AVATARVLN: AGENTIC WORLD MODELING FOR VISION-AND-LANGUAGE NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models provide a promising way for vision-and-language navigation (VLN) agents to anticipate the consequences of candidate actions before physically executing them. However, imagined futures are not inherently actionable. A synthetic trajectory may span multiple instruction-relevant events, omit events that would occur, fabricate events that would not, and provide increasingly unreliable geometry over long rollouts. We present **AvatarVLN**, an agentic world modeling framework that turns imperfect imagined futures into actionable evidence for VLN. Rather than relying on a task-specific evaluator, AvatarVLN leverages the reasoning capability of the navigation agent itself. The agent decomposes an instruction into ordered events and tracks its progress through them. Specifically, we introduce **Anticipated Progress Grounding (APG)**, which models navigation progress as a latent state. Rather than committing to a single progress estimate, APG weights possible progress states according to real history and imagined rollouts to estimate the reduction in remaining navigation cost. In addition, this formulation uses instruction order to accommodate omitted intermediate events while discounting unsupported progress suggested by fabricated content. Beyond semantic grounding, we introduce **Uncertainty-aware Geometric Grounding (UGG)**, which estimates the reliability of imagined depth from accumulated motion along each rollout and selectively projects trustworthy geometry into candidate maps for spatial reasoning. The world model is trained on a navigation corpus derived from VLNVerse and transfers zero-shot to unseen VLN benchmarks. Experiments on three VLN benchmarks demonstrate the effectiveness of AvatarVLN, with ablations showing consistent gains on success rate over the base agents on R2R-CE.

1 INTRODUCTION

Imagine a robot navigating a 3D environment given a natural-language instruction (Anderson et al., 2018; Krantz et al., 2020). To understand where different actions may lead and what landmarks may appear along the way, the robot would normally need to physically explore alternative routes. A world model offers an appealing alternative. It allows the robot to imagine multiple possible futures, as if deploying virtual avatars to explore different routes without physical execution. Given a candidate action, the world model can generate a sequence of future observations depicting what the robot may encounter along that route (Koh et al., 2021; Bar et al., 2025; Zhou et al., 2025). Yet these imagined futures do not explicitly reveal how they correspond to the ordered intent expressed by the navigation instruction. More importantly, generating plausible futures is not equivalent to knowing which action to take. How can an agent convert imperfect imagined trajectories into actionable evidence about its progress toward an instruction?

Existing approaches ground imagined futures using task-specific scorers (Koh et al., 2021; Wang et al., 2023; 2024), direct goal matching (Bar et al., 2025; Huang et al., 2025a;b), or textual narration (Wang et al., 2025b). The first relies on additional supervision and external evaluators, the second largely reduces navigation intent to compatibility with a single goal, and the third compares narrated rollouts only with the current sub-instruction. None explicitly models how far an imagined trajectory advances through the ordered events of an instruction, even though one rollout may span several of them. Moreover, imagined futures are inherently **imperfect** (Koh et al., 2021; Zhou et al., 2025; Zhu et al., 2026). A world model may omit events that would occur or

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.

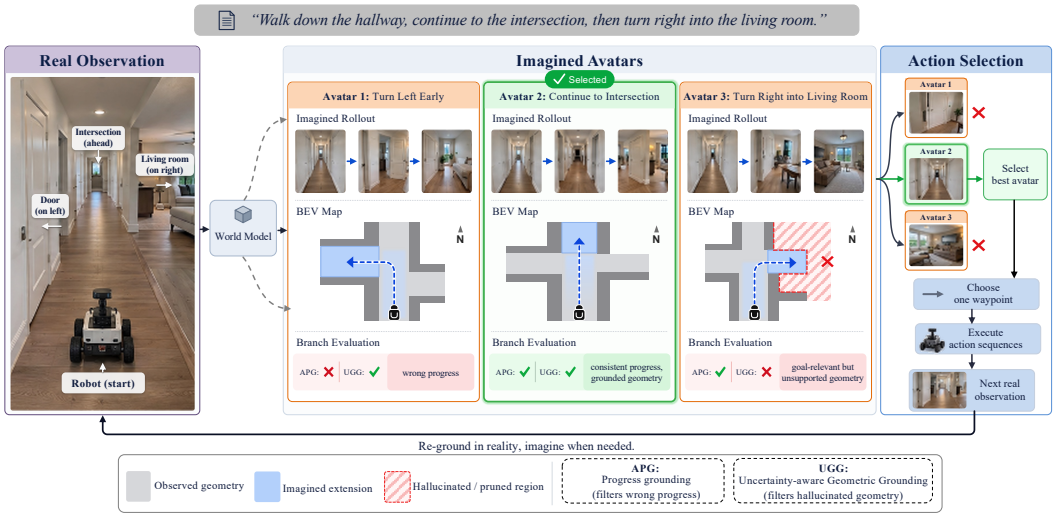


Figure 1: **AvatarVLN: grounding imperfect imagination into action**. A world model generates imagined rollouts for candidate action sequences, as if deploying virtual avatars to explore alternative futures. To interpret these imperfect imaginations, APG (Sec. 3.3) evaluates progress through the instruction's ordered events, while UGG (Sec. 3.4) calibrates and filters imagined geometry. They provide grounded evidence for the frozen agent to select the waypoint that best advances the instruction. Imagination remains a plug-and-play, optional tool. The agent decides when to invoke it and re-grounds in real observations after each move (Sec. 3.5).

fabricate instruction-relevant events that would not (Li et al., 2025b), requiring progress inference to be tolerant to missing evidence yet conservative against spurious progression. This uncertainty extends beyond semantics to geometry, as imagined depth becomes increasingly unreliable over longer rollouts and more complex motions. Turning imagination into action therefore requires grounding imagined futures along two complementary dimensions, *i.e.*, their semantic progress with respect to the navigation instruction, and their geometric reliability with respect to the physical environment.

To address these challenges, we present **AvatarVLN** (Fig. 1), an agentic world modeling framework that turns imagined futures into actionable evidence for navigation. Intuitively, the world model spawns multiple virtual avatars, each exploring a possible future, while the navigation agent must decide which avatar is worth following. Our key idea is to let the agent itself reason about which imagined future places it further along the instruction, thereby reducing the remaining navigation burden. To this end, we introduce **Anticipated Progress Grounding (APG)**, which decomposes the instruction into ordered events and models navigation progress as a latent state (Icarte et al., 2018; Li et al., 2024) inferred from imagined observations. Rather than scoring each frame independently, APG aggregates evidence along the instruction progression to estimate how much progress each avatar is expected to achieve. Importantly, the formulation is designed around the imperfect nature of world-model rollouts, remaining tolerant to omitted intermediate events while conservatively suppressing progress supported by potentially fabricated evidence. Complementing semantic progress, we introduce **Uncertainty-aware Geometric Grounding (UGG)** to determine which imagined geometry can be trusted. UGG calibrates imagined depth against the agent's current real observation and accounts for uncertainty induced by rollout horizon and motion complexity, allowing only sufficiently reliable geometry to contribute to the agent's spatial memory.

