbb{E}_x[\|e(x)\|^2]$, the fraction of output energy that depends on the input. Table 6. The peer embedding's norm remained 0.315 that of the action embedding, with a standard deviation of 2% of its mean: a fixed vector the model absorbed as a bias. The channel's inputs were not degenerate (per-dimension s.d. 3.5–4.0; |corr| with ego action ≤ 0.06).

Table 6: Input dependence of the two conditioning embeddings on EgoTeam, CDiT-S/2.

	peer embedding	ego-action embedding
at initialisation	0.174	0.023
after 40k steps	0.0045	0.549

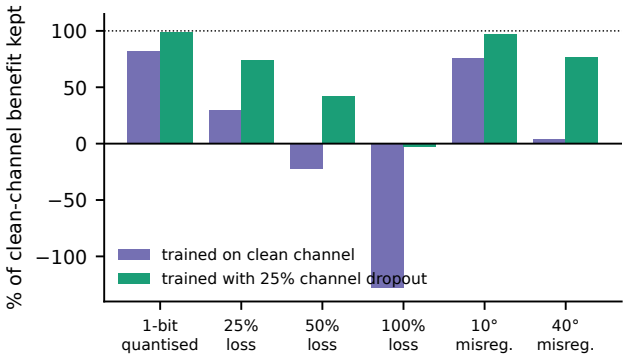

Figure 5: Channel degradation applied at test time to the trained FULL arm; bars show how much of the clean broadcast's LPIPS advantage over NONE survives (640 paired predictions). Below zero, the link is worse than no link. Training with 25% peer-channel dropout turns packet loss from catastrophic into graceful.

The state head on EgoTeam. We train the OBSERVED arm with the state head of Section 6 on the two-robot clips (same recipe, 40k steps) and ask it, on 2,397 held-out (anchor, goal) pairs with the goal 4–8 frames ahead, where the neighbour will be. Median error 0.35 m against 0.38 m for holding the neighbour still and 0.23 m for constant-velocity extrapolation from its last two poses (paired Wilcoxon: head better than holding, $p = 10^{-7}$; worse than extrapolation, $p < 10^{-80}$; the head wins on 30% of pairs). Within 2 m: 0.33 / 0.42 / 0.27 m. The head receives the neighbour's current pose and the frames but not its velocity; on recorded robots that mostly travel in straight lines, velocity is what predicts the future, and the head recovers only part of it from the frames. In the simulator the same head beat extrapolation because it learned the doorway behaviour (a neighbour that slows and yields), which is exactly the population-specific knowledge Section 6 shows to be a liability against a neighbour that does not.

The observational interaction test and its confound. We regress the neighbour's world-frame displacement over 8 steps on (i) its own momentum and current geometry and (ii) additionally the ego's future displacement, and report the incremental R^2 with a 100-permutation null (89,099 anchors). Incremental R^2 is 0.71% overall and rises to 1.56% within 1.5 m, a clean dose-response, all far above the permutation 95th percentile of 0.03%. We initially read this as weak-but-real interaction. The simulator's control world, in which an intervention on the ego moves the neighbour by exactly 0.0000 m, returns 0.73% from the same regression. Co-location in a shared room around shared obstacles produces this much correlation on its own, and the dose-response is equally explained by nearer pairs sharing more scene context. The observational number does not bound interaction from below.

C WHAT SHOULD THE CHANNEL CARRY?

Since sharing should happen at the level of poses, we ask what the link must provide, degrading the clean broadcast to the trained FULL arm without retraining and measuring the fraction of its LPIPS advantage over NONE that survives (640 paired predictions; Fig. 5, full tables in Appendix D). *Bandwidth is not the constraint*: quantising each of the six scalars to one bit retains 82%. *Packet loss is*: replacing the vector by zeros on 25% of steps retains 30%, on 50% it is worse than silence (−22%), and on 100% it is far worse (−128%): an all-zero peer vector never occurred in training and is out of distribution. Training FULL with 25% peer-channel dropout removes the pathology: 100% loss then costs −3% (the no-link floor), 25% loss retains 74%, at a 12% reduction in clean-channel benefit. *Registration* of the shared frame must be within 10° (76% retained; 4% at 40°). Staleness could not be measured on this data because neighbours in a doorway negotiation move a median 0.14 m over the horizon, and displacement is collinear with distance (Appendix D).

