EMBODIED XAI: TRACING HIERARCHICAL AGENT DECISIONS VIA 4D SPATIO-TEMPORAL SALIENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Robots steered by vision-language models are being proposed for spaces people occupy, where an unaccountable choice carries real cost. Yet when these physical AI agents silently pick a target, per-frame 2D saliency cannot say what in the physical world drove them: the planner is re-queried continuously while the robot drives, so each map sits in its own moving camera frame, leaving thousands of disconnected images that cannot be pooled, sliced, or compared. We present *4D spatiotemporal saliency*, a model-agnostic framework that relocates attribution from the camera to the scene: an attribution stage computes importance scores with respect to the planner's sampled steering token, each attributed pixel is lifted into world coordinates through depth and pose, and the results accumulate across queries and trials into one heat field growing through time. The explanation thereby becomes an object that can be added up, grouped by outcome, differenced against a reference, replayed, and physically intervened upon. We instantiate it with chosen-action integrated gradients on a quadruped driven by ambiguous instructions, and use the field to certify scene-level attribution patterns as repeatable, separate the signatures of success and failure, watch a decision harden, and run closed-loop occlusion experiments that persistently target the same physical region as the camera moves.

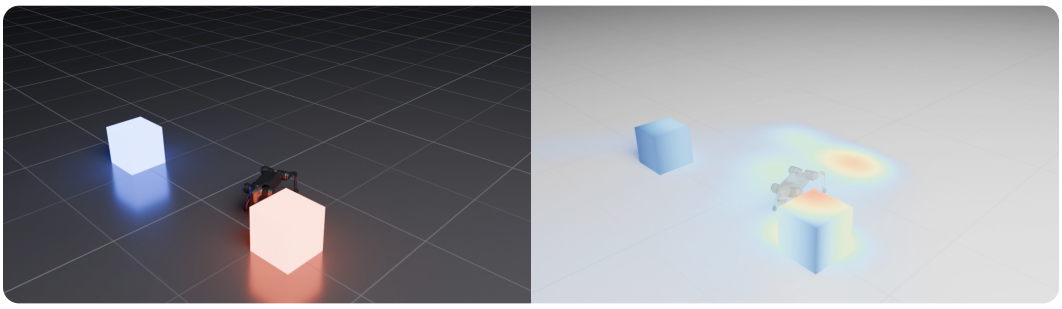
1 INTRODUCTION

Embodied AI systems — robots that perceive, reason, and act in the physical world — are increasingly proposed for deployment in shared human spaces. Much of their appeal comes from a hierarchical design pairing a general-purpose VLM for semantic reasoning with a specialized low-level controller for physical execution (Ahn et al., 2022; Driess et al., 2023; Brohan et al., 2023; Kim et al., 2024; Guaman Castro et al., 2025), combining the flexibility to follow open-ended instructions with reliable motor behavior. When deployed among people they make real-time decisions that affect them — which way to walk, which object to approach — and a robot that silently picks a target is a black box: its decisions need to be traceable back to what it actually looked at, so its behavior can be audited, debugged, and ultimately trusted (Jacovi & Goldberg, 2020; Gong et al., 2026). Explainable AI has matured around static, disembodied predictors, yet physical AI is the frontier that XAI has not yet caught up with.

The obstacle: explanation does not scale to embodied agents. The standard tool for attributing a vision model's output to its input is per-frame 2D saliency (Simonyan et al., 2013; Selvaraju et al., 2017; Sundararajan et al., 2017; Abnar & Zuidema, 2020), and it does not survive contact with a deployed physical agent. A robot re-queries its planner continuously; each query in a memory-equipped agent attributes over a *window* of frames; and every resulting map is anchored to a different, constantly moving camera pose, so nothing says whether the hot patch in frame 12 and the hot patch in frame 47 are the same physical spot. Auditing one condition therefore means eyeballing thousands of disconnected images with no shared frame of reference and no way to separate a repeatable influence pattern from single-trial noise (Adebayo et al., 2018); Appendix M shows the pile, and how it grows with a memory window. Per-frame saliency localizes to the *camera*; an auditor needs attribution localized to the *world*.

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.

Why it matters here: there is something worth explaining. Web-pretrained VLMs carry semantic associations — between colors and moods, or systematic spatial preferences — that can silently dominate behavior precisely when instructions are ambiguous, the regime open-ended language deployment invites (Liu et al., 2025; Ziliotto et al., 2026). An agent that resolves "go to the friendly one" by quietly always turning left is acting on a latent prior rather than on evidence, and no amount of outcome logging will say so.



(a) the decision as the world saw (b) as the planner's gradients saw

Figure 1: **From behavior to physically grounded explanation.** (a) A VLM-steered quadruped commits to a target under an ambiguous instruction. (b) The same trial with the accumulated 4D saliency of Section 5 projected back into the scene, localizing influence to the physical surfaces and moments that drove the decision. This is a full-fidelity path-traced render of the same heat-field data; the later diagnostic figures (Figures 4–8) use a simpler point-cloud style suited to pooling, slicing, and side-by-side comparison.

Our contributions, in order of centrality:

(1) 4D spatiotemporal saliency for embodied decisions. A three-stage framework that (i) scores the steering command the agent actually sampled while driving, (ii) unprojects each attributed pixel into world coordinates through depth and pose, and (iii) accumulates the result across queries and trials into one time-stamped volume over the arena (Figure 1). Stage (i) is a slot, not a fixture: stages (ii)–(iii) need only a per-pixel scalar map, so a decade of 2D saliency work transfers to embodied agents rather than being replaced by it (Appendices I and K).

(2) Four diagnostics that follow from world-anchoring. Fields in a shared frame are arithmetic objects, and each instrument is a different operation on them: *convergence sweeps* certifying a scene-level pattern as repeatable rather than single-trial noise; *outcome-split fields* isolating the influence signature of each choice and of failure; *time-resolved accumulation* making the moment of commitment watchable; and a *localization-noise sweep* asking how much of the reading survives once the camera's position is known only approximately. None has a per-frame analogue.

(3) An experiment world-anchoring makes definable at all: closed-loop occlusion auditing. An image-coordinate mask cannot follow a physical region as the camera moves, so occlusion-auditing a driving agent is not merely harder than for a static image — it is undefined. A hot set made of *world voxels* can instead be re-projected into every frame of a *fresh* rollout, keeping the same surfaces covered while the robot drives. Section 6 runs the experiment with area- and fill-matched controls in both mirrored arrangements, and reports what it measured.

Scope of the claim. We do not propose a new definition of pixel importance. The framework separates two questions that per-frame work answers together: whether a pixel attribution is faithful to a planner output, and whether those attributions can be organized into a persistent object that supports embodied audit operations (Bekkemoen, 2024; Atrey et al., 2020). We address the second directly and probe the first through chosen-action targeting, cross-trial reproducibility, alternative attributors, and closed-loop intervention, without claiming that any single diagnostic is a complete test of faithfulness (Jacovi & Goldberg, 2020). Pairing a behavioral probe that establishes *that* something latent is steering an agent with a spatiotemporal field that localizes *where* and *when* it took hold is, we argue, the minimum transparency standard for a VLM-controlled robot operating around people.

2 RELATED WORK

Attribution and saliency, from 2D to 3D. Gradient-based attribution (Simonyan et al., 2013; Selvaraju et al., 2017), integrated gradients (Sundararajan et al., 2017), and attention rollout (Abnar & Zuidema, 2020) explain a single output in pixel space, as does saliency for deep RL policies (Greydanus et al., 2018); our framework treats any such attributor as an interchangeable front end. Validation comes from complementary directions: perturbation asks how outputs respond to a specified input intervention (Fong & Vedaldi, 2017; Hooker et al., 2019), while cross-example aggregation asks whether a recurring strategy exists at the population level (Lapuschkin et al., 2019). Both are posed over static images, and in sequential control saliency is best used to generate and organize testable hypotheses about agent behavior rather than as a complete causal account (Atrey et al., 2020); Section 6 accordingly re-poses deletion in *closed loop*, where masking an observation changes the observations the agent later receives. The closest scene-anchored saliency is in graphics: Sitzmann et al. (2018) aggregate human gaze onto a viewing sphere for static panoramas — a projection, not unprojection into a world frame, and with no moving camera. We contribute that missing step, and with it the shared scene representation needed to pose embodied hypotheses and interventions consistently across time and rollouts: attribution on the planner's own sampled action token, unprojection via depth and pose, and accumulation into a convergent per-condition heat field.

Hierarchical VLM/VLA control. Grounding web-scale semantics in physical action has progressed from language-conditioned affordance ranking (Ahn et al., 2022) and embodied multimodal LLMs (Driess et al., 2023) to VLA models that emit actions directly (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024), with a recurring pattern decoupling a slow semantic planner from a fast embodiment-specific policy (Tang et al., 2023; Guaman Castro et al., 2025; Chen et al., 2026). Our agent follows this two-timescale pattern but deliberately uses a *continuous-command* interface to expose a numeric action token to attribute against.

Behavioral evaluation and latent bias in embodied VLMs. Controlled 3D probes are emerging as diagnostics of what physical AI agents ground decisions in: mirror self-recognition benchmarks separate perception-rooted behavior from priors and confabulation (Ziliotto et al., 2026), and RoboView-Bias documents systematic visual biases in VLM-driven manipulation (Liu et al., 2025). Our mirrored two-target design with a same-color control extends this line to *navigation-level target selection*, and is, to our knowledge, the first to statistically decompose positional from color bias in a closed-loop VLM-steered robot.

3 TESTBED: A HIERARCHICAL VLM–PPO AGENT WITH HYBRID MEMORY

Figure 2 summarizes the agent: a slow semantic loop and a fast physical loop, with the attribution tap of Section 5 drawn in place. All experiments run a simulated Unitree Go2 quadruped inside the Genesis physics engine (Genesis Authors, 2024); Figure 3 shows the arena and the planner's onboard view. All hyperparameters are in Appendix B.

Low-level controller (PPO locomotion policy). A gait controller is trained from scratch with PPO (Schulman et al., 2017) via `rsl-rl` (Rudin et al., 2022) across 4,096 parallel Genesis environments, purely on randomly resampled velocity commands (v_x, v_y, ω_z): it never sees the VLM, the camera, or the task, and its only job is to track whatever continuous command it is given at 50 Hz (reward terms in Appendix B).

High-level planner (frozen Qwen2.5-VL-3B). The planner (Bai et al., 2025), queried at 3 Hz, reads the forward-facing camera and emits a JSON contract (Appendix A) with its believed target, a one-sentence rationale, and — critically — a *continuous* (v_x, v_y, ω_z) command rather than a discrete action label, low-pass filtered ($\alpha = 0.35$) into the controller's command input: the VLM steers the robot like a slow joystick. Real-valued velocities let the planner reason about *what* while the controller handles *how*, and, decisively for Section 5, make every steering decision a concrete numeric token span attribution can target.

Hybrid memory (adapted MEM). Without memory the agent cannot remember which target it committed to or integrate motion cues across time. We adapt Multi-Scale Embodied Memory (Torne et al., 2026) to a *frozen* off-the-shelf VLM (the original needs a custom-trained video encoder and a

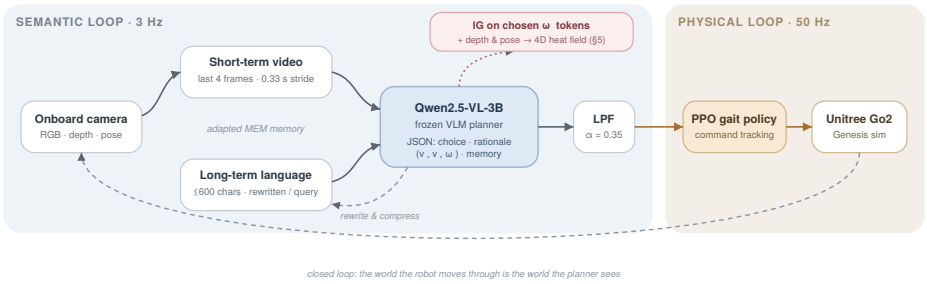
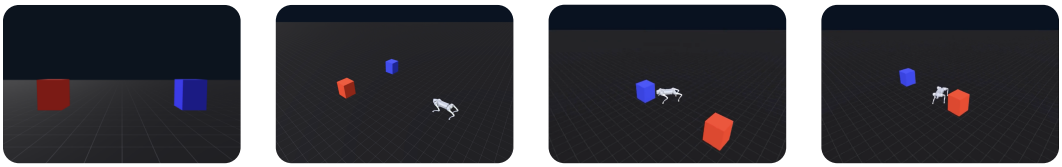


Figure 2: **System architecture.** The 3 Hz semantic loop, the 50 Hz physical loop, and the attribution tap of Section 5 (dashed, red).

learned rewriting policy, neither feasible here): *short-term*, the last 4 onboard frames ∼0.33 s apart are fed to Qwen as one native video clip; *long-term*, a running summary is *rewritten and compressed by the VLM itself* every query, hard-capped at 600 characters so it neither dominates the prompt nor drifts. Every query logs the planner's exact inputs — frame window and summary snapshot — alongside the sampled command tokens, which is what lets Section 5 attribute each decision over precisely the evidence that produced it.

4 The Behavioral Probe: What Needs Explaining

Before an explanation method is worth anything, there must be a decision worth explaining. We generate one by making the instruction genuinely ambiguous — "go to the friendly one," with no stated color mapping — and asking whether the agent's choice reflects a *color* bias (red or blue regardless of side) or a *positional* bias (left or right regardless of color). The arena is deliberately bare: not a navigation benchmark but a calibrated probe — analogous to a synthetic psycholinguistic stimulus — that isolates latent priors from environmental clutter (Liu et al., 2025). Separating the two biases requires two designs. **Mirror testing**: every semantic condition runs in both the (L=red, R=blue) and (L=blue, R=red) arrangements, so pooling disentangles color-tracking from side-tracking — without the mirror, "always picks red" and "always picks left" are indistinguishable whenever red happens to be on the left (Ziliotto et al., 2026). **A same-color control**: setting both targets red removes the color cue entirely, isolating pure positional bias. Pooled, the two signatures differ: a RED rate far from 50% with LEFT near 50% is a color-driven choice; the reverse is a side-driven one.



(a) POV at reset (b) arena at reset (c) end at blue (d) end at red

Figure 3: **Task environment.** (a) The agent's onboard camera at reset — exactly what the planner sees: two candidate cubes (here L=red, R=blue) 3 m ahead at ±1.2 m lateral offset, no other scene cues.

Protocol. We run N=50 trials per category — 300 total across 4 semantic and 2 control categories — each with a 6 s timeout and a 0.7 m commit radius. Trials that never enter a commit radius are labeled *drift* and conservatively scored against whichever target the agent's final camera position was nearest, so no trial is discarded. Choice rates carry Wilson 95% confidence intervals (Wilson, 1927) and exact binomial tests against chance.

Results. The dominant effect is a strong *positional (left) bias*: in the color-neutral control the agent chose the left target 87% of the time (95% CI [79, 92], $p \approx 1.3 \times 10^{-14}$) — a genuine positional

Table 1: **Mirror decomposition and control baseline.** Semantic conditions pooled across L/R arrangements (n=100 each); p-values are exact binomial tests vs. chance.

Condition	n	RED% [95% CI]	p_{RED}	LEFT% [95% CI]	p_{LEFT}
"Go to the friendly one."	100	30% [22, 40]	7.9×10^{-5}	74% [65, 82]	1.7×10^{-6}
"Go to the hostile one."	100	50% [40, 60]	1.0	58% [48, 67]	0.133
Control (identical targets)	100	—	—	87% [79, 92]	1.3×10^{-14}

artifact no color-only audit would reveal. For "go to the friendly one," the mirror decomposition shows a large LEFT effect (74%, $p \approx 1.7 \times 10^{-6}$) *alongside* a real color effect toward blue (RED $= 30\%$, $p \approx 7.9 \times 10^{-5}$); "go to the hostile one" shows no significant color effect and a smaller, non-significant positional lean. A chi-squared test confirms the prompt *wording itself* shifts color choice ($\chi^2 = 7.52$, $p = 0.006$; contingency [red, blue] = [[30, 70], [50, 50]]) on pixel-identical images. Two co-mingled biases are thus separated — the dominant one positional, with a significant secondary blue-association for "friendly" (un-pooled per-category outcomes in Appendix C). What neither table can say is *where in the scene*, or *when in the trial*, those biases were realized. That is the rest of the paper.

5 4D SPATIOTEMPORAL SALIENCY VIA CHOSEN-ACTION INTEGRATED GRADIENTS

5.1 ATTRIBUTING THE DECISION THE AGENT ACTUALLY MADE

Integrated Gradients (Sundararajan et al., 2017) attributes a scalar model output F to input pixels by integrating gradients along a straight-line path from a baseline x' (all-black here) to the real input x:

$$\text{IG}_i(x) \;=\; (x_i - x'_i) \int_0^1 \frac{\partial F(x' + \alpha\,(x - x'))}{\partial x_i}\, d\alpha, \tag{1}$$

approximated with an 8-step Riemann sum. Pixels whose gradient consistently pushes the output in the same direction across the interpolation receive high attribution; pixels the model never used receive near-zero attribution. Gradients enter here and nowhere else: later stages consume only a per-pixel score, so a perturbation attributor such as RISE (Petsiuk et al., 2018) drops in unchanged and carries the framework to API-only planners (Appendix I).