We further construct a navigation world-model training corpus from the VLNVerse training split (Lin et al., 2025) and train a world model on it to support agentic imagination for VLN. Despite seeing only VLNVerse during training, the resulting world model transfers zero-shot to unseen navigation benchmarks, including R2R (Anderson et al., 2018) and RxR (Ku et al., 2020). For example, without training any task-specific decision module, AvatarVLN reaches a 52% success rate on the full R2R-CE `val_unseen` split with an open-weight Qwen3.8-27B agent (Qwen Team, 2026), on par with zero-shot methods built on GPT-5-family models (Li et al., 2026c;a). These results suggest that the key to exploiting a world model for agentic navigation is not simply generating more plausible futures, but grounding imperfect imagination into reliable evidence for action. Our contributions are summarized as follows:

- We introduce AvatarVLN, an agentic world modeling framework for VLN, and propose Anticipated Progress Grounding (APG) to infer latent navigation progress by aligning imagined trajectories with ordered instruction events. Unlike task-specific evaluators, APG leverages the reasoning capability of the navigation agent itself and is designed to tolerate omitted events while remaining conservative against fabricated progress.
- We propose Uncertainty-aware Geometric Grounding (UGG), which calibrates imagined depth with real observations and filters hallucinated imagined geometry, allowing only reliable geometry to enter the agent's spatial memory.
- We construct a navigation world-model training corpus from VLNVerse and show that a world model trained solely on this corpus transfers zero-shot to unseen benchmarks including R2R and RxR. AvatarVLN consistently improves a frozen open-source navigation agent without training any task-specific decision module.

2 Related Work

From specialist policies to agentic navigation. Vision-and-language navigation (VLN) (Anderson et al., 2018) requires an agent to follow natural-language instructions through visual environments. Early approaches such as Seq2Seq and CMA (Krantz et al., 2020) learn specialist policies. Recent work increasingly builds navigation systems on pretrained vision-language models. NaVILA (Cheng et al., 2024), StreamVLN (Wei et al., 2025b), and DualVLN (Wei et al., 2025a) adapt broad vision-language priors for embodied control, while NavFoM (Zhang et al., 2025) scales this paradigm across navigation tasks and embodiments. A complementary line uses frozen foundation models, organizing their reasoning through maps, action proposals, and instruction constraints (Chen et al., 2024b; Qiao et al., 2024; Chen et al., 2024a;c; Yin et al., 2025). More recently, GTA (Li et al., 2026c) provides an interactive metric world representation, and AgenticNav (Li et al., 2026b) exposes action, depth, and memory as tools that the agent can call as needed. AvatarVLN follows this agentic direction, adding imagination as an optional tool and grounding imagined outcomes for a frozen agent, rather than learning a new navigation policy.

World models are not new to VLN. Early works use predicted future observations or features for lookahead planning, including Pathdreamer (Koh et al., 2021) and HNR (Wang et al., 2024). In broader visual navigation, world models are often used to roll out candidate trajectories and rank them by goal compatibility (Bar et al., 2025; Zhou et al., 2025; Zhu et al., 2026). For language-guided navigation, recent methods use imagined goals, views, or trajectories to score or select candidate actions (Huang et al., 2025b;a; Zhao et al., 2025; Wang et al., 2025a;b), with related ideas appearing in spatial reasoning and autonomous driving (Yang et al., 2025; Wang et al., 2026). More recently, UniWM (Dong et al., 2025), WAM-Nav (Yang et al., 2026), UniNav (Zhou et al., 2026), and WorldVLN (Zhao et al., 2026) jointly integrate world prediction and action generation within trained navigation models. While visual imagination itself is well explored, many existing approaches rely on goal-specific matching, learned decision modules, or task-specific navigation models to turn imagined futures into actions. Our focus is instead on how a frozen navigation agent can structurally interpret imperfect rollouts against the ordered intent of an instruction. APG infers latent progress while mitigating *omitted* and *fabricated* evidence, while UGG filters *hallucinated* geometry by grounding imagined structure in real observations and motion-dependent reliability. This requires neither a goal image nor an additional learned task-specific ranking or decision network.

3 Method

AvatarVLN equips a frozen navigation agent, *e.g.*, Qwen3.8-27B or GPT-5.5, with imagined alternatives to physical exploration. The world model predicts the observations along candidate routes. Anticipated Progress Grounding (APG) interprets their consequences for the instruction, while Uncertainty-aware Geometric Grounding (UGG) determines which predicted geometry can support spatial reasoning. The agent uses both forms of evidence to choose its next move.

3.1 Problem formulation and overview

Given a navigation instruction I, the agent observes an RGB-D panorama o_t and its pose at decision step t. Let $\mathcal{H}_t$ denote its real observation and action history, and $\mathcal{M}_t^{\text{real}}$ its accumulated spatial

memory. A waypoint proposer supplies candidates $\mathcal{A}_t$; the agent can move to a waypoint, stop, or request imagination for a subset $\mathcal{B}_t \subseteq \mathcal{A}_t$. Each requested candidate b specifies a local action sequence and hence a nominal camera trajectory $P^b_{0:H_b}$. Conditioned on the current view facing that candidate, the world model generates

$$\hat{\tau}^b = \{(\hat{x}^b_k, \hat{D}^b_k)\}_{k=0}^{H_b} \sim W_\theta\big(\cdot \mid x^b_t, D^b_t, P^b_{0:H_b}\big), \tag{1}$$

where x^b_t, D^b_t are measured RGB and depth, and $k = 0$ is the conditioning view. Different waypoint distances yield different effective horizons H_b. APG therefore reasons over the instruction progress supported by each rollout, rather than comparing frames at matched temporal indices. UGG produces a corresponding candidate BEV map $\hat{\mathcal{M}}^b_t$.

3.2 Navigation world model

We construct a navigation prediction corpus from the VLNVerse training split (Lin et al., 2025), recording egocentric RGB, metric depth, and camera poses along navigation trajectories. Temporally sampled clips provide aligned RGB-D targets and camera control. Our predictor uses the joint RGB-D video generation implementation of MemoryWorld (Zhou et al., 2025), built on CogVideoX-2B (Yang et al., 2024). Note that AvatarVLN does not aim to design a new world-model architecture. Instead, we focus on grounding its imperfect predictions into actionable evidence for navigation. As evidenced in Tab. 3 of Appendix A.1, with the same world model, imagination alone improves SR by only $+3$ to $+4$ points, whereas the proposed APG and UGG raise the gain to $+10$ to $+12$ SR. The world model is trained once on VLNVerse and reused on R2R-CE and RxR-CE without adaptation. Training details appear in Appendix A.5.

3.3 Anticipated Progress Grounding

Ranking rollouts by remaining cost. Given the instruction I, real history $\mathcal{H}_t$, and an imagined rollout $\hat{\tau}^b$, we seek to determine how much navigation work this candidate would leave unfinished. Let $C^*_I(h)$ denote the ideal minimum expected navigation cost needed to complete I from a history h. The ideal anticipated progress is therefore

$$\Delta P^*_b = C^*_I(\mathcal{H}_t) - C^*_I(h_b), \tag{2}$$

where $h_b = \mathcal{H}_t \oplus \hat{\tau}^b$ and $\oplus$ is the concatenation operator. A larger ΔP^*_b indicates a future from which less navigation work remains.

Event-by-event tracking meets hallucinations. Estimating remaining cost requires inferring how much of the instruction has already been satisfied. An intuitive approach is to split the instruction into ordered *events*, *i.e.*, route milestones that mark progress along the instruction, such as passing a landmark or entering a room, and track them one by one. However, imagined rollouts can be imperfect, *omitting* or *fabricating* such events. For "pass the sofa, turn left after the table, and enter the kitchen," an omitted sofa may prevent the tracker from recognizing progress even when subsequent frames support the turn and room entry. Conversely, fabricated evidence of passing the sofa may cause premature advancement. Thus, committing to a single discrete event boundary can make progress estimates brittle under imperfect imagination.

Latent progress. We treat instruction progress as a latent state inferred from evidence across the imagined trajectory together with the order specified by the instruction. Rather than committing to a single progress state, we assign belief weights over possible states supported by real history, imagined observations, and instruction order (Kaelbling et al., 1998). We estimate the remaining navigation cost under each state and aggregate these costs using the belief weights.