D COMMUNICATION SWEEP DETAILS

All impairments are applied at test time to the trained FULL arm on the interactive world; "kept" is $(LPIPS_{\text{impaired}} - LPIPS_{\text{NONE}})/(LPIPS_{\text{FULL}} - LPIPS_{\text{NONE}})$. Quantisation maps each of the six normalised scalars to 2^b uniform levels; loss replaces the whole vector with zeros on a fraction of anchors; registration rotates the neighbour's reported frame by a fixed angle; staleness substitutes the pose from k horizons earlier.

Table 7: Channel degradation. Left: 640 paired predictions, of which 384 with the neighbour at 1–2 m; the clean broadcast is worth -0.0136 LPIPS overall and -0.0253 at 1–2 m. Right: the arm trained with 25% peer-channel dropout (clean value -0.0119, -0.0186 at 1–2 m).

impairment	level	kept	kept, 1–2 m
quantise	4 bit	105%	104%
quantise	2 bit	89%	91%
quantise	1 bit	82%	84%
loss	25%	30%	46%
loss	50%	−22%	12%
loss	100%	−128%	−74%
registration	10°	76%	80%
registration	40°	4%	16%

impairment	level	kept	kept, 1–2 m
quantise	1 bit	99%	98%
loss	25%	74%	75%
loss	50%	42%	44%
loss	100%	−3%	1%
registration	10°	97%	96%
registration	40°	77%	77%

Staleness. Substituting a pose 0.25, 0.5 or 1.0 horizons old keeps 93%, 83% and 86% of the benefit (n=1600), which looks like robustness but is not a measurement: the neighbour travels a median 0.144 m over the horizon (90th percentile 0.867 m), so a stale pose is nearly the current one. Stratifying by travel gives 95/87/92% for < 0.3 m, and unstable values (268–537%, then -296 to -2131%) for the two sparse moving bands, where the denominator is near zero and displacement is collinear with distance. The doorway negotiation that makes the broadcast valuable is also what makes the neighbour stand still; staleness would need a world in which the broadcast matters while the neighbour moves.

E CAPACITY CHECK

We repeat the recipe with CDiT-B/2 (12 blocks, width 768, 12 heads; 195M parameters, $4\times$ CDiT-S), training the FULL and OBSERVED arms for the same 40k steps on the same data. Training loss at 40k steps is 0.0546 for CDiT-S FULL and 0.0474 for CDiT-B FULL: the larger model fits the data better.

Planning. In the two-agent doorway against the assertive neighbour, CDiT-B FULL collides 43.3% of the time with the hand readout and 53.3% with R1, against 46.7% and 40.0% for CDiT-S and 40.0% for BLIND on the same 30 seeds. The detection rate in imagined frames is unchanged (9.3% vs. 8.4%). Capacity does not move the planning result.

Anticipating reactions. We rerun the protocol of Section 6 on the same 1281 candidates with the FULL arm of each size (the declared-future channel is zeroed at scoring time, so the arm sees the neighbour's current pose only; the S-OBSERVED result of Section 6 is reproduced by S-FULL under this protocol). Table 8 gives the numbers. CDiT-S is at the placebo: 52.2% against 54.8% with the two distance strata weighted equally, paired Wilcoxon $p = 0.82$. CDiT-B is not: 54.7% against 50.7%, $p = 0.009$; it prefers the reacting future where the placebo does not in 243 candidates against 195 the other way, and its advantage over the placebo holds in both strata (46.2 vs. 41.3% when the reacting neighbour ends farther; 63.3 vs. 60.2% when nearer). The OBSERVED arm of the same size is weaker and mixed: balanced preference below the placebo, a paired p of 0.043 driven by margin sizes rather than by which candidates it wins (191 vs. 201). So a $4\times$ larger model may acquire a small sensitivity to how the neighbour's future depends on the ego's action, at most four points above the distance confound and not consistent across arms. It is not enough to plan with: the same model's imagined frames leave the planner at the blind floor above. The state head of Section 6 finds the dependence in the CDiT-S trunk directly, which locates the failure in the readout rather than in capacity.