The target token. Rather than attributing to a generic "steering" concept, IG is run against the *exact sign-inclusive token span* the model generated for the yaw-rate command ω_z it actually chose while driving: that span is located in the model's own sampled output and teacher-forced back through the model, and its summed log-probability is the scalar F in Eq. 1 (tokenization in Appendix D). Including the sign is essential — attributing only the leading digit makes $+0.5$ and -0.5, turn left versus turn right, indistinguishable. This is a deliberate faithfulness choice (Jacovi & Goldberg, 2020): it is *not* a fresh greedy re-decode, so every saliency map explains a real steering action the agent committed to.

5.2 FROM A 2D PIXEL MAP TO A 4D (3D + TIME) HEAT FIELD

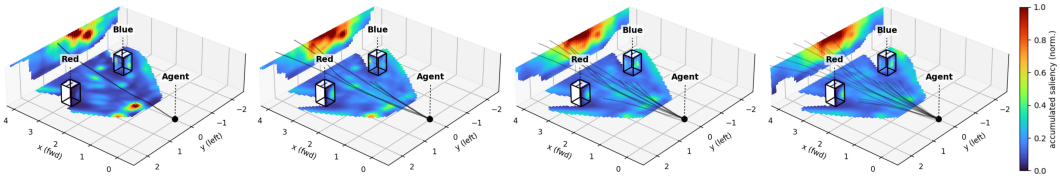
Run per frame, Section 5.1 hands back a stack of 2D heatmaps, each anchored to its own camera pose. One straightforward but consequential unprojection bridges the gap. For each attributed pixel (u, v) with saliency weight s_{uv}, the simulator's depth map d_{uv} and the camera's known intrinsics K and pose (R, t) cast a ray from the camera center through the pixel and place the saliency mass at its 3D world-space hit point:

$$\mathbf{p}_{uv} \;=\; R\, d_{uv}\, K^{-1}[u, v, 1]^\top + t. \tag{2}$$

Pixels landing on the floor or a target cube become weighted 3D points on that surface; pixels with no finite depth (horizon/sky) are intersected with a back-wall plane rather than discarded, preserving

the planner's reliance on background context and with it attribution completeness (Sundararajan et al., 2017) instead of silently dropping valid gradient mass. In the multi-frame modes, *each of the four window frames is unprojected with its own camera pose and depth*, since the robot moved between frames. Repeating this at every sampled query and accumulating the point clouds over the trial turns a stack of 2D maps into a genuinely 4D object: a heat field over the arena that builds up over time, faint where the agent glanced once and bright where it fixated repeatedly.

Every attribution from every query of every trial now lands in one shared frame. That single property is what the four instruments of Sections 5.4–6 are built on; nothing below requires anything more of the attribution stage.



(a) n=1 (b) n=5 (c) n=10 (d) n=20

Figure 4: **Convergence of the accumulated 4D heat field** over pooled trials (L=red, R=blue, "friendly"). Grey lines are the trajectory fan.

5.3 Theoretical properties of the audit framework

The shared world frame admits statistical and geometric guarantees, while a query-level model clarifies the interpretation of mirror diagnostics.

Proposition 1 (Concentration of pooled fields). For one condition, let $Z_1, \dots, Z_n \in [0, B]^M$ be independent, identically distributed trial fields on M fixed bins, after fixed bounded preprocessing. For $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \right\|_\infty \leq B \sqrt{\frac{\log(2M/\delta)}{2n}}. \tag{3}$$

The sampling unit is a complete trial; within-trial frames may be dependent. This bounds a quantitative trial average, with display rescaling treated separately. Appendix N.2 gives the proof and extensions to regional summaries and outcome groups.

Proposition 2 (Geometric stability). Let ν be a fixed, normalized nonnegative attribution measure on retained pixel records, and let $U, \widehat{U}$ be valid world-space projection maps whose pushforwards have finite first moments. Then

$$W_1(U_\# \nu, \widehat{U}_\# \nu) \leq \int \|U(p) - \widehat{U}(p)\|_2 \, \nu(dp), \tag{4}$$

where $U_\# \nu$ denotes the projected attribution distribution and W_1 is Wasserstein distance. Appendix N.3 bounds the displacement for finite-depth points and non-grazing rays intersecting a known wall, and accounts for changed valid projections. The bound also controls centre-of-mass displacement; stability of thresholded hot sets requires additional conditions.

Proposition 3 (Mirror calibration). At planner-query times, let $\varepsilon_{\text{phys}}$ and $\varepsilon_{\text{plan}}$ bound the uniform total-variation mirror defects of the physical loop and the joint command–memory output, respectively, as defined in Appendix N.4. For reflected initial states $s_0, M_s s_0$, horizon H, and terminal event E,

$$|\Pr(S_H \in E \mid s_0) - \Pr(S_H \in M_s E \mid M_s s_0)| \leq \min\{1, H(\varepsilon_{\text{phys}} + \varepsilon_{\text{plan}})\}, \tag{5}$$

where M_s reflects the augmented state, including directional memory. The controller audits in Appendix J support interpretation on scripted commands; they do not estimate these uniform defects. Appendix N.5 further separates planning and execution contributions to an outcome gap. These guarantees concern sampling, localization, and spatial symmetry; they establish neither semantic correctness nor causal faithfulness of the attribution method.

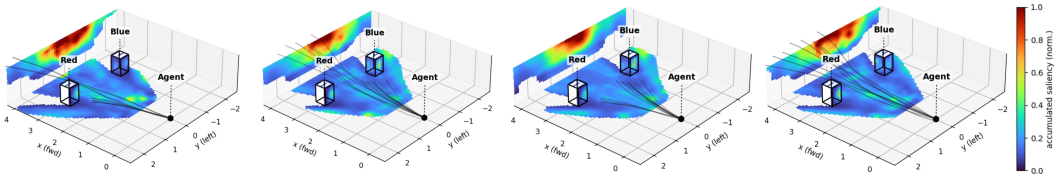
5.4 DIAGNOSTICS THE HEAT FIELD ENABLES

Each condition keeps a shared 20-trial drive cache — trajectory, frames, depth, poses, chosen tokens, memory snapshots — so the attribution pass of Sections 5.1–5.2 can be rerun and re-sliced arbitrarily without re-driving the robot.

Convergence sweeps: is the pattern real? Accumulating the field at $n = 1, 5, 10, 20$ pooled trials (Figure 4) tests cross-trial reproducibility: whether the world-space pattern recurs across trials rather than reflecting one observation sequence. The n=1 field is idiosyncratic — scattered hot patches an auditor could over-read; by n=20 it has stabilized into sustained heat around the target region, the approach corridor traced by the trajectory fan, and a persistent back-wall lobe on the favored side. A background lobe is not, by itself, evidence of attribution failure: driving policies may read heading, bearing, and free space off the horizon line or other non-object geometry. But explanation methods have also exposed background reliance that is purely spurious (Lapuschkin et al., 2019), and the map alone does not say which case this is; here the question is whether that background dependence is repeatable, behaviorally associated, and robust to intervention.

Does world-anchoring survive imperfect localization? Eq. 2 consumes pose and depth, which the simulator supplies exactly and a real robot would estimate. Holding the attributions of the n=20 clean drives fixed and re-placing them under injected pose, yaw, and depth error separates the field from the reading taken off it. The field degrades gently — cosine $0.99/0.98/0.98$ and hot-set IoU $0.93/0.91/0.88$ at $2/5/10$ cm of drift, surface voxels faster than projected horizon walls (Appendix G). The *reading* does not degrade at all: the left/right saliency mass ratio that Sections 5 and 6 rest on holds between 0.73 and 0.75 in every cell (noise-free 0.743), and the fields look identical throughout (Figure 16). Aggregation buys this — pooling 20 trials averages independent pose errors out before they can distort the audit — but it is a robustness bound, not a real-robot validation.

Outcome-split fields: why this target, and why did it fail? Because every trial's attribution lands in the same world frame, one cached condition can be sliced *post hoc* by outcome — by choice (nearer-red vs. nearer-blue) and by success (reached vs. drifted) — yielding one field per group from the same drives (Figure 5). This turns saliency into a choice-audit instrument: the majority nearer-blue group — consistent with the "friendly" blue association of Table 1 — concentrates heat on the blue cube and approach corridor while the minority nearer-red group localizes around the opposite cube, exposing which part of the environment grounded each choice. The success slice isolates failure's population signature: reached trials show a tight fan with heat consolidated on the committed target, drifted trials a visibly broader, less committed field — which the next subsection develops at the single-trial level.

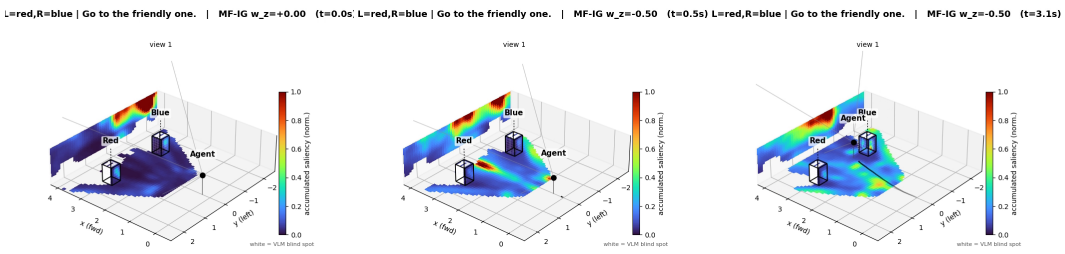


(a) nearer red (5) (b) nearer blue (15) (c) reached (12) (d) drifted (8)

Figure 5: **Outcome-split accumulated saliency**: the same N=20 cached drives re-sliced by choice (a, b) and by success (c, d), group sizes in parentheses (L=red, R=blue, "friendly").

5.5 WATCHING INFLUENCE EVOLVE THROUGH TIME

Static fields show where attribution *ended up*, not how it got there. The framework's final deliverable is a rendered video of the same 4D accumulation per condition, in which the heat field grows frame-by-frame exactly as computed. Figure 6 snapshots a representative reached trial at $t = 0.0, 0.5, 3.1$ s, the last its commit moment: attribution consolidates onto the blue side and its traversed corridor exactly as the sampled commands settle, so by commitment influence and action agree. The matched *drifted* trial, sampled at the same instants, is Figure 20 in Appendix H.



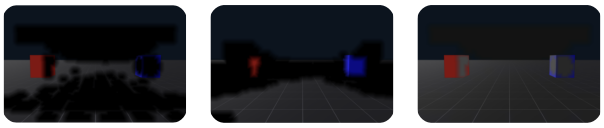
(a) $t = 0.0\,\text{s}$, $w_z = 0$ (b) $t = 0.5\,\text{s}$, $w_z = -0.5$ (c) $t = 3.1\,\text{s}$ (end), $w_z = -0.5$

Figure 6: **Time-resolved accumulation: the moment of commitment** (one reached trial; L=red, R=blue, "go to the friendly one"). Token sequences in Appendix D; the matched drifted trial is Figure 20.

Diagnosing individual trials. Because every map is anchored to a chosen token, each trial also carries an ordered record of its attributed ω_z spans (Appendix D), so field and commands can be read together. The reached trial settles in both channels at once: 2 direction reversals in 10 samples, and an attribution that stays on the same world region as the camera moves, compact around the target and its corridor by commit. The matched drifted trial fails in both: 5 reversals across 24 samples, the first inside the opening half second, and an attribution that never consolidates, spanning roughly twice the hot area at the moment it should have committed.

6 CLOSED-LOOP OCCLUSION AUDITING

Represented as world voxels rather than image pixels, a selected region can be re-projected into every frame of a *fresh* rollout, so the same physical surfaces stay covered while the camera and the trajectory change. That makes the intervention well posed; it does not make it a proof. Perturbation tests are not complete tests of attribution faithfulness: their conclusions depend on the replacement operator, can introduce artifacts or distribution shift, and can understate importance when multiple cues are redundant (Fong & Vedaldi, 2017; Hooker et al., 2019). Our goal here is therefore narrower and specifically embodied.

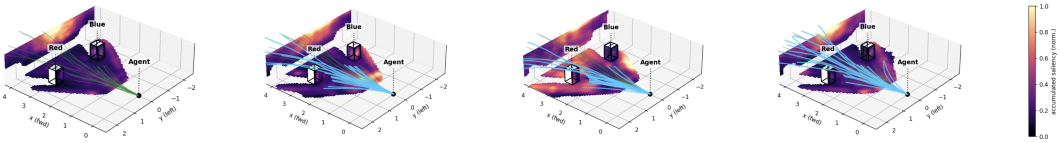


(a) BLACK-HOT (b) BLACK-COLD (c) GRAY-HOT

Figure 7: The three interventions, shown on one clean frame (p=0.70). Per-frame sanity renders and threshold calibration are in Appendix E.

Setup. We threshold a fresh n=20 clean field at p=0.70 and take the resulting world voxel set as HOT. Two controls separate what a mask confounds (Figure 7): GRAY-HOT fills the same voxels mid-gray, matching region and area but not contrast; BLACK-COLD fills a region elsewhere that calibration expanded into hot's complement. Each arm re-drives N=20 fresh rollouts on seeds paired to its own panel's clean arm, mask re-projected online at every frame, and the probe runs twice, once per mirrored arrangement, so an effect of *where* the saliency was cannot be read as an effect of *which side* it was on. Table 6 in Appendix E reports all eight arms as measured.

Two things worth noticing. First, attribution *relocates* rather than simply fading: after a region is deleted, the fresh closed-loop rollout places saliency mass on alternative evidence (Figure 8). This points to redundancy rather than a flawed original attribution: if two cues carry the same information, deleting one need not change the output—which is also why deletion should be tested at several budgets and against controls (Hooker et al., 2019). Occlusion in BLACK-HOT and GRAY-HOT directs mass toward the near-field ground structure between the agent and targets, whereas masking the complementary COLD region yields the least structured field of the four.



(a) clean (reference) (b) BLACK-HOT (c) BLACK-COLD (d) GRAY-HOT

Figure 8: **Re-accumulated fields under closed-loop occlusion** (L=red, R=blue; n=20 fresh drives per arm, sharing no rollout with Figures 4 or 5). Green = clean trajectories, light blue = the same seeds re-driven under that arm's occlusion. Metrics in Table 6A; mirrored counterpart in Figure 11.

Second, in both arrangements the arm whose field is least like its own clean reference is also the arm whose behavior is most one-sided with highest reach. We do not interpret improved reach under occlusion as improved semantic reasoning or as a contradiction of saliency. Rather, the intervention changes the evidence available to a closed-loop policy and can expose a simpler fallback strategy: in this task, removing candidate-region information reduces dithering under an ambiguous instruction and leaves the positional prior measured in Table 1. At n=20 per arm, the result is exploratory rather than definitive; its central methodological point is that physically corresponding regions and resulting fields can now be intervened on and compared in one shared frame.

7 LIMITATIONS

Simulation. All results use the simulator's exact depth and pose. Appendix G bounds the cost of giving that up, but the bound assumes a known wall plane, and a real port would also face depth dropout, rolling shutter, and calibration error. **Attribution configuration.** IG with a black baseline and 8 Riemann steps is pragmatic; four other attributors are in Appendix I. **Coverage.** While our application focuses on the primary setup, evaluating across diverse architectures and environments is critical: we include low-level controller audits in Appendix J, 4D fields for three further planners in Appendices I and K, and a new task in a dynamic arena whose target itself moves in Appendix L.

8 CONCLUSION

An explanation tied to the camera dies with the frame that held it. Tying it to the scene leaves an auditor holding one accumulating quantity in the arena rather than a pile of pictures to page through. On our quadruped, that quantity traced two co-mingled biases back to particular surfaces and seconds, certified the pattern as repeatable rather than single-trial, and gave failure a look of its own; the occlusion arms then lifted a hot region out of a driving robot's view — possible only because it is a place and not a patch of image — and showed how the closed-loop policy reorganizes under missing evidence. Larger probes and a hardware port are the next steps. Auditing an embodied agent means asking what it used and not only what it did — a question no representation can answer at scale unless it can be aggregated.

Reproducibility statement

Our implementation — simulation environment, planner, attribution engine, and experiments here — is submitted as executable notebooks in the supplementary material. The master configuration fixing every hyperparameter is in Appendix B, the planner's output contract in Appendix A, and the occlusion mask construction and threshold calibration in Appendix E.

Use of AI statement

We used generative AI tools to get feedback on experiments, implement methods, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat datasets, or assist with translation; formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. We take responsibility for the final content of this work.