To represent this state, the frozen navigation agent itself decomposes the instruction text I into ordered events $\mathcal{E}_I = (e_1 \prec \cdots \prec e_K)$ once per episode. Each event e_k is typically an action anchored to a landmark or region, such as *passing the sofa*, *turning left after the table*, and *entering the kitchen*. The latent state $r \in \{0, \ldots, K\}$ represents how many ordered events are believed to have been completed. In state r, the first r events are considered satisfied and e_{r+1} is next, with $r = K$ denoting completion. In the example above, $r = 1$ means that the sofa has been passed, while turning left after the table is the next event and entering the kitchen remains further ahead. Importantly, a progress state does not require explicit evidence for every preceding event, as discussed below.

Remaining work conditioned on progress. For a progress state r, we approximate the ideal remaining cost $C_I^*(h)$ (Eq. (2)) with the following state-conditioned surrogate:

$$C(h, r) = \begin{cases} d(h, e_{r+1}) + K - r - 1, & r < K, \\ 0, & r = K, \end{cases} \tag{3}$$

where $d(h, e_{r+1}) \in [0, 1]$ measures the relative work remaining to complete the next event, and $K - r - 1$ counts the events that remain afterward. For a given state r, the agent estimates d from history h and each avatar's imagined rollout with respect to the same next event e_{r+1}. Rather than metric distance, d expresses relative proximity to completing that event across candidate rollouts, with smaller values indicating less remaining work.

In the running example, $K = 3$ and $r = 1$ give $C(h, 1) = d(h, e_2)+1$, *i.e.*, the sofa has already been passed, $d(h, e_2)$ measures the work remaining to complete the left turn after the table, and the final unit accounts for entering the kitchen. A rollout can therefore reduce remaining cost by bringing the agent closer to completing the turn, even before the progress state advances to $r = 2$.

Inferring progress from imperfect imagination. Progress states should contribute according to how well they are supported by the imagined evidence. We represent this support with a belief weight $q_k^b(r)$, the agent's confidence that exactly the first r events have been completed after k imagined steps. Real history provides the initial belief $q_t(r)$, shared by all candidate rollouts:

$$q_k^b(r) \approx p\big(r_k^b = r \mid I, \mathcal{H}_t, \hat{x}_{1:k}^b\big), \qquad q_0^b = q_t. \tag{4}$$

where r_k^b denotes the progress state after k imagined steps of candidate b, and p is its posterior probability given the instruction, the real history, and the imagined frames. APG approximates this posterior with q_k^b. Rather than directly predicting numerical confidence, the agent first extracts *categorical evidence* for each event from the imagined observations: COMPLETED, IN-PROGRESS, NOT-YET, or NO-EVIDENCE. These judgments are combined with instruction order and temporal consistency to infer the belief weights. Motion-dependent events such as turns and passes require temporal evidence. An omitted landmark yields NO-EVIDENCE, whereas NOT-YET requires positive evidence that the event remains unsatisfied. Our belief-based formulation helps mitigate two types of error in imperfect imagination:

- **Omission.** Missing evidence does not imply that an event remains unfinished. For example, a coherent left turn followed by kitchen entry can support later progress states even when passing the sofa was not rendered. The agent assigns more conservative belief weights when intermediate events lack direct evidence, rather than ruling out progress solely because a landmark is omitted. Missing evidence alone cannot justify advancement, and explicit evidence of an unmet earlier constraint blocks it.
- **Fabrication.** Apparent evidence of a later event does not by itself justify a high belief in a later progress state. An isolated kitchen-like frame provides weak support for $r = 3$ unless the surrounding trajectory and real history also support completion of the preceding events in order. Evidence that conflicts with motion, geometry, or earlier observations is therefore down-weighted, limiting the effect of potentially fabricated content on the progress belief.

Which avatar makes more progress? For each candidate rollout, we weight the remaining cost of each progress state by its belief weight and sum over all states. The anticipated progress is the reduction in this weighted remaining cost:

$$\Delta_b = \sum_{r=0}^{K} q_t(r) C(\mathcal{H}_t, r) - \sum_{r=0}^{K} q_{H_b}^b(r) C(h_b, r). \tag{5}$$

The first term measures the work remaining before the rollout, and the second measures the work remaining afterward. A larger reduction indicates greater anticipated progress.

In practice, we decompose this reduction into two sources: progress toward completing the next event while holding the progress belief fixed (Δ_b^{within}), and progress from shifting belief toward states where more events have been completed ($\Delta_b^{\text{advance}}$). We then form a guarded comparison score:

$$S_b = \text{clip}\big(\Delta_b^{\text{within}}, -B_b, B_b\big) + \rho_b \Delta_b^{\text{advance}}. \tag{6}$$

Here B_b bounds within-event progress by what the candidate's motion can plausibly support on the surrogate cost scale, while $\rho_b \in [0, 1]$ discounts event advancement when its supporting evidence is uncertain. Larger S_b indicates a stronger progress-based preference. The score structures the agent's comparison of candidate rollouts rather than serving as a calibrated value estimate. Appendix A.6 details the decomposition and scoring safeguards.

3.4 Uncertainty-aware Geometric Grounding

Imagined depth is not metrically reliable by default. UGG calibrates it against real observations, weights it by accumulated motion, and filters geometrically inconsistent predictions.

You only anchor once. Given the measured depth at the current viewpoint, we propagate this *anchor* to successive imagined frames by projecting the accumulated real point cloud into each nominal camera pose P_k^b. This yields sparse reference depths $D_k^{b,\mathrm{ref}}$ over overlap regions Ω_k^b. Specifically, we fit a joint affine calibration whose scale and shift vary with the prediction step:

$$\tilde{D}_k^b(u) = a_b(k)\hat{D}_k^b(u) + b_b(k), \qquad a_b(k) = a_{b,0} + \alpha_b k, \quad b_b(k) = b_{b,0} + \beta_b k. \tag{7}$$

We jointly estimate the four parameters from reference-depth correspondences across the rollout.

Spatiotemporal weighting. Geometric reliability of world-model imagination tends to degrade as motion accumulates. Hence, we weight each imagined frame according to the translation and rotation accumulated since its most recent geometric anchor. For an anchored frame j, confidence c_j^b is estimated from its calibrated residuals and reconstruction quality. For a subsequent unanchored frame k, we propagate this confidence as

$$c_k^b = \mathrm{clip}_{[c_{\min},1]}\left[c_j^b \exp\left(-\lambda_L L_{j:k}^b - \lambda_\theta \Theta_{j:k}^b\right)\right], \tag{8}$$

where $L_{j:k}^b$ and $\Theta_{j:k}^b$ accumulate translation and rotation along the rollout, respectively. Using generated rollouts with reference depth, we fit the decay coefficients offline by regressing relative depth-quality degradation against cumulative translation and rotation using nonnegative Huber regression, yielding $\lambda_L = 0.424\,\mathrm{m}^{-1}$ and $\lambda_\theta = 0.619\,\mathrm{rad}^{-1}$. The coefficients remain fixed at deployment.

Hallucination filtering. We treat geometric contradiction, rather than lack of verification, as evidence of hallucination. Beyond observed regions, we use candidate rollout poses to check cross-view geometric consistency, rejecting surfaces contradicted by neighboring views when unsupported by any eligible neighbor. Regions without overlapping evidence are retained as speculative geometry and weighted by c_k^b, rather than discarded solely for being unverified.

After calibration, spatiotemporal weighting, and filtering, imagined geometry is fused into a candidate BEV map $\hat{\mathcal{M}}_t^b$ and overlaid on the real map. Real measurements always take precedence, while imagined layers remain local to the current decision and are rebuilt after each move. Appendix A.7 provides implementation details.

3.5 Agentic action selection

Imagination as a tool, not a constraint. Imagination is available to the frozen agent as an optional tool. APG guides the interpretation of imagined progress, while UGG supplies grounded geometric evidence. The agent decides when imagination is useful and whether structured comparison is needed. For an unambiguous move, it may act directly. When candidates compete or imagined evidence conflicts with the instruction, it can invoke structured APG comparison.