Table 8: Anticipating reactions at two model sizes; protocol of Section 6, n=1281 candidates. "Balanced" weights the nearer/farther strata equally (bootstrap 95% CI); the paired test compares the arm's preference margin with a placebo whose candidate actions are permuted against the same target pairs.

model	arm	prefers reacting	balanced	placebo balanced	paired p
CDiT-S/2 (49M)	FULL	55.0%	52.2% [49.4, 55.1]	54.8% [51.9, 57.7]	0.82
CDiT-S/2 (49M)	OBSERVED	55.7%	53.1% [50.3, 55.9]	53.5% [50.6, 56.2]	0.41
CDiT-B/2 (195M)	FULL	57.2%	**54.7%** [52.0, 57.5]	50.7% [47.9, 53.5]	**0.009**
CDiT-B/2 (195M)	OBSERVED	55.5%	52.5% [49.7, 55.4]	54.0% [51.3, 56.9]	0.043

F EXPERIMENTAL DETAILS

Model. CDiT-S/2 (12 blocks, width 384, 6 heads; 49M parameters) on $4 \times 14 \times 14$ SD-VAE latents of 112 px frames, context of 4 frames at stride 1 (0.4 s apart), target offsets ± 8 frames. Conditioning: ego action (translation, yaw) and relative time through the NWM embedders; the peer channel through an MLP of the same design (six inputs), summed into the conditioning vector. Training: AdamW, lr 10^{-4}, batch 32 anchors $\times$ 4 goals, bf16, gradient clip 10, 40k steps (one pass over the simulated training split), seed 0 unless stated. The NONE arm is trained with the peer channel zeroed; the arms share initialisation, data order and noise.

Simulator. Scenes: hall of $W \in [12, 16]$ m by $H \in [10, 13]$ m, a dividing wall at $W/2$ with a gap of width $\in [0.95, 1.25]$ m at a random position, and 2–4 rectangular pillars; walls coloured per segment; floor a checkerboard. Agents: radius 0.25 m, cruise speed V_0, simulator step $\Delta t = 0.2$ s, heading relaxes toward velocity. Policy: goal attraction via a doorway-routed subgoal, wall repulsion within 0.75 m, and a peer term proportional to reactivity that acts within 2.2 m and pushes perpendicular to the relative velocity in proportion to the inverse time-to-closest-approach; reactivity 4 unless stated. Rendering: one ray per column, 90° horizontal FOV, eye height 1.15 m, neighbour a cylinder of height 1.5 m; the peer's silhouette is the nearest cylinder hit per column, in front of the wall hit. Training data: 400 episodes of 320 steps per world (interactive, control), both agents recorded as egos, frames saved every second step (0.4 s apart), 15% of episodes held out; the model's horizon of ± 8 frames is therefore ± 3.2 s. At planning time the model receives its context at that same 0.4 s spacing and horizons converted from simulator steps to frames.

Paired evaluation. 1600 (context, goal, target) triples from held-out episodes, 4 goals per anchor, one diffusion noise per triple shared across arms; LPIPS (AlexNet) on the decoded prediction against the true target frame; two-sided Wilcoxon signed-rank on per-triple differences; bootstrap CIs over triples (4000 resamples).

Matching for the difference-in-differences. Anchors are kept when the neighbour is at 1–2 m at decision time, inside the ego's 90° view, and displaces < 0.3 m over the horizon (interactive 1786, control 950 predictions). The DiD is the difference of per-world mean Δ; CIs by bootstrap over anchors within each world.

Reaction test. From each anchor, alternatives to the true ego action: veer $\pm 30^\circ$, half speed, stop, each held over the horizon. For each alternative the neighbour's true reaction is obtained by re-simulating from the anchor state; the non-reacting counterpart is the neighbour trajectory from the control world with the same seed. The model scores each target frame by its denoising loss averaged over 8 timesteps with shared noise; the placebo arm receives the same targets with actions permuted across anchors.