Ethics statement

This work is an interpretability and bias-auditing study of an embodied agent in simulation; no human subjects or personal data are involved. Surfacing latent biases in VLM-steered robots is intended to support safer deployment. We note the dual-use consideration that saliency tooling could be used to optimize deceptive behavior; we believe the transparency benefits to auditors outweigh this risk.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

Akanksha Atrey, Kaleigh Clary, and David Jensen. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yanzhe Bekkemoen. Explainable reinforcement learning (XRL): A systematic literature review and taxonomy. *Machine Learning*, 113:355–441, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

William Chen, Jagdeep Bhatia, Catherine Glossop, Nikhil Mathihalli, Ria Doshi, Andy Tang, Danny Driess, Karl Pertsch, and Sergey Levine. Steerable vision-language-action policies for embodied reasoning and hierarchical control. *arXiv preprint arXiv:2602.13193*, 2026.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond. `https://github.com/Genesis-Embodied-AI/Genesis`, December 2024.

Shizhan Gong, Minda Hu, Qiyuan Zhang, Chen Ma, and Qi Dou. Saliency-R1: Enforcing interpretable and faithful vision-language reasoning via saliency-map alignment reward. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Google DeepMind. Gemini 3.5 flash-lite model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/`, 2026.

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding Atari agents. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Mateo Guaman Castro, Sidharth Rajagopal, Daniel Gorbatov, Matt Schmittle, Rohan Baijal, Octi Zhang, Rosario Scalise, Sidharth Talia, Emma Romig, Celso de Melo, Byron Boots, and Abhishek Gupta. VAMOS: A hierarchical vision-language-action model for capability-modulated and steerable navigation. *arXiv preprint arXiv:2510.20818*, 2025.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

Enguang Liu, Siyuan Liang, Liming Lu, Xiyu Zeng, Xiaochun Cao, Aishan Liu, and Shuchao Pang. RoboView-Bias: Benchmarking visual bias in embodied agents for robotic manipulation. *arXiv preprint arXiv:2509.22356*, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Daniela Rus. LIO-SAM: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Yujin Tang, Wenhao Yu, Jie Tan, Heiga Zen, Aleksandra Faust, and Tatsuya Harada. Saytap: Language to quadrupedal locomotion. In *Conference on Robot Learning (CoRL)*, 2023.

Marcel Torne, Karl Pertsch, Homer Walke, Kyle Vedder, Suraj Nair, Brian Ichter, Allen Z. Ren, Haohuan Wang, Jiaming Tang, Kyle Stachowicz, Karan Dhabalia, Michael Equi, Quan Vuong, Jost Tobias Springenberg, Sergey Levine, Chelsea Finn, and Danny Driess. MEM: Multi-scale embodied memory for vision language action models. *arXiv preprint arXiv:2603.03596*, 2026.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. FAST-LIO2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Filippo Ziliotto, Ciro Beneduce, Bruno Lepri, Luciano Serafini, Massimiliano Luca, and Tommaso Campari. Mirror, mirror on the wall: Can VLM agents tell who they are at all? *arXiv preprint arXiv:2605.08816*, 2026.

A Planner output contract

The planner is instructed to read the onboard clip and long-term summary, commit early to a single target, avoid flip-flopping, and respond with exactly one JSON object:

```
{"reasoning": "<one short sentence>", "choice":
"left|right|none",
"v_x": <float>, "v_y": <float>, "w_z": <float>, "memory":
"<updated long-term summary>"}
```

The prompt specifies the robot body-frame coordinate convention (positive ω_z turns left, positive v_y strafes left, positive v_x drives forward), gives sign-mapping examples for targets left/center/right of frame, and mandates the memory-rewrite rules of Section 3. The full contract text is included with the code release.

B Hyperparameters and configuration

PPO reward terms (all scaled by Δt): exponentiated tracking of commanded planar linear velocity and yaw rate ($\exp(-\mathrm{err}/\sigma)$), penalties on vertical base velocity (bounce), action rate (jerk), joint deviation from the default stance (posture), and squared base-height error.

Table 2: Master configuration used for all experiments.

Simulation / PPO		**VLM planner**	
Physics timestep	0.02 s (50 Hz)	Model	Qwen2.5-VL-3B-Instruct (frozen)
Parallel envs	4,096	Query rate	3 Hz
PPO iterations	600	Sampling temperature	0.7
Action scale	0.25	Max new tokens	192
Joint gains (k_p, k_d)	20.0, 0.5	Command low-pass α	0.35
Termination roll/pitch	10°	v_x range	$[-1.0, 2.5]$ m/s
Episode length	20 s	v_y range	$[-0.8, 0.8]$ m/s
Command resample	4 s	ω_z range	$[-1.5, 1.5]$ rad/s
Arena / evaluation		**Memory (adapted MEM)**	
Target distance	3.0 m	Short-term window	4 frames
Lateral offset	$\pm$1.2 m	Frame stride	$\sim$0.333 s
Target box	$0.4\times0.4\times0.5$ m	Input mode	native Qwen video
Commit radius	0.7 m	Long-term summary cap	600 chars
Trial timeout	6 s (8 s for IG video)	Rewrite policy	replace, compress
Trials per category	50 (batched $\times$6)	Per-query logging	frames, summary, tokens
Attribution (IG)		**Occlusion probe (§6)**	
Riemann steps	8	Threshold quantile p	0.70
Baseline	all-black	Mask domain	world voxels, re-projected
Attributed component	ω_z (sign-inclusive)	Fills	black, mid-gray
Frame long-side cap	392 px ($\div$28)	Trials per arm	20 (fresh drives)
Query subsampling	3rd query (static) / every (video)	Arms	clean / hot $\times$ fill / black-cold
Normalization	per-query, robust q=0.99	Arrangement	L=red, R=blue / mirror
Attribution trials per condition	20 (shared cache)	Convergence checkpoints	$n \in \{1, 5, 10, 20\}$

C PER-CATEGORY BEHAVIORAL OUTCOMES

Table 3 gives the un-pooled breakdown behind the mirror decomposition of Table 1. Reading the two mirrored "friendly" rows together is what separates the effects: choice tracks the left side across both arrangements (94% and 54% left) rather than tracking red, while the same-color control rows isolate the positional effect with no color cue present at all.

Table 3: **Per-category outcomes** (300 trials; N=50 per category). *Reach* is the fraction entering a commit radius before timeout; brackets are Wilson 95% CIs.

Arrangement	Instruction	n	Reach	Red% [95% CI]	Left% [95% CI]
L=blue, R=red	friendly	50	66%	6% [2, 16]	94% [84, 98]
L=blue, R=red	hostile	50	54%	42% [29, 56]	58% [44, 71]
L=red, R=blue	friendly	50	62%	54% [40, 67]	54% [40, 67]
L=red, R=blue	hostile	50	46%	58% [44, 71]	58% [44, 71]
L=red, R=red (control)	friendly	50	78%	100% [93, 100]	96% [87, 99]
L=red, R=red (control)	hostile	50	64%	100% [93, 100]	78% [65, 87]

D ATTRIBUTED ω_z TOKEN SEQUENCES FOR THE TRIALS OF FIGURES 6 AND 20

Table 4: Chosen ω_z token spans for the matched reached and drifted trials, in attribution order (both L=red, R=blue, "friendly"; positive = turn left).

Trial	Attributed ω_z sequence
Reached (10 samples, 3.1 s)	`0.0`, $-$`0.5`, $-$`0.5`, $-$`0.5`, +`0.5`, $-$`0.5`, `0.0`, `0.0`, `0.0`, $-$`0.5`
Drifted (24 samples, 7.9 s)	$-$`0.5`, +`0.5`, `0.0`, `0.0`, $-$`0.5`, $-$`0.5`, $-$`0.5`, $-$`0.5`, `0.0`, $-$`0.5`, $-$`0.5`, `0.0`, $-$`0.5`, $-$`0.5`, `0.0`, +`0.5`, `0.0`, +`0.5`, `0.0`, +`0.5`, +`0.5`, $-$`0.5`, `0.0`, +`0.5`

Because every saliency map targets the exact sign-inclusive token span of a sampled command (Section 5.1), each trial carries a complete, ordered record of the decisions being explained. Table 4 lists the chosen ω_z spans in attribution order. These two trials come from the *time-resolved* mode of Section 5.5, which runs MF-IG on *every* planner query so the accumulation is smooth frame-to-frame; at the 3 Hz query rate that is one sample per query, and the sequence lengths (10 samples over 3.1 s, 24 over 7.9 s) follow directly. The static fields of Sections 5.4 and 6 attribute every third query instead, which is $3\times$ cheaper and, because those fields pool 20 trials, loses nothing that pooling does not restore (Table 2). Spans are quoted as tokenized, sign included — Qwen splits `-0.5` into [" −", "0", ".", "5"], a 4-token span, and `0.0` into a 3-token span. The reached trial's sequence *converges*: an opening neutral sample, a held rightward turn, a single left-glance-and-correct, then a settled straight-drive phase ending at commitment (2 direction reversals in 10 samples). The drifted trial's sequence *never settles*: 16 value changes and 5 direction reversals across 24 samples, 7 of those changes falling in the final 9 samples — persistent dithering that matches its non-consolidating heat field.

E Occlusion probe: mask construction and threshold calibration

The hot voxel set is obtained by thresholding the accumulated clean field at a quantile p of accumulated saliency. Table 5 reports how the set composition and the induced masking load vary with p. Two facts drove the choice of p=0.70 as the operating point. First, the set is dominated by back-wall voxels at every threshold, and overwhelmingly so at p=0.50. Second, the masking load must be large enough to matter and small enough to leave a scene: p=0.50 masks only 14% of the frame on average, while p=0.90 masks 61%, leaving too little of the arena for the trial to remain the same task. At p=0.70 the mask covers roughly a third of the frame.

Table 5: Hot voxel set composition and induced masking load against threshold quantile p (n=20 clean "friendly" field, L=red, R=blue). *% wall* is the share of hot voxels on the back-wall plane; frame area is averaged over the drive after re-projection.

p	Hot voxels	% wall (hot set)	Mean masked frame area
0.50	508	97.2%	14.0%
0.70	1511	58.8%	33.4%
0.90	5368	39.9%	61.4%

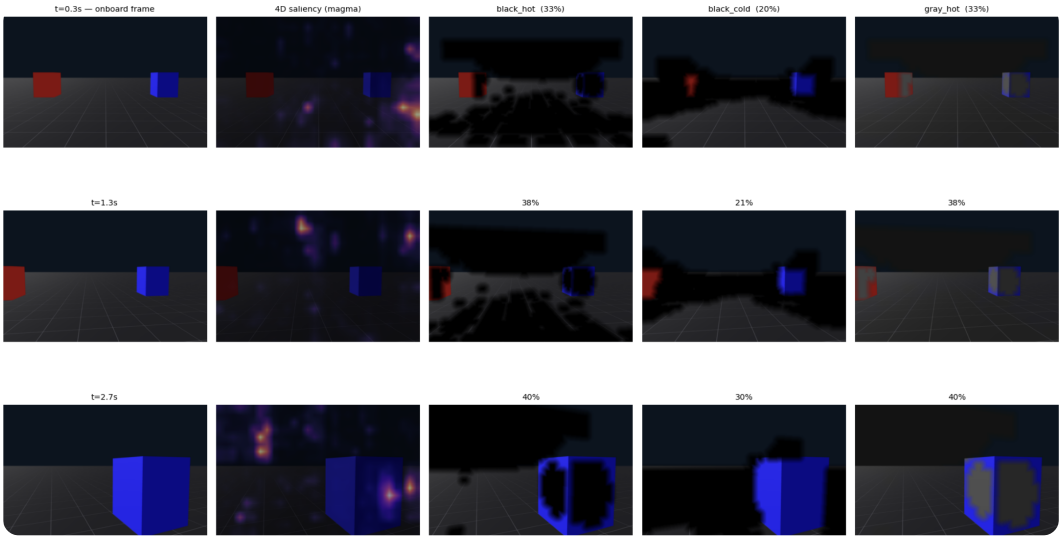


Figure 9: Per-frame occlusion sanity render, main arrangement (clean trial, p=0.70).

What the COLD control actually became. COLD was meant to be a compact region somewhere else in the arena, grown away from the hot voxels until it hid as much of the frame as HOT does. It did not stay compact. The area target was only reached once the region had expanded to take in nearly everything the hot set does not, so in practice BLACK-COLD is close to "everything but the hot set" (Figure 9); the mirrored replication is only slightly milder. It therefore still bounds the cost of blacking out a comparable slice of the view, but it does not isolate *where* the hot set is, and BLACK-HOT against BLACK-COLD should not be read as localizing the decision. The region is sampled in the world and never mirrored, because the finding under audit is itself a left/right asymmetry. Realized areas over the actual drives (Table 6) differ from these nominal ones because each arm visits different parts of the arena — a closed-loop effect with no analogue in static deletion tests. Figure 9 shows the full per-frame sanity render behind Figure 7: left to right, the onboard frame the planner received, its 4D saliency in magma, and the BLACK-HOT, BLACK-COLD and GRAY-HOT maskings of that same frame as actually presented. Because the masks are world-space voxel sets re-projected per frame, the same physical surfaces stay covered as the camera moves.

Table 6: **Closed-loop occlusion, both mirrored arrangements**, reported as measured (N=20 fresh drives per arm; "go to the friendly one"). Arms are ordered by decreasing cosine within each panel. *Mask area* is the realized mean masked fraction of the frame over the drive and *Rev.* the mean direction reversals per trial; field metrics compare each arm to its *own* panel's fresh clean field (cosine 1.000 = identical, CoM shift is a divergence). LEFT is the red cube in panel A and the blue cube in panel B.

		Behavior			Field vs. own clean	
Arm	Mask area	LEFT % [95% CI]	Reach	Rev.	cosine	CoM (m)
A. *L=red, R=blue*	*(the arrangement of Figures 7–8)*					
clean (re-drive)	—	30% [15, 52]	55%	2.15	1.000	0.000
GRAY-HOT	31.7%	40% [22, 61]	45%	2.75	0.894	0.117
BLACK-HOT	33.0%	60% [39, 78]	70%	3.10	0.866	0.475
BLACK-COLD	32.0%	90% [70, 97]	75%	2.30	0.795	0.557
B. *L=blue, R=red*	*(mirrored replication; renders in Appendix F)*					
clean (re-drive)	—	90% [70, 97]	45%	2.80	1.000	0.000
GRAY-HOT	30.0%	95% [76, 99]	65%	1.65	0.937	0.373
BLACK-COLD	20.5%	95% [76, 99]	85%	1.05	0.858	0.299
BLACK-HOT	32.1%	100% [84, 100]	90%	1.15	0.795	0.739

F THE MIRRORED OCCLUSION REPLICATION

Section 6 reports the mirrored arrangement (L=blue, R=red) alongside the main one in Table 6. This appendix gives its construction (Figure 10) and its fields (Figure 11).

Table 7: Hot voxel set composition and induced masking load in the *mirrored* arrangement (L=blue, R=red), on its own n=20 clean "friendly" field. Compare Table 5.

p	Hot voxels	% wall (hot set)	Mean masked frame area
0.50	503	98.2%	13.5%
0.70	1406	65.8%	26.3%
0.90	5397	43.9%	50.8%

The mirrored clean field accumulates to 59,905 voxels at a 0.05 m edge, of which the back-wall plane carries 75.5% of the saliency mass (89.1% of the voxels) — against 72.8% and 88.3% in the main arrangement. The wall share and the shape of the threshold curve replicate those of Table 5. At p=0.70 the hot set is 1406 voxels masking 26.3% of the frame, and the COLD ball area-matches it to within 0.0% error, spanning 34,437 voxels.

Reading the two panels against each other. The association noted in Section 6 — lowest cosine against clean, highest reach, most one-sided choice — holds in both arrangements, and is the stronger of the two here (BLACK-HOT's 90% reach against clean's 45%).