Grounded action selection. The agent selects actions from the instruction, real observation history, and available waypoints. When imagination is used, it additionally considers predicted RGB rollouts, APG progress evidence, and UGG-grounded candidate BEV maps, without introducing a learned decision head to combine them. The agent executes one waypoint move, discards the speculative branches, and re-grounds itself in the next real observation. Task completion is determined from real observations only.

4 Experiments

Benchmarks and metrics. AvatarVLN is evaluated on R2R-CE (Krantz et al., 2020), RxR-CE (Ku et al., 2020), and VLNVerse (Lin et al., 2025). We evaluate Qwen3.8-27B on the 100-episode R2R-

Table 1: **R2R-CE and RxR-CE `val_unseen`.** Results use the full split unless marked otherwise. § marks a 100-episode subset, which on R2R-CE is the subset of Qiao et al. (2024) and on RxR-CE a size-matched English subset that we sample, and ‡ marks results whose episode ids are not released.

Method	Backbone	R2R-CE					RxR-CE				
		NE↓	OSR↑	SR↑	SPL↑	nDTW↑	NE↓	OSR↑	SR↑	SPL↑	nDTW↑
Trained policies											
Seq2Seq (Krantz et al., 2020)	–	7.77	37	25	22.0	–	12.10	–	14	11.9	30.8
CMA (Krantz et al., 2020)	–	6.20	52	41	36.0	–	8.76	–	27	22.1	47.0
NaVILA (Cheng et al., 2024)	–	5.22	63	54	49.0	–	6.77	–	49	44.0	58.8
StreamVLN (Wei et al., 2025b)	–	4.98	64	57	51.9	–	6.22	–	53	46.0	61.9
JanusVLN (Zeng et al., 2026)	–	4.78	65	61	56.8	–	6.06	–	56	47.5	62.1
DualVLN (Wei et al., 2025a)	–	4.05	71	64	58.5	–	4.58	–	61	51.8	70.0
Zero-shot agents											
MapGPT-CE§ (Chen et al., 2024b)	GPT-4o	8.16	21	7	5.0	27.4	–	–	–	–	–
Open-Nav§ (Qiao et al., 2024)	GPT-4	6.70	23	19	16.1	45.8	–	–	–	–	–
CA-Nav (Chen et al., 2024c)	GPT-4	7.58	48	25	10.8	–	10.37	–	19	6.0	13.5
AO-Planner (Chen et al., 2024a)	Gemini-1.5-Pro + GPT-4o	6.95	38	26	16.6	–	10.75	–	22	15.1	33.1
SmartWay§ (Shi et al., 2025)	GPT-4o	7.01	51	29	22.5	–	–	–	–	–	–
HiMemVLN§ (Lyu et al., 2026)	Qwen2-VL-72B	6.65	36	30	26.9	52.8	–	–	–	–	–
InstructNav‡ (Long et al., 2024)	GPT-4 + GPT-4V	6.89	–	31	24.0	–	–	–	–	–	–
DreamNav‡ (Wang et al., 2025b)	GPT-4o + Qwen-VL-Max	7.06	41	33	29.0	–	–	–	–	–	–
GC-VLN (Yin et al., 2025)	DeepSeek-R1	7.30	42	34	16.3	–	8.80	44	34	13.8	–
Three-Step Nav§ (Zheng et al., 2026)	GPT-5	5.87	39	34	29.1	57.7	9.21	–	22	16.1	45.7
STRIDER§ (He et al., 2025)	Qwen-VL-Max + GPT-4o	6.91	39	35	30.3	51.8	11.19	–	21	9.6	30.1
LightZeroNav§ (Luo et al., 2026)	Qwen3-VL-32B	7.65	51	38	22.0	–	–	–	–	–	–
BZS-VLN (Li et al., 2025a)	GPT-5-mini	6.12	55	41	25.4	–	7.56	–	36	21.7	42.4
HSGM‡ (Li et al., 2026a)	GPT-5	5.42	59	48	32.8	–	7.43	–	42	25.1	54.9
GTA (Li et al., 2026c)	GPT-5.1	4.95	56	49	41.8	60.4	**6.29**	–	46	**39.3**	**57.4**
P2DNav§ (Sheng et al., 2026)	Qwen3-VL-32B	6.61	65	50	30.6	–	–	–	–	–	–
AgenticNav§ (Li et al., 2026b)	GPT-5.5	5.19	65	55	**48.4**	**63.4**	–	–	–	–	–
AvatarVLN (ours)											
AvatarVLN§	Qwen3.8-27B	5.27 ±0.24	66 ±3	51 ±2	34.3 ±1.8	53.4 ±2.6	9.12 ±0.47	44 ±3	34 ±3	24.7 ±2.1	41.5 ±3.0
AvatarVLN	Qwen3.8-27B	5.23	67	52	34.9	48.1	9.31	43	33	24.2	40.9
AvatarVLN	Opus 5	4.61	**76**	**60**	39.7	57.2	6.64	53	50	36.1	51.9
AvatarVLN	GPT-5.5	**4.50**	72	58	35.5	53.8	6.38	**55**	**52**	38.6	53.4

CE `val_unseen` subset introduced by Qiao et al. (2024) and a size-matched English subset of RxR-CE. We additionally evaluate all three agents (Qwen3.8-27B, Opus 5, and GPT-5.5) on both full `val_unseen` splits. For Qwen3.8-27B, we report the mean and half-range over three runs on each 100-episode subset. On VLNVerse, we evaluate AvatarVLN on the fine-grained test split following (Lin et al., 2025). We report navigation error (NE, in metres), oracle success rate (OSR), success rate (SR, stopping within 3 m of the goal), success weighted by path length (SPL), and normalised dynamic time warping (nDTW).

Agents and implementation. Our base agent follows MapGPT (Chen et al., 2024b). We use SmartWay's waypoint predictor (Shi et al., 2025) on R2R-CE and RxR-CE, and train a waypoint predictor on VLNVerse. Each episode is limited to 30 waypoint moves or 40 minutes. Imagination is a plug-and-play module with two scoring variants. Forced APG always invokes the scoring function when imagination is used, whereas adaptive APG† lets the agent decide whether scoring is needed. The world model is the only trained component added by the imagination module (Sec. 3.2).

4.1 Comparison with prior work

R2R-CE. In Table 1, AvatarVLN achieves the highest reported SR (60) and OSR (76) among zero-shot methods with Opus 5, and the lowest NE (4.50 m) with GPT-5.5. Using GPT-5.5, AvatarVLN reaches 58 SR and 72 OSR on the full split, compared with AgenticNav's 55 SR and 65 OSR on its 100-episode subset. With Qwen3.8-27B, AvatarVLN achieves 52 SR on the full split, exceeding the reported SR of open-weight baselines with larger backbones. P2DNav and LightZeroNav use Qwen3-VL-32B (50 and 38 SR), while HiMemVLN uses Qwen2-VL-72B (30 SR). It also exceeds some GPT-5-class methods in SR, such as GTA and HSGM (49 and 48). Without fine-tuning the navigation agent, the Opus 5 variant approaches the trained JanusVLN policy in SR (60 vs. 61).

RxR-CE. AvatarVLN achieves 52 SR with GPT-5.5 and 50 SR with Opus 5, both exceeding the best prior zero-shot result of 46 SR reported by GTA. Qwen3.8-27B achieves 33 SR on the full split, compared with 52 on R2R-CE. The larger performance gap between Qwen3.8-27B and the proprietary agents highlights the challenge of longer, more detailed instructions.