Planner. Candidates: headings $\{-90^\circ, -45^\circ, 0, 45^\circ, 90^\circ\}$ from the direction to the current subgoal at speeds $\{V_0, V_0/2\}$, plus waiting; held for 4 simulator steps (0.8 s) then replanned; lookahead 8 steps (1.6 s), separation evaluated at steps 4 and 8 (model horizons of 2 and 4 frames). Goal term: distance-to-go through the doorway, identical for all conditions. Constraint: minimum estimated separation ≥ 0.75 m (collision threshold 0.5 m); if no candidate is feasible, the one with the largest estimated separation is taken. Pixel readouts use margin 1.25 m. Episodes: 250 steps, seeds 0–29 for the pixel-readout conditions and 0–89 for the model-free and state-head conditions (the first 30

seeds are shared); a collision is any step with separation < 0.5 m; success is reaching within 0.5 m of the goal.

Hand detector. Pixels with red chromaticity above threshold form the neighbour mask; bearing from the mask's column centroid through the pinhole model; distance from the silhouette height when the mask touches neither frame edge, and from height alone (a lower bound on size, hence an upper bound on distance) when it does.

State head. LayerNorm–Linear($D\to D$)–GELU–Linear($D\to 2$) on the mean of the target-frame tokens after the last block (D=384; 0.3M parameters), predicting the neighbour's position at the target time in the ego frame at decision time, in the same normalised units as the ego action. Trained jointly from scratch with the diffusion loss, smooth-L1 with weight 1, same schedule as the other arms; the target is the unmasked future-position channel of the peer vector, so it is available to every arm. At inference the model is run once at the last diffusion timestep (t=999) with a Gaussian target latent and the head's output is read; the planner converts it to a separation against the candidate's own end pose. Arms: NONE (frames only), OBSERVED (frames and current pose); the control-world head is the OBSERVED arm trained on the zero-interaction data. Paired LPIPS of the head arms against their headless counterparts on the 1600 held-out targets: OBSERVED 0.275 vs. 0.227; NONE 0.473 vs. 0.235 (worse than copying the last context frame, 0.358); with no pose input, the frames are the head's only source of the target and it takes them. *Probe.* The frozen headless trunks of Section 4 (NONE, OBSERVED, and the control-world NONE) with only a head trained, smooth-L1, AdamW 3×10^{-4}, 4 epochs (12.6k steps). Mean pooling did not train on the frozen trunk (loss flat from step 0), so the probe keeps the tokens in place: LayerNorm–Flatten–Linear($196D\to 256$)–GELU–Linear($256\to 2$), 19M parameters against the joint head's 0.3M. Its converged loss (0.51 for OBSERVED, 0.83 for NONE, in normalised units) is 50$\times$ the joint head's (0.011). Sensitivity is evaluated with one Gaussian target latent shared across the alternatives of an anchor, so that the difference between alternatives is the action's; the probe's prediction then moves by a median 0.24 m between alternatives and the joint head's by 0.10 m, against a median true reaction of 0.12 m.

Interventional branches. For each of the 290 training episodes with seeds outside the reaction test's range, six branch points t_0 are drawn uniformly in simulator steps $[40, 250)$ and an alternative ego command (veer $\pm30°$, half speed, or stop, held from t_0 on) replaces the ego's; the episode is re-simulated from the same seed with the neighbour reacting, and the window of 8 saved frames before t_0 and 16 after is written as a short trajectory (1,740 branches, 41,760 frames, a 38% addition). The intervention displaces the neighbour by a median 0.13 m by the end of the window, above 0.1 m in 57% of branches. The test split is unchanged.

Learned readouts. A 5-layer strided CNN (channels 32–128, GroupNorm, GELU; 0.9M parameters) mapping a 112 px frame to a visibility logit and the nearest neighbour's position in the ego frame in metres; loss is binary cross-entropy plus smooth-L_1 on position where visible. R1: 80k true renders with 1–2 neighbours at log-uniform distances 0.3–8 m and uniform bearings, 12 epochs, AdamW 10^{-3}. R2: R1 fine-tuned (15 epochs, 3×10^{-4}) on frames imagined by each arm from 60 episodes disjoint from the evaluation seeds, half driven by the blind planner, half by a driver that approaches the neighbour to populate the near range, each labelled with the neighbour's true position at the imagined time.