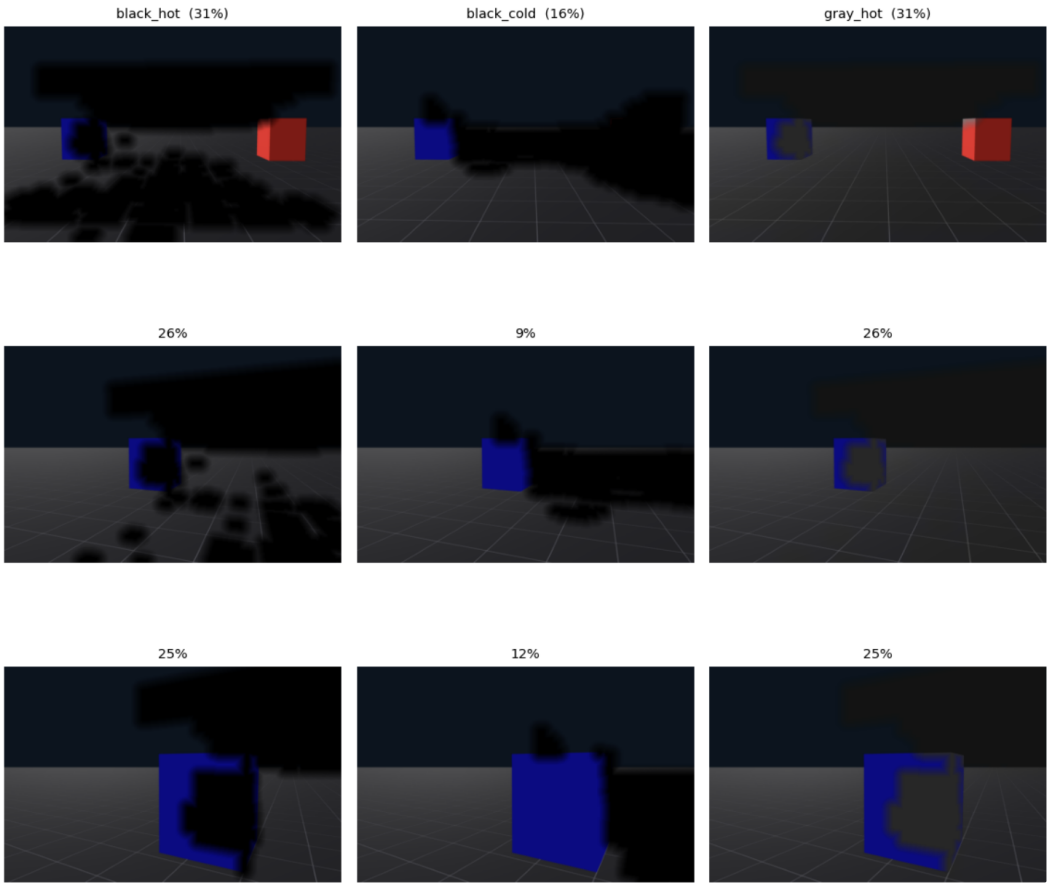
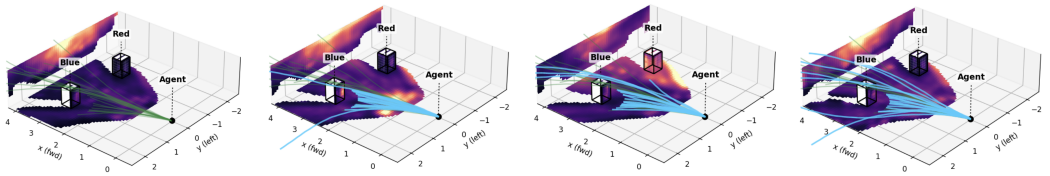


Figure 10: Per-frame occlusion sanity render for the mirrored arrangement, with the blue cube on the left; columns as in Figure 9.



(a) clean (reference) (b) BLACK-HOT (c) BLACK-COLD (d) GRAY-HOT

Figure 11: **Re-accumulated fields under closed-loop occlusion, mirrored arrangement** (n=20 fresh drives per arm; magma scale). Green = the mirrored clean trajectories, light blue = the same seeds re-driven under that arm's occlusion. Arm ordering in Table 6B.

G LOCALIZATION-NOISE ROBUSTNESS

Eq. 2 consumes camera pose and depth. In simulation both are exact; on a real platform both are estimated, and the honest question is not whether the field changes under that error but whether the *reading* an auditor takes from it does. So we run the IG pass over n=20 fresh clean drives once and cache the geometry, then re-place it algebraically under injected error — no VLM call, no re-attribution, so the saliency is held fixed and only *where it lands in the world* moves.

Two branches, cached separately. Section 5.2 splits every attributed pixel in two, and the two halves respond to localization error in completely different ways, so we cache and score them apart. A *surface* pixel has finite depth; we store its camera-frame offset $c = [(u - c_x)d/f_x,\ (v -$

$c_y)d/f_y,\ d]$, which carries its own range, and a noise draw moves it by a rigid transform,

$$\mathbf{p}' \;=\; (\mathbf{t}+\delta) \;+\; R_z(\delta_{\text{yaw}})\, R\, R_x(\delta_{\text{pitch}})\, \big(c\,(1+\varepsilon)\big). \tag{6}$$

A *horizon* pixel has no finite depth; we store only its camera-frame ray direction $\mathbf{r}$, and a noise draw re-casts that ray from the perturbed pose and re-intersects it with the same fixed back-wall plane,

$$\mathbf{d}' = R_z(\delta_{\text{yaw}})\, R\, R_x(\delta_{\text{pitch}})\, \mathbf{r}, \qquad t = \frac{X_{\text{WALL}} - (\mathbf{t}+\delta)_x}{d'_x}, \qquad \mathbf{w}' = (\mathbf{t}+\delta) + t\,\mathbf{d}', \tag{7}$$

re-validated against the plane bounds, then snapped to the plane. Because $\mathbf{r}$ carries no range, multiplicative depth error ε touches Eq. 6 only.

Ladders. Two translation models are swept: IID per-frame sensor jitter, and a WALK model integrating a per-trial drift direction scaled so end-of-trial error is $\sim\sigma_t$ — the SLAM-like case, which is the one we quote. Rotation is yaw-dominant with a 20% pitch component; roll is not representable in the level look-at camera model used throughout, so none is injected. Depth error is multiplicative, so it grows with range exactly as a stereo or time-of-flight error does. Points are thresholded at 5% of each map's peak and strided by 2; voxel edge 0.05 m, hot set at p=0.70 of accumulated mass; five noise seeds per cell, each reproducible from its seed. Every metric is computed three ways — combined, surface-only, wall-only — against the σ=0 field of the same component.

Table 8: Localization-noise sweep on n=20 clean (friendly, L=Red, R=Blue) drives (mean over 5 seeds/cell). Cosine is reported for the combined field and separately for its two branches; the remaining metrics are for the combined field. ***L/R*** is the left/right saliency mass ratio, noise-free value 0.743.

		voxel cosine					
Perturbation	σ	combined	surface	wall	CoM disp. (cm)	hot-set IoU	L/R
translation (IID)	1 cm	0.991	0.869	0.992	0.25	0.916	0.745
	2 cm	0.989	0.793	0.991	0.37	0.900	0.745
	5 cm	0.966	0.661	0.970	0.68	0.886	0.748
	10 cm	0.967	0.498	0.971	0.80	0.865	0.753
translation (WALK)	1 cm	0.999	0.933	1.000	0.10	0.934	0.742
	2 cm	0.988	0.888	0.989	0.41	0.930	0.749
	5 cm	0.983	0.795	0.985	0.62	0.909	0.747
	10 cm	0.975	0.684	0.978	0.87	0.875	0.747
rotation (WALK)	1°	0.998	0.903	0.999	0.34	0.920	0.739
	2°	0.997	0.843	0.999	0.69	0.908	0.736
	5°	0.992	0.727	0.995	1.51	0.893	0.726
depth (multiplicative)	1%	0.999	0.897	1.000	0.22	0.902	0.743
	3%	0.998	0.792	1.000	0.40	0.890	0.743
	5%	0.997	0.744	1.000	0.43	0.885	0.743

What the sweep says. The two branches separate cleanly and in the direction geometry predicts. Surface points inherit the pose error whole and lose agreement steadily under every perturbation — worst under IID translation, where independent per-frame jitter never cancels along a trajectory (0.50 at $\sigma_t = 10$ cm). Wall points barely move, because intersecting a re-cast ray with a plane discards the component of the error along the ray and keeps only what is tangent to the plane; under multiplicative depth error they are exactly invariant, since Eq. 7 never uses a range. The combined field sits close to the wall curve, because the wall carries about three quarters of the accumulated mass. The centre of mass moves by well under a centimetre everywhere except $5°$ of yaw error, and the left/right ratio stays in $[0.726, 0.753]$ against a noise-free 0.743 across all fourteen cells — the only visible trend being a slow slide under rotation, which is the perturbation that actually mixes the two sides.

Limits of this analysis. The wall's stability is real geometry, but it is geometry that assumes a *known* plane: here the wall is a simulator constant, whereas a real system would fit that plane from the same noisy sensing, adding an error term Eq. 7 does not carry. The combined-field numbers

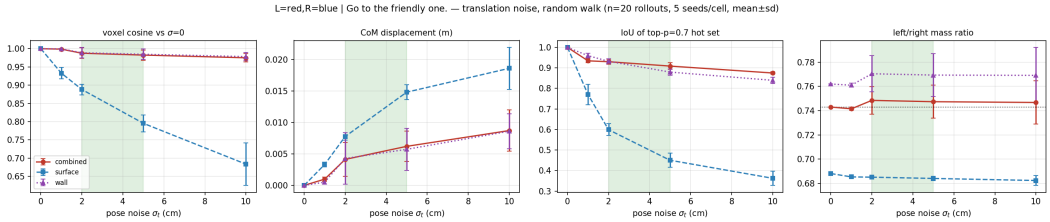


Figure 12: Translation ladder, random-walk model: each metric against pose error, split into combined, surface-only and wall-only fields. The shaded 2–5 cm band is a reference band shown for scale only: reported indoor lidar-inertial ATE varies widely across sequences and often exceeds 10 cm (Shan et al., 2020; Xu et al., 2022), so capping the test ladder at 10 cm is optimistic. The gap between the blue and purple curves is the main purpose of this appendix.

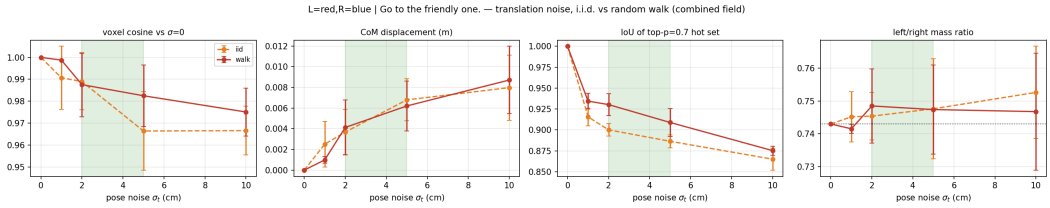


Figure 13: Translation ladder, IID per-frame jitter against the integrated random walk, combined field only. The two models are close on the combined field — the wall component dominates it and is insensitive to which model generated the pose error.

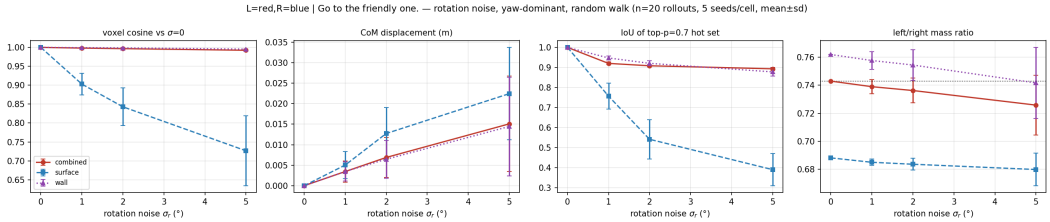


Figure 14: Rotation ladder (yaw-dominant, 20% pitch, random walk). Rotation is the only perturbation that rotates rays across the arena's midline, so it is the one that moves the left/right ratio at all.

should therefore be read as an upper bound, and the surface-only column as the closer analogue of a scene with no clean background plane at all. Real-world depth errors are also spatially correlated, so our per-pixel independent noise averages out too efficiently during accumulation. This is a sensitivity bound on the unprojection, not a real-robot validation — a physical port would additionally face depth dropout, rolling shutter and calibration error that Eqs. 6–7 do not model.

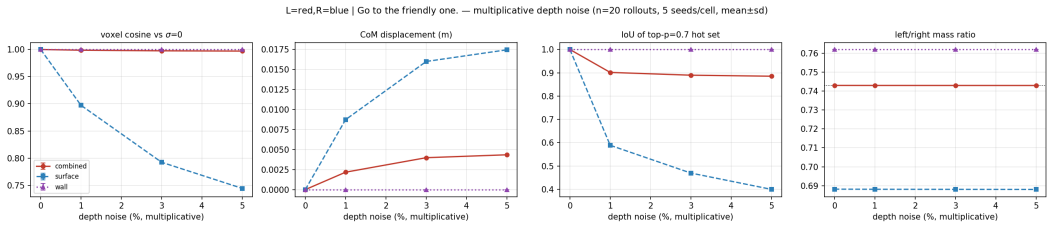
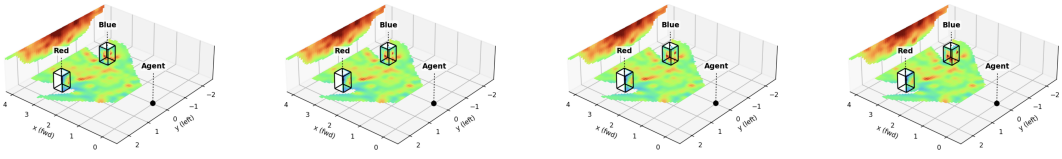
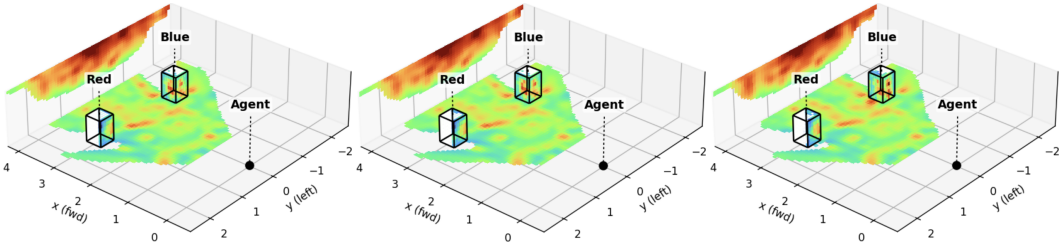


Figure 15: Multiplicative depth ladder. Horizon rays carry no range, so Eq. 7 is untouched by ε and the wall field is bit-identical to the reference. All of the combined field's movement here comes from the surface branch.



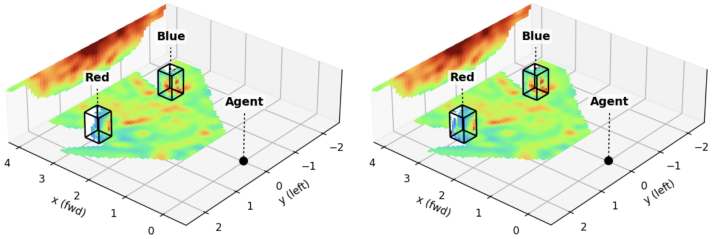
(a) $\sigma_t = 0$ (reference) (b) $\sigma_t = 2\,\text{cm}$ (c) $\sigma_t = 5\,\text{cm}$ (d) $\sigma_t = 10\,\text{cm}$

Figure 16: Accumulated 20 fresh fields under random-walk translation noise. Visually there is little to no difference; see *Limits of this analysis* above.



(a) $\sigma_r = 1^\circ$ (b) $\sigma_r = 2^\circ$ (c) $\sigma_r = 5^\circ$

Figure 17: Rendered fields under rotation error, the rotation counterpart of Figure 16. Compare against the σ=0 reference, Figure 16a.

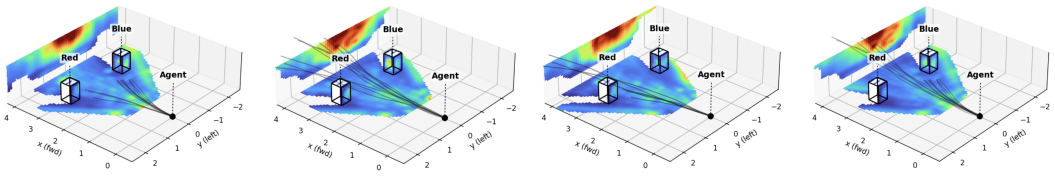


(a) 3% depth error (b) 5% depth error

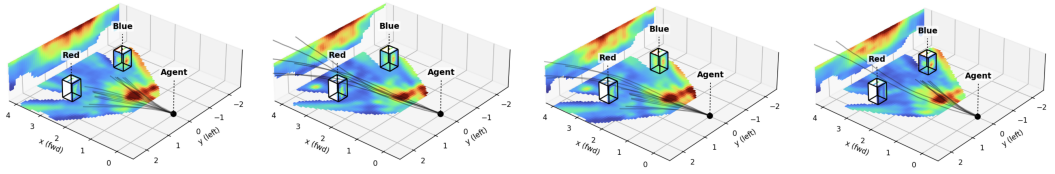
Figure 18: Rendered fields under multiplicative depth error, against the same σ=0 reference. Only floor and cube-face structure can move here; the wall lobe is unchanged.

H ADDITIONAL SALIENCY RENDERS

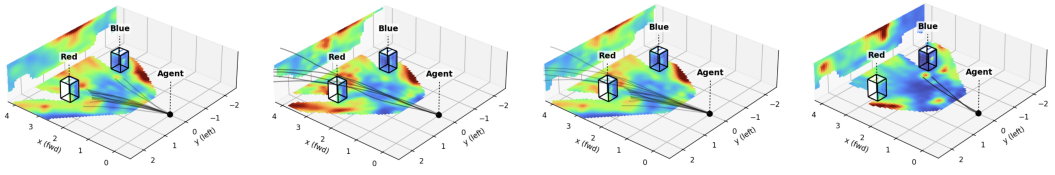
Figure 19 slices every occlusion arm of Section 6 by outcome (for L=red, R=blue): the choice split shifts from 6/14 under clean observations to 18/2 under BLACK-COLD, the prior reversion reported in Table 6.