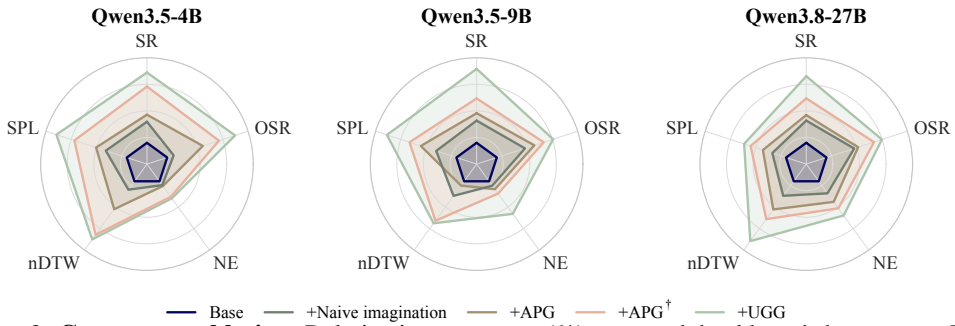


Figure 2: **Component ablation.** Relative improvement (%) over each backbone's base agent. NE is plotted as a relative reduction. Radial scales are normalised independently per panel.

Table 2: Results on the VLNVerse fine-grained test split.

Agent	NE↓	OSR↑	SR↑	SPL↑	nDTW↑
Trained navigation foundation models					
InternVLA-N1 (Intern Robotics, 2025)	5.68	39	29	25.0	–
UniNavid (Zhang et al., 2024)	4.11	67	46	26.9	–
NavFoM (Zhang et al., 2025)	4.13	70	52	32.4	–
Qwen-RobotNav-4B (Zhang et al., 2026a)	3.05	70	63	56.2	–
Qwen-RobotNav-8B (Zhang et al., 2026a)	3.00	73	64	57.9	–
Zero-shot benchmark agents					
Text-history (Zhou et al., 2024)	6.10	61	19	8.1	52.6
Map-history (Chen et al., 2024b)	5.62	55	26	7.6	43.4
AvatarVLN (ours)					
AvatarVLN (Opus 5)	**3.65**	**76**	59	37.8	61.5
AvatarVLN (GPT-5.5)	3.83	73	**61**	**41.5**	**63.0**

VLNVerse. In Table 2, AvatarVLN achieves 59 SR and the highest OSR (76) with Opus 5, while GPT-5.5 reaches 61 SR. Both variants exceed the reported SR of the benchmark's zero-shot agents (19 and 26) and several trained navigation foundation models. With its navigation agent frozen, AvatarVLN approaches the reported SR of Qwen-RobotNav (Zhang et al., 2026a) (59–61 vs. 63–64), which is fine-tuned end-to-end on 15.6M navigation and vision-language samples.

4.2 Ablation

Fig. 2 shows relative improvement over the base agent as we introduce each component across three backbones. Overall, each component promotes base agent's navigation performance. *Naive imagination*, which directly selects among imagined rollouts, provides modest gains. *Forced APG* increases these gains while *adaptive APG*† improve performance by allowing the agent to decide whether scoring is needed. With UGG grounding imagined geometry, the complete AvatarVLN achieves the strongest performance across all five metrics and all three backbones. Tab. 3 of Appendix A.1 provides the absolute scores and paired comparisons.

4.3 Case studies

We analyse how the base agent and +APG† behave on the same episodes (Appendix A.8 provides more examples) and summarise our findings in four remarks.

Remark 1. *Imagination supplies what the panorama cannot show.* Candidates that look alike from the current viewpoint can lead to different places, and imagining them reveals where each one goes (Fig. 3). The benefit depends on which candidates the agent imagines, as the imagination-only agent never requests the rollout that would have corrected it.

Remark 2. *A rollout is only as useful as the agent's judgement of it.* GPT-5.5 recognises rollouts that stall or run into a wall and changes its choice, whereas a 9B agent can read an arrival into a nearly static rollout.

Remark 3. *Event order keeps imagination aligned with the instruction.* Reading rollouts against the ordered events of the instruction lets APG reject a candidate that heads for a visible goal while

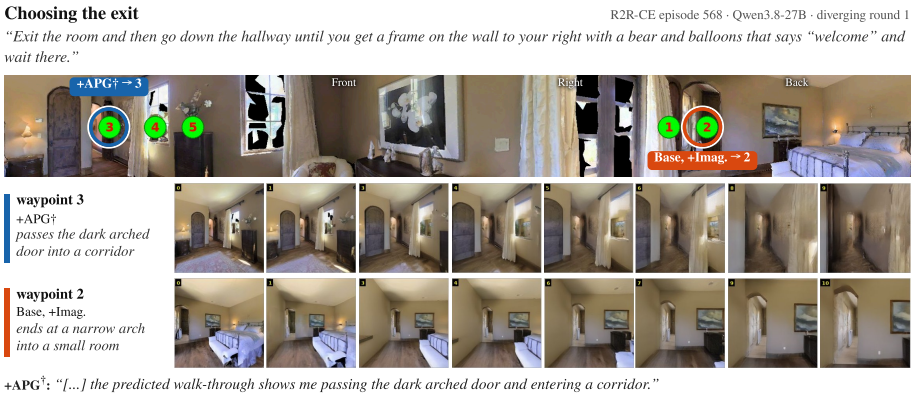


Figure 3: **Imagining beyond the doorway.** The two exits look alike in the panorama, but only the rollout of waypoint 3 shows a corridor. +APG† imagines it and reaches the goal (NE 0.6 m), whereas the base agent and the imagination-only agent, which never imagines waypoint 3, both end in a bathroom (NE 14.5 m and 12.7 m).

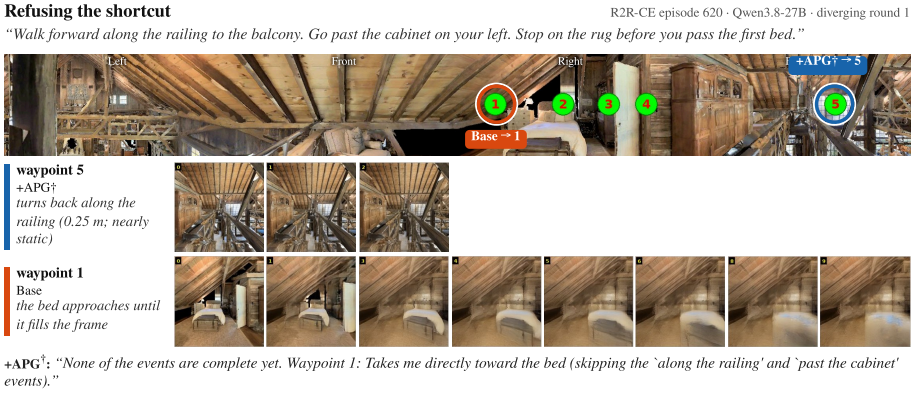


Figure 4: **Progress before proximity.** The goal bed is visible at waypoint 1, but the instruction first requires the railing and the cabinet. +APG† rejects waypoint 1 for skipping these events and reaches the goal (NE 1.8 m), while the base agent heads for the bed and stops too early (NE 6.3 m).

skipping earlier events (Fig. 4), and keeps an agent at the foot of a staircase from believing it has reached the top. Small agents still sometimes stop on an imagined arrival, which the protocol forbids.

Remark 4. *Imagination costs time.* Every request adds a decision round, and several episodes that only the base agent solves are lost to the wall-clock limit while +APG† is within 3 m of the goal.

These effects pull in opposite directions. Imagination alone gives only a small gain in Tab. 3 of Appendix A.1, which grows substantially once APG† structures how rollouts are read.

5 Conclusion

We presented AvatarVLN, an agentic world modeling framework that lets a *frozen* VLN agent imagine the consequences of candidate actions before physically executing them. Rather than designing a new world model, we focus on grounding these imperfect predictions as actionable evidence. APG estimates latent progress through the instruction's ordered events, while UGG calibrates, weights, and filters imagined geometry. As a result, the agent can compare candidates without a separately trained scorer. The world model is trained once on VLNVerse and transfers zero-shot to R2R-CE and RxR-CE. We evaluate AvatarVLN on three benchmarks and three backbones, with ablations showing consistent gains over base agents. For example, on R2R-CE, the open-weight Qwen3.8-27B variant surpasses several GPT-5-class agents in success rate. Imagination incurs additional inference time (Appendix A.4), motivating future work on faster generation and more selective use of imagined rollouts. A real-world robot demonstration is provided in the Supplementary Material.

AI USE STATEMENT

Large language models are used as the navigator in our system, as described in Section 3. We also used large language models to polish the writing of the manuscript. The authors have verified all AI-assisted text and assume full responsibility for the final content, claims, and results.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018. URL https://arxiv.org/abs/1711.07280.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.03572.

Neel P. Bhatt, Yunhao Yang, Rohan Siva, Pranay Samineni, Daniel Milan, Zhangyang Wang, and Ufuk Topcu. VLN-Zero: Rapid exploration and cache-enabled neurosymbolic vision-language planning for zero-shot transfer in robot navigation. *arXiv preprint arXiv:2509.18592*, 2025.

Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*, 2024a.

Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. MapGPT: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *ACL*, 2024b. doi: 10.18653/v1/2024.acl-long.529.

Kehan Chen, Dong An, Yan Huang, Rongtao Xu, Yifei Su, Yonggen Ling, Ian Reid, and Liang Wang. Constraint-aware zero-shot vision-language navigation in continuous environments. *arXiv preprint arXiv:2412.10137*, 2024c.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

Yifei Dong, Fengyi Wu, Guangyu Chen, Lingdong Kong, Qiyu Hu, Yuxuan Zhou, Xu Zhu, Jingdong Sun, Jun-Yan He, Qi Dai, Alexander G. Hauptmann, and Zhi-Qi Cheng. Towards unified world models for visual navigation via memory-augmented planning and foresight. *arXiv preprint arXiv:2510.08713*, 2025. URL https://arxiv.org/abs/2510.08713.

Diqi He, Xuehao Gao, Hao Li, Junwei Han, and Dingwen Zhang. STRIDER: Navigation via instruction-aligned structural decision space optimization. *arXiv preprint arXiv:2511.00033*, 2025.

Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. VISTAv2: World imagination for indoor vision-and-language navigation. *arXiv preprint arXiv:2512.00041*, 2025a. URL https://arxiv.org/abs/2512.00041.

Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. VISTA: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868*, 2025b. URL https://arxiv.org/abs/2505.07868.

Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2107–2116. PMLR, 2018. URL https://proceedings.mlr.press/v80/icarte18a.html.

Intern Robotics. InternVLA-N1: An open dual-system vision-language navigation foundation model with learned latent plans. Technical report, Shanghai AI Laboratory, 2025. URL https://internrobotics.github.io/internvla-n1.github.io/.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14738–14748, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Koh_Pathdreamer_A_World_Model_for_Indoor_Navigation_ICCV_2021_paper.html.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4392–4412, 2020. doi: 10.18653/v1/2020.emnlp-main.356.

Andrew Li, Zizhao Chen, Toryn Q. Klassen, Pashootan Vaezipoor, Rodrigo Toro Icarte, and Sheila McIlraith. Reward machines for deep RL in noisy and uncertain environments. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110341–110368. Curran Associates, Inc., 2024. doi: 10.52202/079017-3502. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c71769e2715835d37c3e25cc1173bd62-Paper-Conference.pdf.

Boqi Li, Siyuan Li, Weiyi Wang, Anran Li, Zhong Cao, and Henry X. Liu. Boosting zero-shot VLN via abstract obstacle map-based waypoint prediction with TopoGraph-and-VisitInfo-aware prompting. *arXiv preprint arXiv:2509.20499*, 2025a.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1834. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4ec03ed08a3fcb59e1c815b5598beff1-Paper-Datasets_and_Benchmarks_Track.pdf.

Kailing Li, Tianwen Qian, Lijin Yang, Yuqian Fu, Jingyu Gong, Xiaoling Wang, and Liang He. Bridging the 2D-3D gap: A hierarchical semantic-geometric map for vision language navigation. *arXiv preprint arXiv:2606.00095*, 2026a.

Yijian Li, Changze Li, Hantian Shi, Jiaying Luo, Jiyuan Cai, Ming Yang, and Tong Qin. AgenticNav: Zero-shot vision-and-language navigation as a tool-calling harness, 2026b. URL https://arxiv.org/abs/2606.10577.

Zerui Li, Hongpei Zheng, Fangguo Zhao, Aidan Chan, Jian Zhou, Sihao Lin, Shijie Li, and Qi Wu. One agent to guide them all: Empowering MLLMs for vision-and-language navigation via explicit world representation. *arXiv preprint arXiv:2602.15400*, 2026c.

Sihao Lin, Zerui Li, Xunyi Zhao, Gengze Zhou, Liuyi Wang, Rong Wei, Rui Tang, Juncheng Li, Hanqing Wang, Jiangmiao Pang, Anton van den Hengel, Jiajun Liu, and Qi Wu. VLNVerse: A benchmark for vision-language navigation with versatile, embodied, realistic simulation and evaluation. *arXiv preprint arXiv:2512.19021*, 2025.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. InstructNav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024.

Kun Luo, Xiangyu Dong, Xiaoguang Ma, Haoran Zhao, and Yaoming Zhou. LightZeroNav: Zero-shot vision language navigation in continuous environments based on lightweight VLMs. *arXiv preprint arXiv:2603.16947*, 2026.

Kailin Lyu, Kangyi Wu, Pengna Li, Xiuyu Hu, Qingyi Si, Cui Miao, Ning Yang, Zihang Wang, Long Xiao, Lianyu Hu, Jingyuan Sun, and Ce Hao. HiMemVLN: Enhancing reliability of open-source zero-shot vision-and-language navigation with hierarchical memory system. *arXiv preprint arXiv:2603.14807*, 2026.

Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-Nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source LLMs. *arXiv preprint arXiv:2409.18794*, 2024.

Qwen Team. Qwen3.8-27B model card. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026.

Kai Sheng, Liuyi Wang, Haojie Dai, Jinlong Li, Yongrui Qin, Zongtao He, Chengju Liu, and Qijun Chen. P2DNav: Panorama-to-downview reasoning for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2605.19634*, 2026.

Xiangyu Shi, Zerui Li, Wenqi Lyu, Jiatong Xia, Feras Dayoub, Yanyuan Qiao, and Qi Wu. SmartWay: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2503.10069*, 2025.

Guoqing Wang, Pin Tang, Xiangxuan Ren, Guodongfang Zhao, Bailan Feng, and Chao Ma. Learning vision-language-action world models for autonomous driving. *arXiv preprint arXiv:2604.09059*, 2026. URL `https://arxiv.org/abs/2604.09059`.

Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. DREAMWALKER: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10873–10883, 2023. URL `https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DREAMWALKER_Mental_Planning_for_Continuous_Vision-Language_Navigation_ICCV_2023_paper.html`.

Teng Wang, Xinxin Zhao, Wenzhe Cai, and Changyin Sun. ImagineNav++: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2512.17435*, 2025a. URL `https://arxiv.org/abs/2512.17435`.