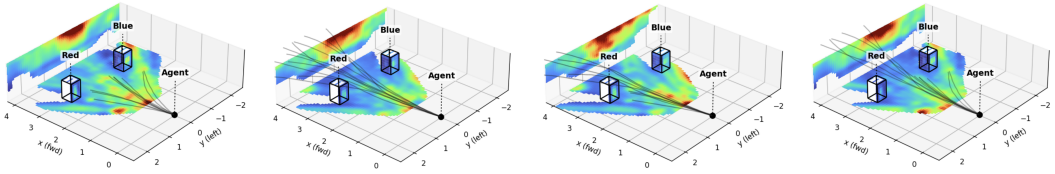
(a) clean reached (11) (b) clean drifted (9) (c) clean nearer red (6) (d) clean nearer blue (14)



(e) BLACK-HOT reached (14) (f) BLACK-HOT drifted (6) (g) BLACK-HOT nearer red (12) (h) BLACK-HOT nearer blue (8)



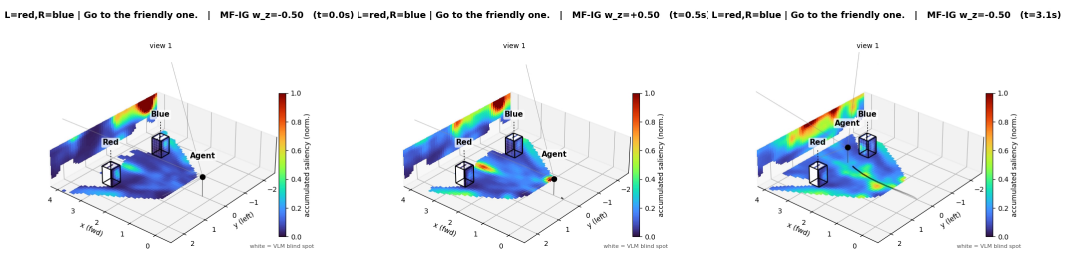
(i) BLACK-COLD reached (15) (j) BLACK-COLD drifted (5) (k) BLACK-COLD nearer red (18) (l) BLACK-COLD nearer blue (2)



(m) GRAY-HOT reached (9) (n) GRAY-HOT drifted (11) (o) GRAY-HOT nearer red (8) (p) GRAY-HOT nearer blue (12)

Figure 19: Outcome-split fields for all four re-drive arms (L=red, R=blue; n=20 per arm; group sizes in parentheses). Rows: clean, BLACK-HOT, BLACK-COLD, GRAY-HOT. Columns: reached, drifted, nearer-red, nearer-blue.

Figure 20 is the drifted counterpart of Figure 6. Both trials steer initially toward blue, and because the columns are aligned to the reached trial's commit moment at t=3.1 s, panel (c) shows where the drifted agent was and what it was attending to *when it should have committed* — a field spanning roughly twice the hot area of Figure 6c, smeared across both candidates, with the agent still mid-arena until timeout. Figure 21 then accumulates the "go to the hostile one" condition: attention is divided far more evenly between the two targets than in the "friendly" field of Figure 4d, mirroring the null color effect for "hostile" in Table 1.



(a) drift, $t = 0.0\,\mathrm{s}$ (b) drift, $t = 0.5\,\mathrm{s}$ (c) drift, $t = 3.1\,\mathrm{s}$

Figure 20: **The matched drifted trial** complementing Figure 6, with columns aligned to the same $t \in \{0.0, 0.5, 3.1\}$ s. Token sequence in Table 4.

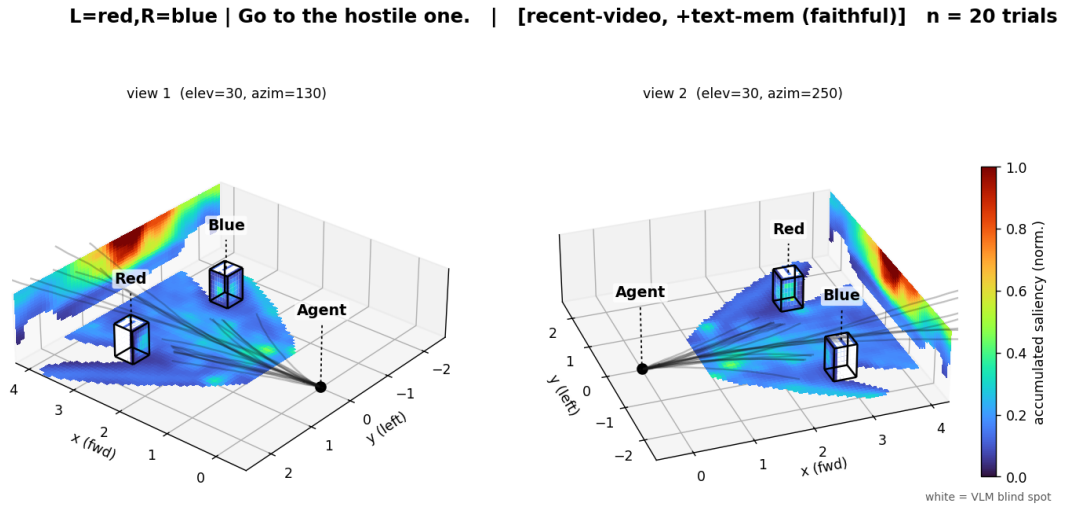
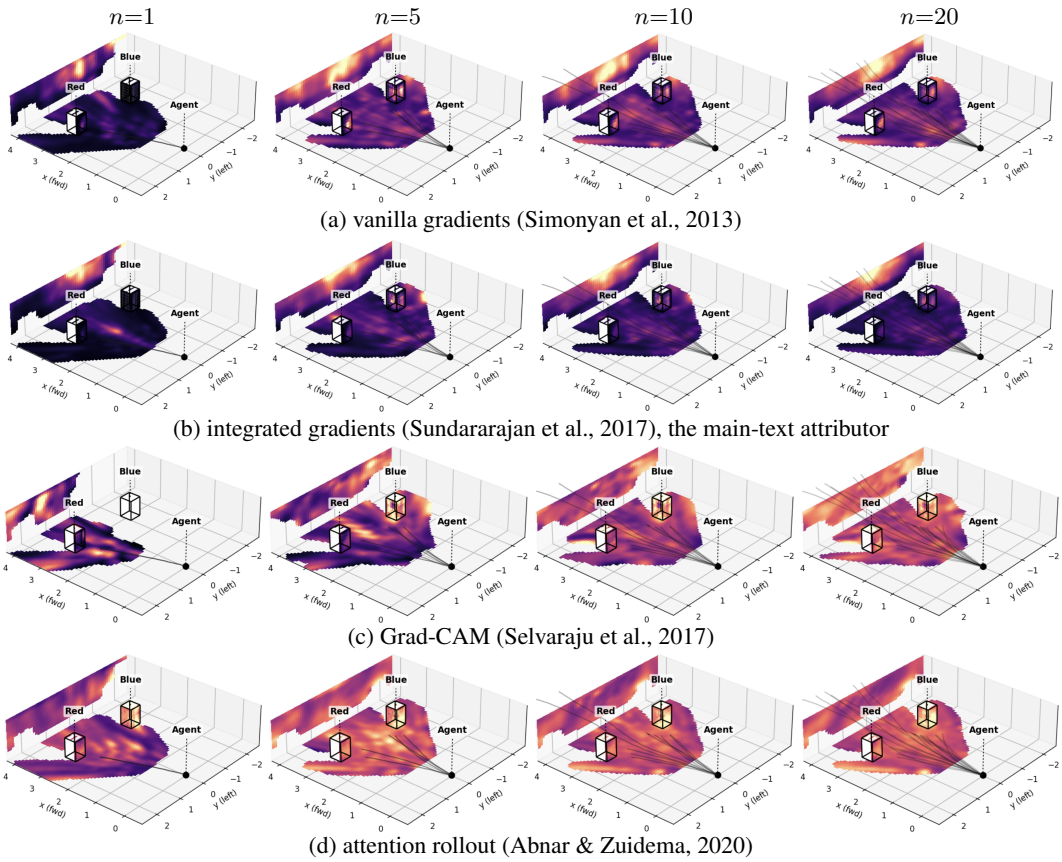


Figure 21: Accumulated heat field at n=20 for the "go to the hostile one" instruction (L=red, R=blue).

I ALTERNATIVE ATTRIBUTION METHODS

Figure 22 swaps the stage-(i) attributor; Figure 23 removes gradients from it entirely.



(a) vanilla gradients (Simonyan et al., 2013)

(b) integrated gradients (Sundararajan et al., 2017), the main-text attributor

(c) Grad-CAM (Selvaraju et al., 2017)

(d) attention rollout (Abnar & Zuidema, 2020)

Figure 22: **The 4D field is attributor-agnostic.** Same cache, instruction and sweep as Figure 4 (L=red, R=blue, "friendly"; magma).

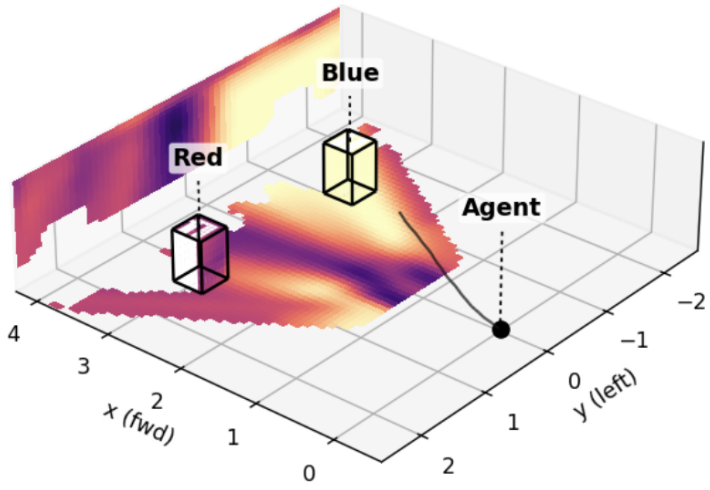
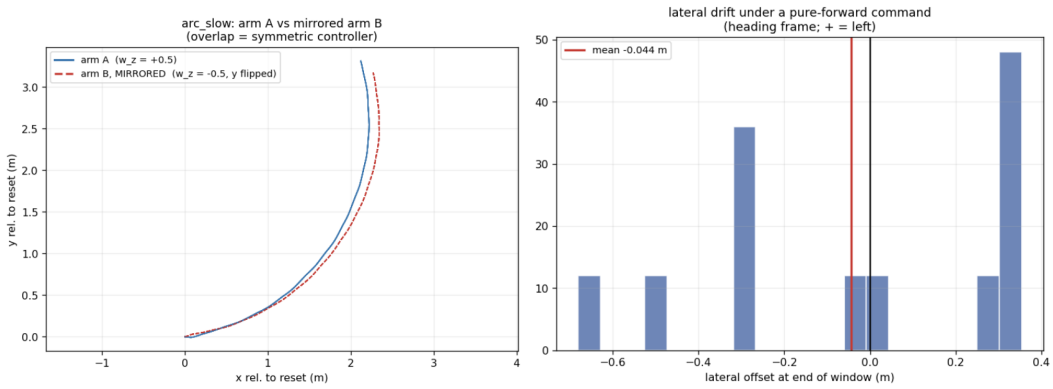


Figure 23: **A black-box planner.** RISE (Petsiuk et al., 2018) attributes an API-only Gemini 3.5 Flash-Lite planner (Google DeepMind, 2026) by random-mask perturbation; the scores unproject unchanged (n=1). White-box access is a property of the attributor, not of the framework. Note not the same cache.

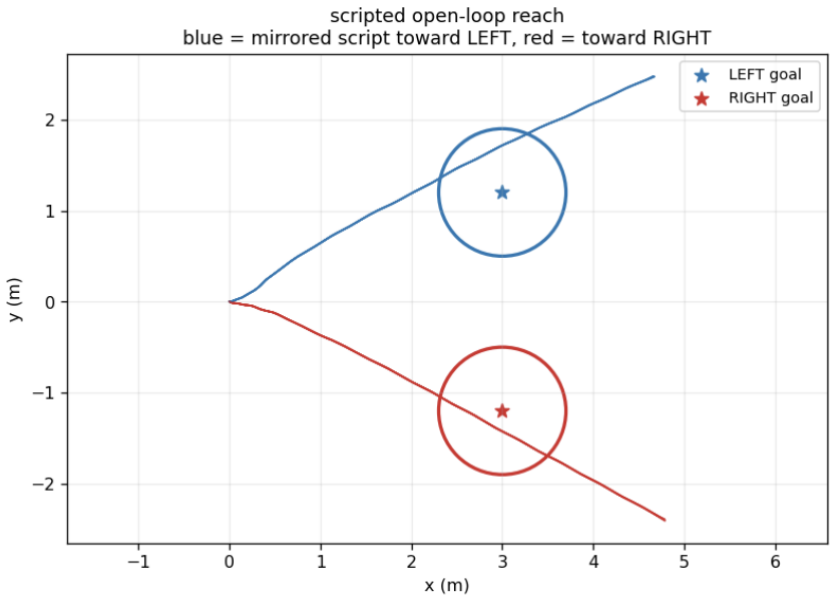
J LOW-LEVEL CONTROLLER AUDITS

The positional bias of Table 1 is attributable to the planner only if the gait policy underneath it is side-symmetric. Figure 24 audits the controller alone, with the VLM removed and commands scripted open-loop.



(a) mirrored constant-yaw arcs

(b) drift under a pure-forward command

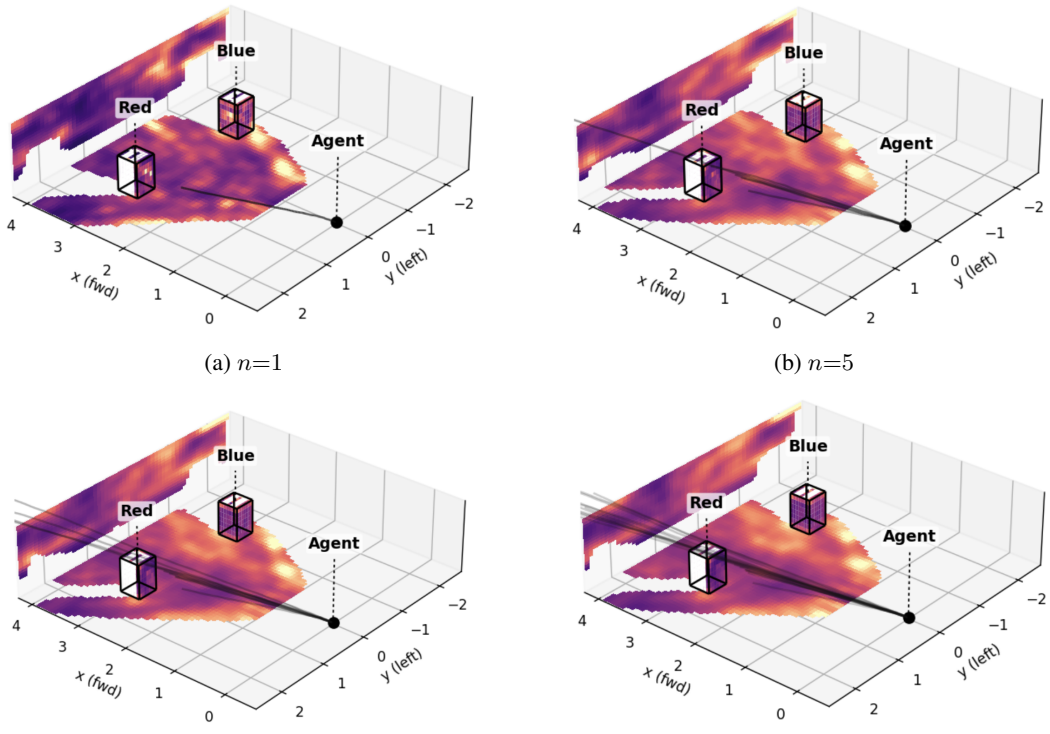


(c) mirrored open-loop reach scripts

Figure 24: **The gait policy is side-symmetric** (VLM removed, commands scripted). Neither turn direction, straight-line drift, nor open-loop reach favors a side, so the left bias enters above the controller.