Yunheng Wang, Yuetong Fang, Taowen Wang, Yixiao Feng, Yawen Tan, Shuning Zhang, Peiran Liu, Yiding Ji, and Renjing Xu. DreamNav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2509.11197*, 2025b.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753–13762, 2024. URL `https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Lookahead_Exploration_with_Neural_Radiance_Representation_for_Continuous_Vision-Language_Navigation_CVPR_2024_paper.html`.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. *arXiv preprint arXiv:2512.08186*, 2025a.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming vision-and-language navigation via SlowFast context modeling. *arXiv preprint arXiv:2507.05240*, 2025b.

Ning Yang, Yan Huang, Kaiwen Peng, Ziheng He, Kai Wang, Cui Miao, Kailin Lyu, Guo Li, Xiaofeng Wang, Zheng Zhu, Jing Liu, and Nianfeng Liu. WAM-Nav: Asymmetric latent world-action modeling for unified visual navigation. *arXiv preprint arXiv:2606.04907*, 2026. URL https://arxiv.org/abs/2606.04907.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-time scaling with world models for spatial reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.12508.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. URL https://arxiv.org/abs/2408.06072.

Hang Yin, Haoyu Wei, Xiuwei Xu, Wenxuan Guo, Jie Zhou, and Jiwen Lu. GC-VLN: Instruction as graph constraints for training-free vision-and-language navigation. *arXiv preprint arXiv:2509.10454*, 2025.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, Yuxin Fan, Wenjun Li, Zhibo Chen, Fei Gao, Qi Wu, Zhizheng Zhang, and He Wang. Embodied navigation foundation model. *arXiv preprint arXiv:2509.12129*, 2025.

Jiazhao Zhang, Gengze Zhou, Hale Yin, Yiyang Huang, Zixing Lei, Qihang Peng, Haoqi Yuan, Jie Zhang, Xudong Guo, Xiaoyue Chen, et al. Qwen-robotnav technical report: A scalable navigation model designed for an agentic navigation system. *arXiv preprint arXiv:2606.18112*, 2026a.

Jiwen Zhang, Zejun Li, Siyuan Wang, Xiangyu Shi, Zhongyu Wei, and Qi Wu. SpatialNav: Leveraging spatial scene graphs for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2601.06806*, 2026b.

Jiwen Zhang, Xiangyu Shi, Siyuan Wang, Zerui Li, Zhongyu Wei, and Qi Wu. SpatialAnt: Autonomous zero-shot robot navigation via active scene reconstruction and visual anticipation. *arXiv preprint arXiv:2603.26837*, 2026c.

Baining Zhao, Jiacheng Xu, Weicheng Feng, Xin Zhang, Zhaolu Wang, Haoyang Wang, Shilong Ji, Ziyou Wang, Jianjie Fang, Zhiheng Zheng, Weichen Zhang, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. WorldVLN: Autoregressive world action model for aerial vision-language navigation. *arXiv preprint arXiv:2605.15964*, 2026. URL https://arxiv.org/abs/2605.15964.

Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. ImagineNav: Prompting vision-language models as embodied navigator through scene imagination. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=vQFw9ryKyK.

Wanrong Zheng, Yunhao Ge, and Laurent Itti. Three-Step Nav: A hierarchical global–local planner for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2604.26946*, 2026.

Changqing Zhou, Yueru Luo, Zeyu Jiang, and Changhao Chen. UniNav: A unified world-action diffusion model for visual navigation. *arXiv preprint arXiv:2608.03244*, 2026. URL https://arxiv.org/abs/2608.03244.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *AAAI*, 2024.

Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3D persistent embodied world models. *arXiv preprint arXiv:2505.05495*, 2025. URL https://arxiv.org/abs/2505.05495.

Deyi Zhu, Haoyu Fan, Yinan Zhu, Weichen Zhang, Shilin Ma, Xinlei Chen, and Yansong Tang. Uncertainty-aware world model for aerial image-goal navigation. *arXiv preprint arXiv:2608.05597*, 2026.

A Appendix

We provide detailed ablation results and paired comparisons in Appendix A.1, clarify the setting of methods with pre-exploration in Appendix A.2, and report per-configuration confidence intervals in Appendix A.3. Appendix A.4 examines the use and cost of imagination. We present world-model implementation details in Appendix A.5 and further explain APG's progress surrogate and reasoning protocol in Appendix A.6, followed by UGG's calibration and fusion procedures in Appendix A.7. Finally, Appendix A.8 analyses additional navigation examples, including successful decisions, failure cases, and the effects of imagination overhead.

A.1 Detailed ablation results

Table 3 reports the absolute navigation scores underlying Figure 2. For each backbone, all configurations are evaluated on the same 100 R2R-CE `val_unseen` episodes, with one run per configuration. We first add imagined rollouts to the MapGPT-style base agent, then introduce forced APG scoring, replace it with adaptive APG†, and finally add UGG. Forced and adaptive APG are alternative scoring protocols, not two simultaneously active components.

Table 3: **Component ablation on R2R-CE.** Components added to a MapGPT-style base: imagined rollouts (Imag.), Anticipated Progress Grounding in forced (APG) or adaptive (APG†) form, and Uncertainty-aware Geometric Grounding (UGG). ΔSR is paired per episode against the base agent of the same block, with a bootstrap 95% CI.

Agent configuration	Components			Navigation					Paired vs. base
	Imag.	APG	UGG	NE↓	OSR↑	SR↑	SPL↑	nDTW↑	ΔSR [95% CI]
Qwen3.5-4B									
Base agent				7.76	31	15	13.4	22.3	–
+ imagination	✓			7.41	33	18	16.2	24.5	+3 [−2, +8]
+ APG	✓	forced		7.33	42	19	17.5	29.6	+4 [−1, +9]
+ APG†	✓	adapt.		6.35	47	23	20.4	36.3	+8 [+2, +14]
+ UGG (AvatarVLN)	✓	adapt.	✓	**6.21**	**52**	**25**	**22.8**	**37.5**	+10 [+4, +16]
Qwen3.5-9B									
Base agent				7.23	36	24	20.9	33.5	–
+ imagination	✓			7.01	42	27	23.4	36.9	+3 [−2, +8]
+ APG	✓	forced		6.83	44	28	25.3	34.5	+4 [−1, +9]
+ APG†	✓	adapt.		6.61	46	30	26.7	42.7	+6 [0, +12]
+ UGG (AvatarVLN)	✓	adapt.	✓	**5.59**	**48**	**34**	**29.5**	**43.3**	+10 [+4, +16]
Qwen3.8-27B									
Base agent				6.55	52	39	28.5	39.8	–
+ imagination	✓			6.10	59	43	30.4	43.1	+4 [−1, +9]
+ APG	✓	forced		5.78	60	44	31.7	46.2	+5 [0, +10]
+ APG†	✓	adapt.		5.54	64	47	33.4	48.4	+8 [+2, +14]
+ UGG (AvatarVLN)	✓	adapt.	✓	**5.27**	**66**	**51**	**34.3**	**53.4**	+12 [+6, +18]

Relative improvements in the radar plot. For metrics where larger values are better, the plotted gain is $100(m_{\text{config}} - m_{\text{base}})/m_{\text{base}}$. For NE, it is $100(\text{NE}_{\text{base}} - \text{NE}_{\text{config}})/\text{NE}_{\text{base}}$, so a positive

value denotes reduced navigation error. The base polygon denotes zero gain. Each panel uses one shared radial scale across its five metrics, but the scale is normalised independently for each backbone; polygon sizes should therefore not be compared directly across panels. These relative percentages differ from the absolute SR differences below, which are reported in percentage points.