K ADDITIONAL PLANNERS

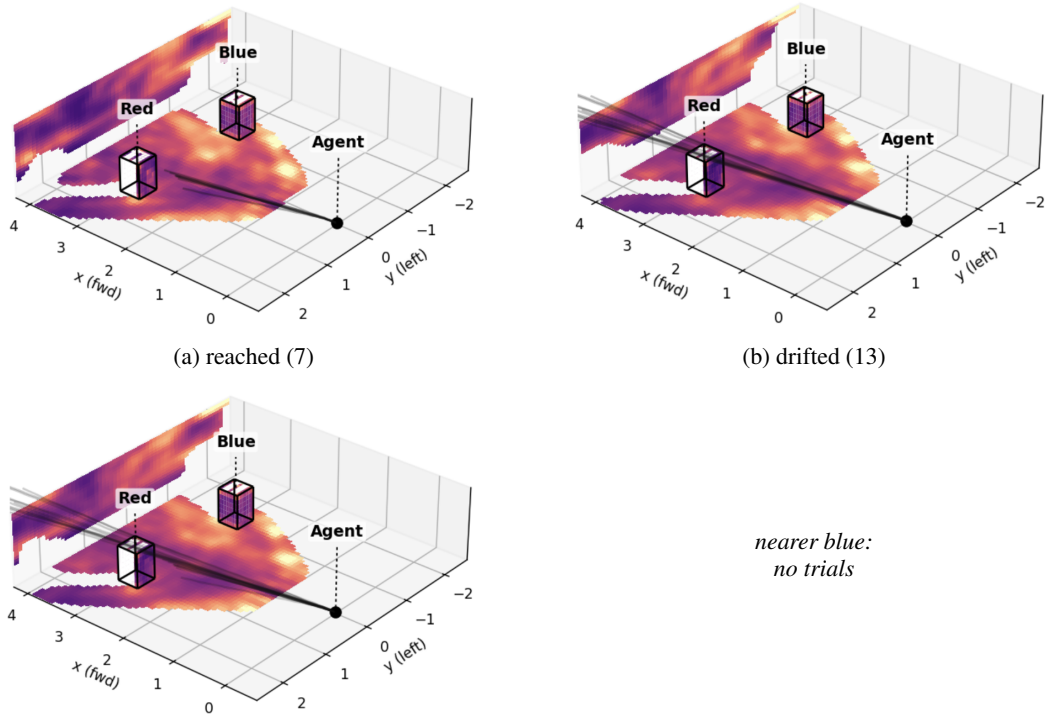
The pipeline taps a sampled numeric token span, not an architecture, so swapping the frozen planner changes nothing upstream of Eq. 1. Figures 25–28 re-run the convergence sweep of Section 5.4 and the outcome split of Figure 5 on two further local planners, InternVL3-2B (Zhu et al., 2025) and SmolVLM2-2.2B (Marafioti et al., 2025). Figure 23 (Appendix I) covers the API-only case, where RISE's per-query mask budget makes multi-trial accumulation impractical. Same arena, same instruction throughout. Only the single arrangement (L=red, R=blue) was driven, so these fields locate attention but do not decompose colour from position the way Table 1 does.



(a) n=1 (b) n=5

(c) n=10 (d) n=20

Figure 25: **InternVL3-2B convergence sweep**, "go to the friendly one" (L=red, R=blue; magma).

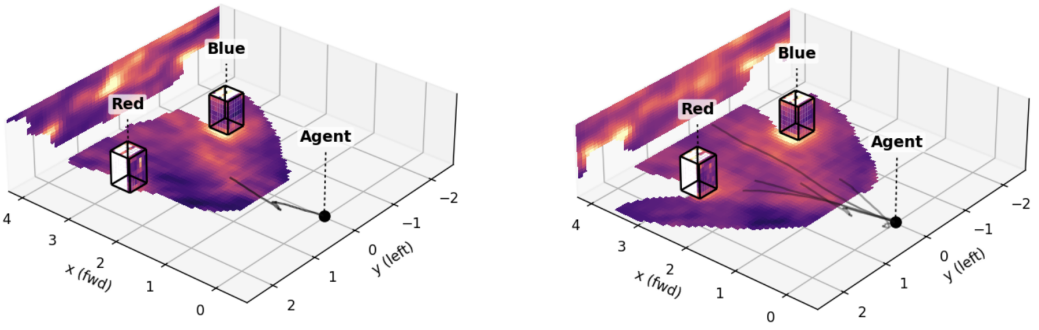


(a) reached (7) (b) drifted (13)

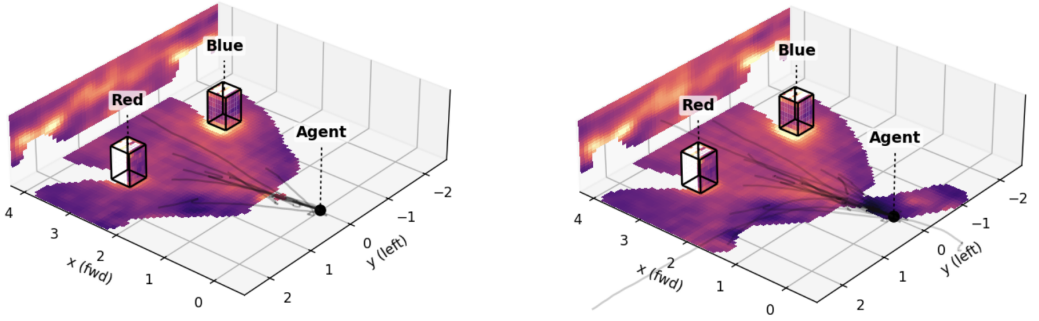
(c) nearer red (20)

nearer blue:
no trials

Figure 26: **InternVL3-2B outcome split**, same n=20 drives as Figure 25; group sizes in parentheses.

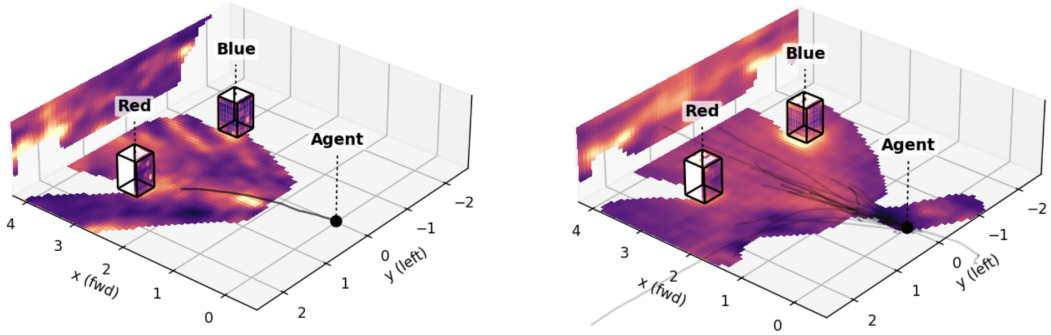


(a) n=1 (b) n=5

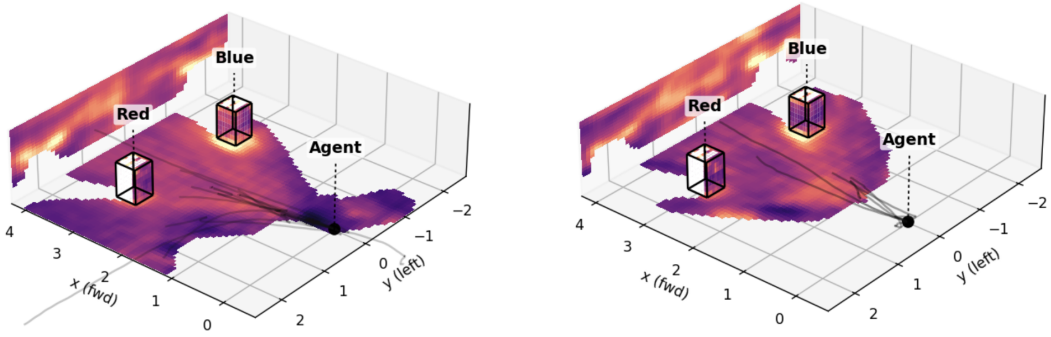


(c) n=10 (d) n=20

Figure 27: **SmolVLM2-2.2B convergence sweep**, same arena and instruction as Figure 25.



(a) reached (1) (b) drifted (19)



(c) nearer red (15) (d) nearer blue (5)

Figure 28: **SmolVLM2-2.2B outcome split**, same n=20 drives as Figure 27.

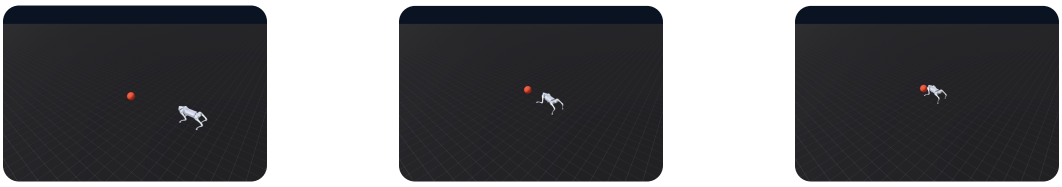
L A DYNAMIC ARENA: CHASING A MOVING BALL

Nothing in Eq. 2 assumes a static scene, but a moving target is the case where world-anchoring is least obviously well posed: the surface carrying the saliency is itself travelling through the frame. We therefore replace the two cubes with a single red ball that flees the agent, and keep the Qwen2.5-VL-3B + PPO stack and the attribution tap otherwise unchanged.

The task. The ball starts straight ahead of the agent and dead centre in frame, so the opening decision is unambiguous and no trial is decided by an initial search. It travels at $\sim$0.28 m/s against the agent's 2.5 m/s ceiling, so it is always catchable in principle and the variable under study is *how the agent tracks it* rather than whether it can win a race. Its heading is re-drawn every $\sim$1.2 s from a von Mises distribution centred on the direction away from the agent, making it evasive without being adversarial or unpredictable in principle. A trial ends the moment the agent is within the 0.60 m commit radius, or at a 10 s timeout; N=20 trials (Table 9).

One deviation from the static setup. The command low-pass filter of Section 3 (α=0.35) is removed here: each sampled (v_x, v_y, ω_z) is passed to the gait policy directly. Filtering exists to keep a commitment to a stationary target from dithering, but against a target that re-picks its heading every $\sim$1.2 s the lag it introduces is the dominant error — unfiltered, a command takes effect within the query interval and the agent can turn sharply enough to stay with the ball. Attribution is unaffected: the token span IG targets is the planner's sampled command either way, and removing the filter only shortens the delay between that command and the motion it explains.

Fields are accumulated in the same world frame as everywhere else; the ball's own path is overlaid as a fading green trail.

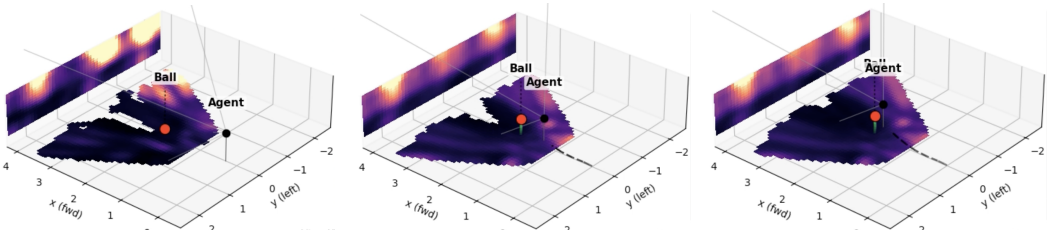


(a) reset (b) pursuit (c) catch

Figure 29: **The chase task.** One trial, agent and fleeing ball.

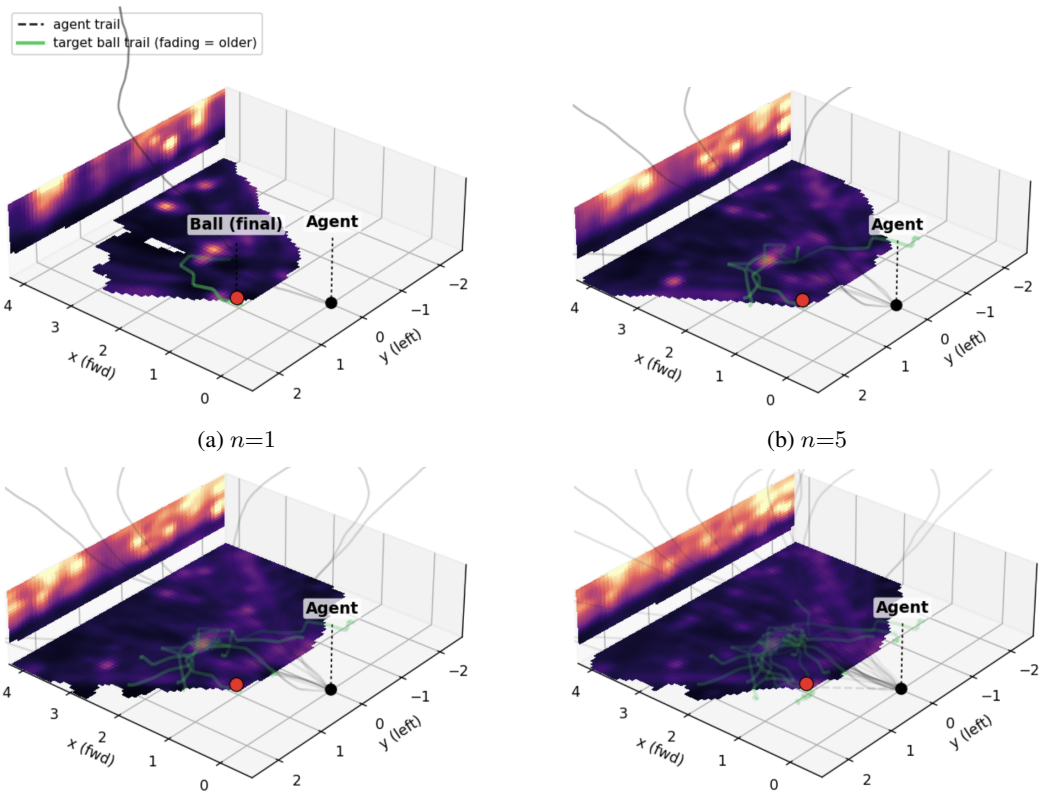
Table 9: **Chase outcomes** (N=20 trials; Qwen2.5-VL-3B + PPO, 4D IG). Bracket is a Wilson 95% CI.

Quantity	Value
Reached (caught within 0.6 m)	6 (30.0% [14.5, 51.9])
Drifted (timed out at 10 s)	14 (70.0%)
Time to catch (reached only)	mean 2.08 s, median 1.47 s, range 1.18–5.40 s
Closest approach (all trials)	mean 0.775 m, median 0.703 m



(a) start, ω_z=0.0 (b) mid, ω_z=−0.8 (c) commit, ω_z=+0.6

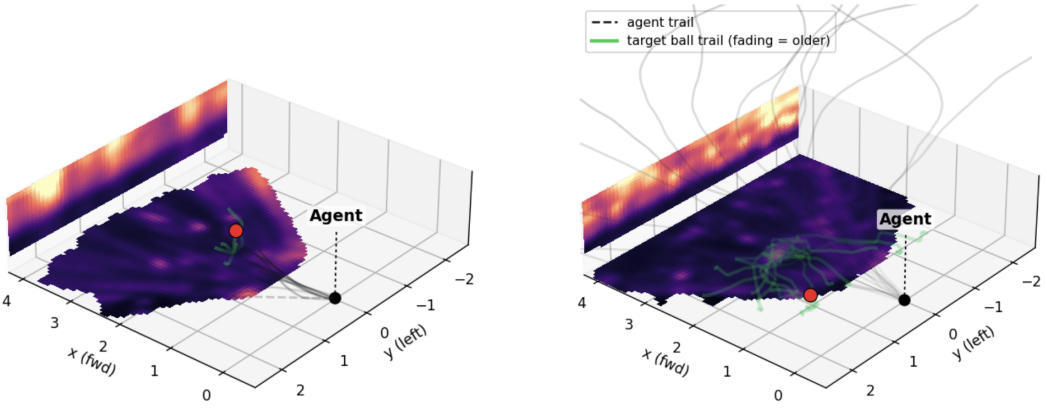
Figure 30: **Time-resolved accumulation with moving target**, one trial, panels labelled by attributed ω_z span.



(a) n=1 (b) n=5

(c) n=10 (d) n=20

Figure 31: **Chase convergence sweep**, pooled over trials.



(a) caught (6) (b) timed out (14)

Figure 32: **Chase outcome split**, the same 20 drives as Figure 31.

M THE 2D PER-FRAME BASELINE, AND WHY IT DOES NOT SCALE

Section 1 argues that per-frame saliency cannot be aggregated for a driving agent. Figures 33 and 34 make the claim concrete: what the auditor is actually handed, and how the pile grows.

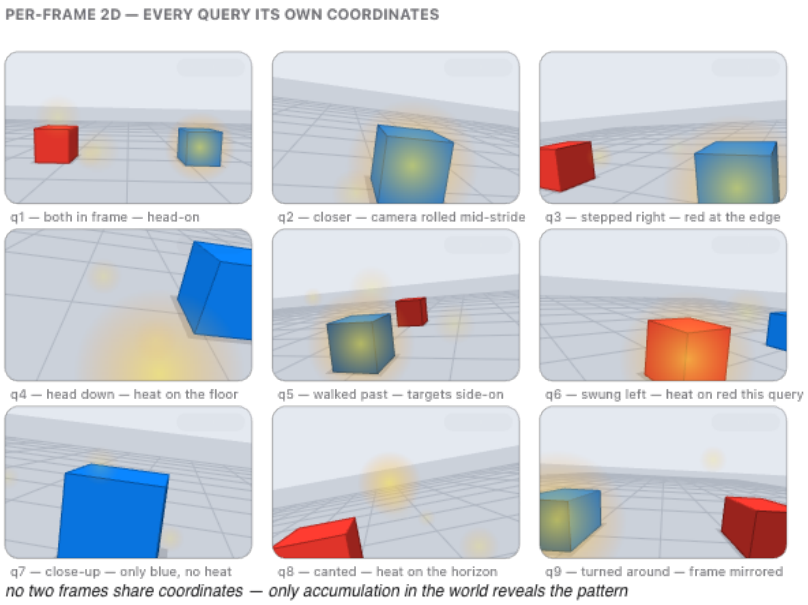


Figure 33: Nine example queries of one trial as per-frame 2D saliency (usually there are more queries/frames to inspect in a trial). Each map lives in its own camera pose, so no two are comparable and none can be summed.

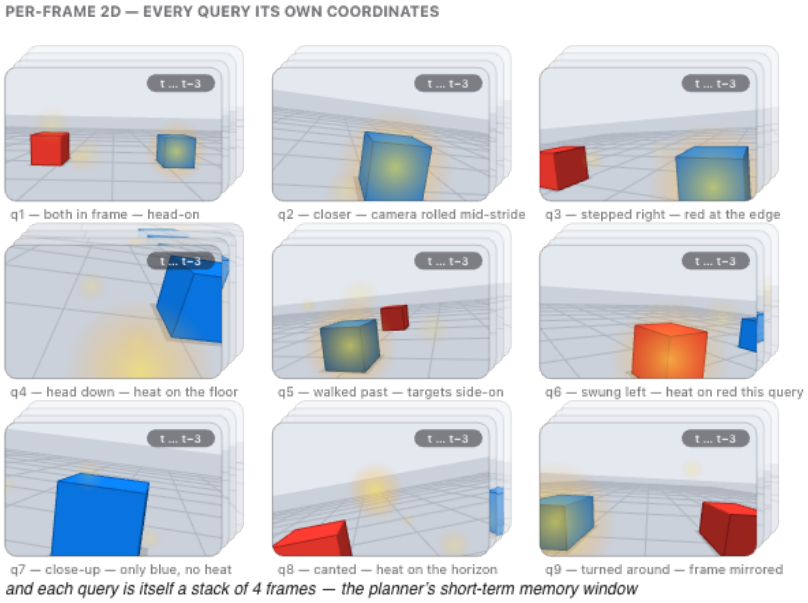


Figure 34: The same nine queries with the 4-frame short-term memory window of Section 3: $9 \rightarrow 36$ maps. The stack grows with query rate, window length and episode duration — worst exactly where long-horizon tasks live.

N THEORETICAL ANALYSIS

We analyze the statistics of a specified attribution field, the geometry of placing that field in the world, and the interpretation of mirror and failure diagnostics. These are properties of an auditing construction with a fixed planner and controller. No learning guarantee for PPO, sufficiency claim for compressed memory, or causal-faithfulness theorem for an attribution method is assumed or established.

N.1 DEFINITIONS, ASSUMPTIONS, AND IMPLEMENTATION CORRESPONDENCE

Trial fields and quantitative normalization. Fix the planner, controller, prompt condition, attribution rule, and all preprocessing hyperparameters. A trial includes its trajectory, sampled planner outputs, memory snapshots, camera frames, and corresponding geometry. A fixed measurable map sends this complete record to $Z = (Z_1, \dots, Z_M) \in [0, B]^M$ on a fixed finite spatial or space–time partition. The deterministic bound $B < \infty$ is an assumption on the *quantitative* field, not a consequence of plotting a color scale.

There are two direct ways to instantiate this assumption. If each retained nonnegative weight is at most b_0 and at most $N_{\max}$ records are kept per trial, unnormalized bin sums satisfy $B = b_0 N_{\max}$. Alternatively, divide each nonempty trial's bin sums by that trial's total retained mass; then $B = 1$ and the bin masses sum to one. Define an empty trial's vector to be zero. Either choice must be fixed before estimating a population field; they weight trials differently and should not be interchanged silently. Per-query quantile scaling alone does not imply a deterministic bound unless clipping or another bound is also specified.

Our estimator is $\bar{Z}_n = n^{-1} \sum_{i=1}^{n} Z_i$, with target $\mu_Z = \mathbb{E}Z_1$. A pooled renderer may subsequently change its color range or divide by a sample-dependent maximum; the concentration statement below concerns $\bar{Z}_n$ before such display operations. It supplies a statistical interpretation of accumulation, without claiming that the existing rendered sweeps already implement a particular confidence band.

Attribution measures and geometry. A pixel record includes its query, its frame within the memory window, its pixel coordinates, and its timestamp. This distinction ensures that each historical frame uses its own depth and pose. After a fixed conversion to nonnegative weights and any chosen thresholding, let ρ be the normalized measure on these records. Geometric propositions require positive retained mass; an empty field has no normalized probability-measure interpretation. Projection U maps a record to a world point. When keeping time explicitly, use $U^{(4)}(p) = (U(p), \tau_p)$ and the metric $d((x,\tau),(y,\sigma)) = \|x - y\|_2 + c_t|\tau - \sigma|$, with fixed $c_t > 0$. If timestamps are unchanged, the spatial displacement bounds below apply unchanged to this space–time representation.

Query-level state and planner output. At query k, let S_k contain the physical state, scene configuration, controller/filter state, camera-frame history, and incoming text memory. Include any further internal variable needed for a Markov description. The frozen planner samples

$$Y_k = (A_k, C_{k+1}) \sim \Pi_k(\cdot \mid S_k), \qquad A_k = (v_{x,k}, v_{y,k}, \omega_{z,k}),$$

where C_{k+1} includes generated memory and any other output that is retained and affects future queries. Conditional on (S_k, Y_k), the kernel $P_k(\cdot \mid S_k, Y_k)$ contains physical execution, filtering, observation updates, and deterministic storage of C_{k+1}. Stochastic memory generation belongs to Π_k, not to a putative controller-only kernel. All spaces are standard Borel, all maps/kernels are measurable, and the state law includes the relevant randomness. This is an analysis state; it need not be fully observed by the VLM or recoverable from its compressed summary.

Reflection and horizons. Let M_s and M_y be measurable involutions on states and planner outputs, respectively. Reflect the physical scene and body variables, the image history, the filter state, and directional content of retained memory consistently. In particular, $M_a(v_x, v_y, \omega_z) = (v_x, -v_y, -\omega_z)$. Color identities remain attached to their reflected objects. A reflection of memory is a prescribed involutive transformation, not a newly sampled VLM paraphrase. Use H queries, padding terminated runs with absorbing states that retain the terminal outcome. Thus reached, drifted, or nearest-target labels can be terminal events even when episode lengths differ.

Distances. We use $d_{\mathrm{TV}}(P,Q) = \sup_E |P(E) - Q(E)|$. For probability measures with finite first moments on a metric space,

$$W_1(P,Q) = \inf_{\Gamma \in \mathcal{C}(P,Q)} \int d(x,y)\, \Gamma(dx,dy).$$

The pushforward $U_{\#}\nu$ is the law of $U(X)$ when $X \sim \nu$.

N.2 Concentration of pooled attribution fields

Proposition N.1 (Uniform concentration)**.** *Let $Z_1, \dots, Z_n$ be independent, identically distributed random vectors in $[0,B]^M$. For $0 < \delta < 1$, with probability at least $1-\delta$,*

$$\|\bar{Z}_n - \mu_Z\|_\infty \le B\sqrt{\frac{\log(2M/\delta)}{2n}}.$$

Proof. For each coordinate m, the bounded-scalar exponential estimate in Lemma N.11 gives

$$\Pr\big(|\bar{Z}_{nm} - \mu_{Zm}| \ge t\big) \le 2\exp(-2nt^2/B^2).$$

A union bound over $m = 1, \dots, M$ gives $\Pr(\|\bar{Z}_n - \mu_Z\|_\infty \ge t) \le 2M\exp(-2nt^2/B^2)$. Set the right-hand side to δ. This proves Proposition 1 as well. □

Corollary N.2 (Fixed regional summaries)**.** *Let T_i be a fixed measurable summary of trial i with $T_i \in [a,b]$. For independent, identically distributed trials, with probability at least $1-\delta$,*

$$\left| n^{-1}\sum_{i=1}^{n} T_i - \mathbb{E}T_1 \right| \le (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

In particular, for fixed nonnegative weights with $\sum_m w_m = 1$, $T_i = \sum_m w_m Z_{im} \in [0,B]$. If Z_i instead records normalized bin masses*, its mass on any fixed union of bins lies in $[0,1]$; the difference between two such masses lies in $[-1,1]$.*

Proof. Apply Lemma N.11 to $T_i - a$. The displayed ranges follow from nonnegative weights and, for mass summaries, total mass at most one. Ordinary unnormalized regional sums are not convex averages: their range must be bounded separately. □

Dependence within a trial and repeated checkpoints. No independence is assumed among a trial's queries, overlapping frame windows, pixels, or bins. Re-running attribution on a cached drive does not create another independent trial. For the four predetermined checkpoints $n \in \{1,5,10,20\}$, replacing $\log(2M/\delta)$ by $\log(8M/\delta)$ gives a simultaneous bound, by a union bound. The checkpoints may be nested prefixes of the same cache. An arbitrary data-dependent stopping time is not covered by that four-checkpoint claim.

Outcome-split fields. For a predetermined per-trial labeling rule $G_i = g(\text{trial } i)$ and a label g of positive probability, condition on the indicator vector $(\mathbf{1}\{G_i = g\})_i$. The selected trial fields are independent, each with law $\mathcal{L}(Z_1 \mid G_1 = g)$. Apply the same proof with the positive observed group size n_g, and target $\mathbb{E}[Z_1 \mid G_1 = g]$. There is no claim for an empty group. A union bound can cover finitely many prespecified groups and checkpoints, even when groups overlap. Thus the reached, drifted, and target-choice slices have different effective sample sizes. A region chosen from the same pooled heatmap is not a fixed region for the scalar corollary; use independent selection/evaluation data or a suitable simultaneous bound for the selection class.

N.3 Stability of world-space attribution

Proposition N.3 (Stability under a common attribution measure)**.** *For a probability measure ν and maps $U, \widehat{U}$ into $\mathbb{R}^3$ whose pushforwards have finite first moments,*

$$W_1(U_{\#}\nu, \widehat{U}_{\#}\nu) \le \int \|U(p) - \widehat{U}(p)\|_2\, \nu(dp) \le \sup_{p \in \operatorname{supp}\nu} \|U(p) - \widehat{U}(p)\|_2.$$

Proof. For $X \sim \nu$, the law of $(U(X), \widehat{U}(X))$ is a coupling of the two pushforwards. Its transportation cost is the displayed integral. Taking an infimum over all couplings proves the first inequality; the second follows because ν has total mass one. The same construction with an unchanged timestamp proves the space–time version and Proposition 2. □

Corollary N.4 (Finite-depth surface points)**.** *For a pixel with homogeneous coordinate $\widetilde{p} = (u, v, 1)^\top$, set*

$$q = K^{-1}\widetilde{p}, \quad \widehat{q} = \widehat{K}^{-1}\widetilde{p}, \qquad U = Rdq + t, \quad \widehat{U} = \widehat{R}\widehat{d}\widehat{q} + \widehat{t}.$$

Assume $R, \widehat{R} \in \mathrm{SO}(3)$, $0 \le d, \widehat{d} \le d_{\max}$, $\|q\|_2 \le q_{\max}$, and $\|\widetilde{p}\|_2 \le p_{\max}$. Suppose, uniformly over retained surface records,

$$|d - \widehat{d}| \le \epsilon_d, \quad \|R - \widehat{R}\|_{\mathrm{op}} \le \epsilon_R, \quad \|t - \widehat{t}\|_2 \le \epsilon_t, \quad \|K^{-1} - \widehat{K}^{-1}\|_{\mathrm{op}} \le \epsilon_K.$$

Then $\|U - \widehat{U}\|_2 \le e_{\mathrm{surf}}$, where

$$e_{\mathrm{surf}} = d_{\max} q_{\max} \epsilon_R + q_{\max} \epsilon_d + d_{\max} p_{\max} \epsilon_K + \epsilon_t. \tag{8}$$

Proof. Write

$$U - \widehat{U} = (R - \widehat{R})dq + \widehat{R}(d - \widehat{d})q + \widehat{R}\widehat{d}(q - \widehat{q}) + (t - \widehat{t}).$$

Rotation matrices have operator norm one, and $\|q - \widehat{q}\|_2 \le p_{\max}\epsilon_K$. Bounding the four terms proves the claim. □

Proposition N.5 (Horizon rays intersecting a known wall)**.** *Fix a plane $\{x : n^\top x = b\}$ with $\|n\|_2 = 1$. Let $v, \widehat{v}$ be unit world-space ray directions and define*

$$\lambda = \frac{b - n^\top t}{n^\top v}, \quad \widehat{\lambda} = \frac{b - n^\top \widehat{t}}{n^\top \widehat{v}}, \qquad U = t + \lambda v, \quad \widehat{U} = \widehat{t} + \widehat{\lambda}\widehat{v}.$$

Assume both intersections are valid forward hits on the same wall face, and, for constants $\kappa > 0$ and $h_{\max} < \infty$,

$$|n^\top v|, |n^\top \widehat{v}| \ge \kappa, \qquad |b - n^\top t|, |b - n^\top \widehat{t}| \le h_{\max}.$$

If $\|t - \widehat{t}\|_2 \le \epsilon_t$ and $\|v - \widehat{v}\|_2 \le \epsilon_v$, then

$$\|U - \widehat{U}\|_2 \le e_{\mathrm{wall}} := (1 + \kappa^{-1})\epsilon_t + h_{\max}(\kappa^{-1} + \kappa^{-2})\epsilon_v. \tag{9}$$

For $v = Rq/\|q\|_2$, $\widehat{v} = \widehat{R}\widehat{q}/\|\widehat{q}\|_2$, and $\|q\|_2, \|\widehat{q}\|_2 \ge q_{\min} > 0$, one may take

$$\epsilon_v = \epsilon_R + 2p_{\max}\epsilon_K / q_{\min}. \tag{10}$$

In particular, changing only finite-depth estimates leaves this branch unchanged.

Proof. The two numerators differ by at most ϵ_t, while the denominators differ by at most ϵ_v. Subtracting the fractions gives

$$|\lambda - \widehat{\lambda}| \le \epsilon_t/\kappa + h_{\max}\epsilon_v/\kappa^2, \qquad |\lambda| \le h_{\max}/\kappa.$$

Use $U - \widehat{U} = (t - \widehat{t}) + \lambda(v - \widehat{v}) + (\lambda - \widehat{\lambda})\widehat{v}$ and $\|\widehat{v}\|_2 = 1$ to obtain equation 9. For nonzero vectors $q, \widehat{q}$, the triangle and reverse triangle inequalities imply

$$\left\| \frac{q}{\|q\|_2} - \frac{\widehat{q}}{\|\widehat{q}\|_2} \right\|_2 \le \frac{2\|q - \widehat{q}\|_2}{\min(\|q\|_2, \|\widehat{q}\|_2)}.$$

Insert this estimate between the two rotations and use $\|q - \widehat{q}\|_2 \le p_{\max}\epsilon_K$ to prove equation 10. The plane intersection contains no finite depth variable, proving the final claim. □

Wall-face boundaries and conditioning. The non-grazing requirement is essential: a denominator approaching zero can amplify arbitrarily small pose or ray errors. For a rectangular wall, if the original hit lies more than e_{wall} from every face boundary within the plane, and the noisy ray remains a forward hit, the displacement bound ensures face validity is preserved. Revalidation can otherwise change the retained pixel set, and a depth-validity switch can change the projection branch of a record that remains accepted. The next proposition accounts for both by coupling only records with controlled displacement. The plane is held fixed, matching the existing noise sweep; uncertainty in an estimated wall plane is not included.

Proposition N.6 (Changed valid projections and renormalization)**.** *Let ρ be the same probability measure on source records before geometric validity tests. Let A and $\widehat{A}$ be the clean and perturbed accepted sets with $a = \rho(A) > 0$, $\widehat{a} = \rho(\widehat{A}) > 0$. Normalize on each accepted set and project, obtaining*

$$\mu = U_{\#}\big(\rho|_A/a\big), \qquad \widehat{\mu} = \widehat{U}_{\#}\big(\rho|_{\widehat{A}}/\widehat{a}\big).$$

Suppose all output points lie in a common region of diameter D. Let $C \subseteq A \cap \widehat{A}$ be a measurable set on which $\|U(p) - \widehat{U}(p)\|_2 \le e$. For example, use records whose projection branch remains unchanged and satisfies the preceding geometric conditions. Writing $c = \rho(C)$ and $\eta = 1 - c/\max(a, \widehat{a})$, we have

$$W_1(\mu, \widehat{\mu}) \le (1-\eta)e + D\eta. \tag{11}$$

If $a, \widehat{a} \ge m_0 > 0$, this implies

$$W_1(\mu, \widehat{\mu}) \le e + D \min\left\{1, \frac{\rho(A \triangle \widehat{A}) + \rho((A \cap \widehat{A}) \setminus C)}{m_0}\right\}.$$

For validity changes alone, take $C = A \cap \widehat{A}$; then the second mass term vanishes.

Proof. The two normalized source measures share the submeasure $\rho|_C/\max(a, \widehat{a})$, of mass $1 - \eta$. Couple this mass by matching the same source record. Couple the remaining equal masses η arbitrarily; their projected distance is at most D. This constructs a coupling of cost at most $(1-\eta)e + D\eta$. Furthermore,

$$\eta = \frac{\max(a, \widehat{a}) - c}{\max(a, \widehat{a})} \le \frac{\rho(A \triangle \widehat{A}) + \rho((A \cap \widehat{A}) \setminus C)}{m_0}, \qquad \eta \le 1.$$

Use $(1-\eta)e \le e$ for the last bound. The argument also holds in the space–time metric when D is a diameter in that metric. □

Corollary N.7 (Aggregation and centre of mass)**.** *If the same source weights and branch memberships are used in the clean and perturbed field, with surface mass α_{surf} and wall mass α_{wall} summing to one, then*

$$W_1(\mu, \widehat{\mu}) \le \alpha_{\mathrm{surf}} e_{\mathrm{surf}} + \alpha_{\mathrm{wall}} e_{\mathrm{wall}}.$$

For any matched mixture weights $\beta_j \ge 0$ with $\sum_j \beta_j = 1$,

$$W_1\left(\sum_j \beta_j \mu_j, \sum_j \beta_j \widehat{\mu}_j\right) \le \sum_j \beta_j W_1(\mu_j, \widehat{\mu}_j).$$

For normalized spatial fields, their centres of mass satisfy

$$\left\| \int x\, \mu(dx) - \int x\, \widehat{\mu}(dx) \right\|_2 \le W_1(\mu, \widehat{\mu}).$$

Proof. Split the integral in Proposition N.3 by branch. For mixtures, mix couplings of each pair with weights β_j, using arbitrarily near-optimal couplings and then taking the approximation error to zero. For centres of mass, any coupling $(X, \widehat{X})$ satisfies $\|\mathbb{E}X - \mathbb{E}\widehat{X}\|_2 \le \mathbb{E}\|X - \widehat{X}\|_2$; take the infimum. □

What the metric controls. Every L-Lipschitz spatial readout f obeys $|\int f\,d\mu - \int f\,d\widehat{\mu}| \le LW_1(\mu,\widehat{\mu})$, by the same coupling proof. Sharp region indicators require boundary-mass control, as made explicit in Lemma N.14. Cosine of hard-binned fields and thresholded hot-set IoU are not automatically controlled by W_1: a point can cross a bin boundary under an arbitrarily small displacement, and nearly tied bin intensities can change threshold membership. Their measured stability remains empirical. The centre-of-mass corollary applies directly to normalized continuous point fields; any later bin-centre snapping introduces its own displacement. Matched aggregation does not require independent geometric errors, and it does not imply that systematic localization error decays as $n^{-1/2}$. In particular, the wall's large contribution to the combined field must not be interpreted as a universal guarantee for scenes without a known wall.

N.4 Mirror-calibrated behavioral bounds

For the state, output, and reflection maps from Appendix N.1, define the uniform defects

$$\varepsilon_{\mathrm{plan}} := \sup_{k,s} d_{\mathrm{TV}}\big(M_{y\#}\Pi_k(\cdot \mid s), \Pi_k(\cdot \mid M_s s)\big), \tag{12}$$

$$\varepsilon_{\mathrm{phys}} := \sup_{k,s,y} d_{\mathrm{TV}}\big(M_{s\#}P_k(\cdot \mid s,y), P_k(\cdot \mid M_s s, M_y y)\big). \tag{13}$$

The second defect conditions on the entire planner output, including generated memory; it describes the remainder of the loop. Bounds restricted to a set of states require that both compared processes remain in that set, or a separate escape-probability term.

Proposition N.8 (Mirror calibration with generated memory)**.** *Let the original initial law be ζ_0 and the mirrored initial law be $M_{s\#}\zeta_0$. Put $\bar{\varepsilon}_P = \varepsilon_{\mathrm{phys}} + \varepsilon_{\mathrm{plan}}$. At every $k \le H$,*

$$d_{\mathrm{TV}}\big(M_{s\#}\mathcal{L}(S_k), \mathcal{L}(\widetilde{S}_k)\big) \le \min\{1, k\bar{\varepsilon}_P\}.$$

Consequently, the terminal-event bound of Proposition 3 holds. For an additional bounded query-level task score $r_k(s,y) \in [0, R_{\max}]$, define

$$\varepsilon_r := \sup_{k,s,y} |r_k(M_s s, M_y y) - r_k(s,y)|, \quad \bar{\varepsilon}_r := \varepsilon_r + R_{\max}\varepsilon_{\mathrm{plan}}.$$

The corresponding returns satisfy

$$|J^{\Pi}(\zeta_0) - J^{\Pi}(M_{s\#}\zeta_0)| \le H\bar{\varepsilon}_r + \frac{R_{\max}H(H-1)}{2}\bar{\varepsilon}_P. \tag{14}$$

Proof. Define the policy-averaged kernel and score

$$Q_k(\cdot \mid s) = \int P_k(\cdot \mid s,y)\Pi_k(dy \mid s), \qquad \bar{r}_k(s) = \int r_k(s,y)\Pi_k(dy \mid s).$$

Compare $M_{s\#}Q_k(\cdot \mid s)$ first to the mixture of $P_k(\cdot \mid M_s s, M_y y)$ under $\Pi_k(dy \mid s)$. The difference is at most $\varepsilon_{\mathrm{phys}}$ by convexity of TV. Changing the mixing law from $M_{y\#}\Pi_k(\cdot \mid s)$ to $\Pi_k(\cdot \mid M_s s)$ costs at most $\varepsilon_{\mathrm{plan}}$ by kernel contraction (Lemma N.13). Hence

$$\sup_s d_{\mathrm{TV}}(M_{s\#}Q_k(\cdot \mid s), Q_k(\cdot \mid M_s s)) \le \bar{\varepsilon}_P.$$

Apply Lemma N.13 recursively to the reflected original law and the mirrored law. Their initial difference is zero, so the difference at k is at most $k\bar{\varepsilon}_P$, capped by one. Since M_s is an involution, applying this TV bound to $M_s E$ gives the claimed terminal-event statement.

For scores, add and subtract the reflected-score mixture under the reflected output law. Physical score asymmetry contributes ε_r, and changing the output distribution contributes at most $R_{\max}\varepsilon_{\mathrm{plan}}$ by Lemma N.12. Thus $|\bar{r}_k(M_s s) - \bar{r}_k(s)| \le \bar{\varepsilon}_r$. Comparing state laws at query k contributes at most $R_{\max}k\bar{\varepsilon}_P$. Sum these bounds from $k=0$ to $H-1$. □

Corollary N.9 (Conditional Wasserstein refinement)**.** *Suppose the joint output space has a Polish metric d_Y, the relevant output laws have finite first moments, and, uniformly in k, s, y, y',*

$$d_{\mathrm{TV}}(P_k(\cdot \mid s,y), P_k(\cdot \mid s,y')) \le L_P d_Y(y,y'), \qquad |r_k(s,y) - r_k(s,y')| \le L_r d_Y(y,y').$$

Put $\varepsilon_W = \sup_{k,s} W_1(M_{y\#}\Pi_k(\cdot \mid s), \Pi_k(\cdot \mid M_s s))$. Then Proposition N.8 also holds with

$$\bar{\varepsilon}_P = \varepsilon_{\mathrm{phys}} + L_P\varepsilon_W, \qquad \bar{\varepsilon}_r = \varepsilon_r + L_r\varepsilon_W.$$

Proof. Couple the reflected and mirrored joint output laws. Convexity of TV and the assumed Lipschitz property bound the kernel-mixture difference by L_P times the coupling's expected d_Y distance; the reward argument uses L_r. Take the infimum over couplings, then repeat the preceding proof. □

Relation to the command-only formulation and the audits. When no independently generated memory is retained, the output can reduce to the continuous command A, with $d_Y(a, a') = \|a - a'\|_2$; the preceding corollary recovers the command-Wasserstein formulation. Otherwise memory must remain in the joint output, or be controlled by an additional equivariance assumption. Close deterministic trajectories need not be close in TV: distinct point masses have TV distance one. Thus the Wasserstein refinement requires a genuine kernel regularity assumption, and even the general TV bound may be numerically uninformative for an approximately symmetric deterministic simulator. We do not infer uniform TV constants from the finite scripted audits in Appendix J. Those audits support interpretation on tested commands. Neither bound claims that mirror consistency implies a semantically correct target choice.

N.5 Planner–controller performance decomposition

Fix a bounded task score and two query-level models on the same augmented state and output spaces: an ideal executor $(r_k^{\mathrm{id}}, P_k^{\mathrm{id}})$ and the actual controller-induced executor $(r_k^{\mathrm{ctrl}}, P_k^{\mathrm{ctrl}})$. Both use the same fixed planner kernel Π_k and the same initial law. The ideal executor is a specified comparison model; the following result does not assume that it is available or evaluated in the experiments. Use a bounded high-level evaluation score in $[0, R_{\max}]$; this assumption does not identify that score with the PPO training reward. Define

$$\epsilon_r^{\mathrm{ctrl}} = \sup_{k,s,y} |r_k^{\mathrm{id}}(s,y) - r_k^{\mathrm{ctrl}}(s,y)|, \qquad \epsilon_P^{\mathrm{ctrl}} = \sup_{k,s,y} d_{\mathrm{TV}}(P_k^{\mathrm{id}}(\cdot \mid s,y), P_k^{\mathrm{ctrl}}(\cdot \mid s,y)).$$

Proposition N.10 (Fixed-policy execution gap)**.** *For the finite-horizon returns of the same planner,*

$$|J_{\mathrm{id}}^{\Pi} - J_{\mathrm{ctrl}}^{\Pi}| \le H\epsilon_r^{\mathrm{ctrl}} + \frac{R_{\max} H(H-1)}{2}\epsilon_P^{\mathrm{ctrl}}.$$

If $J_{\mathrm{id}}^{\star}$ is the supremum of ideal-model return over a specified planner class containing Π, and $\Delta_{\mathrm{plan}} = J_{\mathrm{id}}^{\star} - J_{\mathrm{id}}^{\Pi}$, then

$$J_{\mathrm{id}}^{\star} - J_{\mathrm{ctrl}}^{\Pi} \le \Delta_{\mathrm{plan}} + H\epsilon_r^{\mathrm{ctrl}} + \frac{R_{\max} H(H-1)}{2}\epsilon_P^{\mathrm{ctrl}}.$$

Proof. Averaging over the same conditional planner law preserves the uniform score and kernel differences. By Lemma N.13, the state laws at query k differ in TV by at most $k\epsilon_P^{\mathrm{ctrl}}$. Adding and subtracting the ideal-model expected score under the other state law, Lemma N.12 bounds the score difference by $\epsilon_r^{\mathrm{ctrl}} + kR_{\max}\epsilon_P^{\mathrm{ctrl}}$. Sum from $k = 0$ to $H - 1$. For the second statement, use the exact identity

$$J_{\mathrm{id}}^{\star} - J_{\mathrm{ctrl}}^{\Pi} = \Delta_{\mathrm{plan}} + (J_{\mathrm{id}}^{\Pi} - J_{\mathrm{ctrl}}^{\Pi}) \le \Delta_{\mathrm{plan}} + |J_{\mathrm{id}}^{\Pi} - J_{\mathrm{ctrl}}^{\Pi}|.$$

□

This decomposition explains why reached-versus-drifted outcomes alone do not isolate planning error. It neither estimates the terms from saliency nor identifies a unique cause of an individual failed trajectory. The quantity Δ_{plan} is relative to the chosen score and comparison class, not an independent definition of semantic competence.

N.6 Supporting lemmas

Lemma N.11 (Bounded-scalar concentration)**.** *If independent $X_1, \ldots, X_n$ take values in an interval of length $B > 0$, then for $t > 0$,*

$$\Pr\left(\left|n^{-1}\sum_i (X_i - \mathbb{E}X_i)\right| \ge t\right) \le 2\exp(-2nt^2/B^2).$$

Proof. For one bounded variable X, let $h(\lambda) = \log \mathbb{E}e^{\lambda X}$. Differentiation is valid by boundedness, and $h''(\lambda)$ is the variance under exponential tilting. Any law on an interval of length B has variance at most $B^2/4$: its squared distance to the interval midpoint is at most $B^2/4$, and variance minimizes expected squared distance over constant centres. Thus the second derivative of $g(\lambda) = h(\lambda) - \lambda\mathbb{E}X$ is at most $B^2/4$. Since $g(0) = g'(0) = 0$, integration yields $g(\lambda) \le \lambda^2 B^2/8$. Independence and exponential Markov give

$$\Pr\left(\sum_i (X_i - \mathbb{E}X_i) \ge nt\right) \le \exp(-\lambda nt + n\lambda^2 B^2/8).$$

Optimize at $\lambda = 4t/B^2$, obtaining $\exp(-2nt^2/B^2)$. Apply the same argument to $-X_i$ and add the two tails. The case of zero interval length is deterministic. □

Lemma N.12 (Expectation difference in total variation)**.** *If f is measurable with range in an interval of length C, then $|\int f\,dP - \int f\,dQ| \le C d_{\mathrm{TV}}(P, Q)$.*

Proof. Subtract the lower endpoint and, if $C > 0$, divide by C. The resulting function g lies in $[0, 1]$. The layer-cake identity gives

$$\int g\,dP - \int g\,dQ = \int_0^1 \{P(g > t) - Q(g > t)\}\,dt,$$

whose absolute value is at most $d_{\mathrm{TV}}(P, Q)$. Multiply by C; if $C = 0$, the claim is immediate. □

Lemma N.13 (Kernel contraction and perturbation)**.** *For a Markov kernel K, $d_{\mathrm{TV}}(PK, QK) \le d_{\mathrm{TV}}(P, Q)$. If $\sup_s d_{\mathrm{TV}}(K(\cdot \mid s), \widehat{K}(\cdot \mid s)) \le \epsilon$, then $d_{\mathrm{TV}}(PK, Q\widehat{K}) \le d_{\mathrm{TV}}(P, Q) + \epsilon$.*

Proof. For each event E, the function $s \mapsto K(E \mid s)$ lies in $[0, 1]$. Lemma N.12 and a supremum over E prove contraction. Also, for every E, $|(QK)(E) - (Q\widehat{K})(E)| \le \int |K(E \mid s) - \widehat{K}(E \mid s)|Q(ds) \le \epsilon$. The triangle inequality proves the second statement. This argument also proves the convexity-of-mixtures estimate used in the main proofs. □

Lemma N.14 (Region boundaries and ratios)**.** *Suppose a coupling of normalized spatial fields satisfies $\|X - \widehat{X}\|_2 \le e$ almost surely. For a Borel region A, set*

$$A^{-e} = \{x \in A : \mathrm{dist}(x, A^c) > e\}, \qquad A^{+e} = \{x : \mathrm{dist}(x, A) \le e\}.$$

Then

$$\mu(A^{-e}) \le \widehat{\mu}(A) \le \mu(A^{+e}), \qquad |\mu(A) - \widehat{\mu}(A)| \le \mu(A^{+e} \setminus A^{-e}).$$

If left and right region masses each change by at most η_L, η_R, and both right masses are at least $b_0 > 0$, then

$$\left|\frac{\mu(A_L)}{\mu(A_R)} - \frac{\widehat{\mu}(A_L)}{\widehat{\mu}(A_R)}\right| \le \frac{\eta_L}{b_0} + \frac{\eta_R}{b_0^2}.$$

Proof. If $X \in A^{-e}$ then $\widehat{X} \in A$; if $\widehat{X} \in A$ then $X \in A^{+e}$. Taking probabilities proves the sandwich. Since $A^{-e} \subseteq A \subseteq A^{+e}$, it also gives the absolute-difference bound. For masses $p, q, \widehat{p}, \widehat{q}$, add and subtract $\widehat{p}/q$ to obtain

$$|p/q - \widehat{p}/\widehat{q}| \le |p - \widehat{p}|/q + \widehat{p}\,|q - \widehat{q}|/(q\widehat{q}) \le \eta_L/b_0 + \eta_R/b_0^2.$$

Here $\widehat{p} \le 1$ because the field is normalized. □

Lemma N.15 (Preservation of retained mass)**.** *For a finite nonnegative measure ν on retained records and a measurable projection U defined on all of them, $(U_{\#}\nu)(\mathbb{R}^3) = \nu(\text{retained records})$.*

Proof. By definition, $(U_{\#}\nu)(\mathbb{R}^3) = \nu(U^{-1}(\mathbb{R}^3))$, and that preimage is the entire retained domain. □

This statement concerns the geometric map alone. Taking absolute values, thresholding pixels, changing normalization, dropping invalid projections, or discretizing the IG path can alter the attribution values or mass. Preserving retained nonnegative mass does not establish the exact signed IG completeness identity or causal faithfulness. The same distinction applies to alternative attribution methods used with the projection stage.