Paired comparisons against the base agent. For each configuration, Table 3 reports the SR difference from the corresponding base agent, paired by episode, together with a bootstrap 95% confidence interval. The marginal SR interval of each configuration is listed in Table 5 (Appendix A.3). Naive imagination improves SR by 3–4 points, but all three confidence intervals include zero. Forced APG produces gains of 4–5 points over the base, whose intervals still include zero, while adaptive APG† yields gains of 6–8 points with lower bounds of 2, 0 and 2 points. The complete AvatarVLN improves SR by 10, 10, and 12 points for the 4B, 9B, and 27B agents, respectively, with confidence-interval lower bounds of 4, 4 and 6 points. These paired results support the effectiveness of the complete framework relative to the base across model scales. They do not isolate the effect of each adjacent component increment, since each configuration is compared with the base.

Component trends. Every augmented configuration improves all five metrics relative to its own base, and the complete model performs best on every metric at all three scales. Intermediate improvements need not be monotonic: for the 9B agent, forced APG reduces nDTW from 36.9 to 34.5 relative to naive imagination, while remaining above the base score of 33.5. The 4B agent has the largest relative SR gain (15 to 25, or 66.7%), whereas the 27B agent has the largest absolute gain (39 to 51, or 12 percentage points).

A.2 Zero-shot methods with a pre-exploration stage

The methods below are zero-shot in the same sense as ours – no task-specific training – but they explore the scene before the episode starts, so they enter the episode knowing the layout. That is a different problem setting, and we keep them out of Table 1 rather than compare against them directly.

Table 4: Zero-shot methods with a pre-exploration stage, as reported by their own papers. Columns follow Table 1.

	R2R-CE					RxR-CE				
Method	NE↓	OSR↑	SR↑	SPL↑	nDTW↑	NE↓	OSR↑	SR↑	SPL↑	nDTW↑
VLN-Zero (Bhatt et al., 2025)	5.97	52	42	26.3	–	9.13	–	31	19.0	–
SpatialNav (Zhang et al., 2026b)	5.15	66	64	51.1	65.4	7.64	–	32	24.6	55.0
SpatialAnt (Zhang et al., 2026c)	4.42	76	66	54.4	69.5	5.28	–	51	35.6	65.4

A.3 Per-configuration confidence intervals

Table 3 reports paired differences, which is the quantity a reader should use when comparing two rows evaluated on the same episodes. For completeness, Table 5 gives the marginal interval of every configuration: a Wilson score interval on the success count of $n = 100$ episodes. Marginal intervals of two rows may overlap while the paired interval of their difference excludes zero, because the paired comparison only uses the episodes on which the two configurations disagree.

A.4 Cost of imagination

Imagination is optional at every decision, so its cost is set by the agent rather than by the framework. Figure 5 shows how often the open-weight agents request a rollout and how often the adaptive protocol chooses the structured comparison on R2R-CE, where each request adds a call to the world model and each structured comparison adds a longer reasoning round. Smaller agents rely on imagination much more heavily, with 24.9, 12.3 and 7.9 rollout requests per episode for the 4B, 9B and 27B agents, and they also run the structured comparison more often, 11.3, 7.7 and 2.0 times per episode. On our hardware a 27B decision round takes more than 30 s and each imagination request adds 8 to 15 s, which is why the largest agent accounts for most wall-clock time-outs even though it imagines least (Table 7).

Table 5: Marginal 95% Wilson intervals for SR on R2R-CE `val_unseen` ($n = 100$). At this sample size the half-width is about ± 9 points for every row.

Agent configuration	Imag.	APG	UGG	SR↑	95% CI
Qwen3.5-4B					
Base agent (MapGPT-style)				15	[9.3, 23.3]
+ imagination	✓			18	[11.7, 26.7]
+ APG	✓	forced		19	[12.5, 27.8]
+ APG†	✓	adapt.		23	[15.8, 32.2]
+ UGG (**AvatarVLN**)	✓	adapt.	✓	25	[17.5, 34.3]
Qwen3.5-9B					
Base agent (MapGPT-style)				24	[16.7, 33.2]
+ imagination	✓			27	[19.3, 36.4]
+ APG	✓	forced		28	[20.1, 37.5]
+ APG†	✓	adapt.		30	[21.9, 39.6]
+ UGG (**AvatarVLN**)	✓	adapt.	✓	34	[25.5, 43.7]
Qwen3.8-27B					
Base agent (MapGPT-style)				39	[30.0, 48.8]
+ imagination	✓			43	[33.7, 52.8]
+ APG	✓	forced		44	[34.7, 53.8]
+ APG†	✓	adapt.		47	[37.5, 56.7]
+ UGG (**AvatarVLN**)	✓	adapt.	✓	51	[41.3, 60.6]

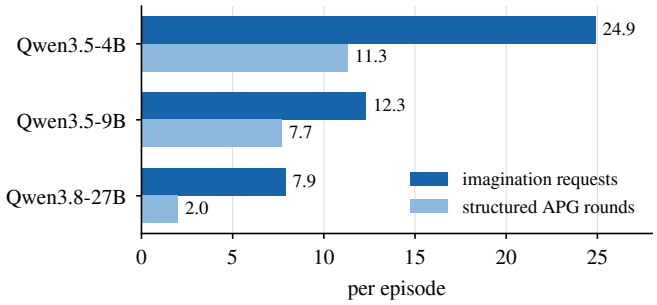


Figure 5: Imagination requests and structured APG rounds per episode for the open-weight agents on R2R-CE. Imagination requests are averaged over all configurations that expose the tool, and structured rounds over the two configurations with adaptive APG.

A.5 World-model implementation

RGB and depth are encoded into separate video latents and concatenated; the diffusion transformer predicts their joint evolution, conditioned on the initial observation and per-frame Plücker camera embeddings, and is fine-tuned with RGB and depth denoising losses. The deployed predictor generates 25 RGB-D frames, including the conditioning frame, at 384×384 resolution. RGB and depth each occupy 16 latent channels. Depth is represented in meters divided by 10, clipped to $[0, 1]$; UGG converts predictions back to meters before calibration. Training samples temporal intervals up to three action steps, and uses the fixed caption "A random room tour." Our world model uses full transformer fine-tuning for 30,000 steps with learning rate 7×10^{-5}, effective batch size 32, and a depth-loss weight of 0.5 relative to RGB. The map-conditioning branch of the world-model architecture is inactive in this checkpoint; UGG's map is an input to the navigation agent.

For waypoint imagination, the current panorama supplies a view facing near the candidate direction. Residual rotations and forward motion are converted to a nominal camera trajectory, using 15° turns and 0.25 m forward increments. The effective sequence ends at the candidate arrival or the generation limit; padded frames are excluded from the agent's comparison. Camera poses specify the proposed motion and do not certify collision-free execution. Inference uses 50 denoising steps with guidance scale 3.

A.6 PROGRESS SURROGATE AND REASONING PROTOCOL

Decomposing anticipated progress. Adding and subtracting $\sum_r q_t(r)C(h_b, r)$ in Eq. (5) gives

$$\Delta_b^{\text{within}} = \sum_r q_t(r)\big[C(\mathcal{H}_t, r) - C(h_b, r)\big], \tag{9}$$

$$\Delta_b^{\text{advance}} = \sum_r \big[q_t(r) - q_{H_b}^b(r)\big]C(h_b, r), \tag{10}$$

$$\Delta_b = \Delta_b^{\text{within}} + \Delta_b^{\text{advance}}. \tag{11}$$

The first term holds the progress belief fixed and compares proximity to the next event. The second evaluates belief advancement at the same endpoint. Under the unit-event convention in Eq. (3), $C(h, r)$ is nonincreasing in r: for $r < K-1$, $C(h, r) - C(h, r+1) = 1 + d(h, e_{r+1}) - d(h, e_{r+2}) \geq 0$, and $C(h, K-1) = d(h, e_K) \geq 0$. Consequently a belief update that only transfers mass forward has nonnegative advancement gain. This justifies discounting that term in Eq. (6); evidence that a candidate moves away from the next event is expressed through the within-event term. The clipping budget B_b is expressed in surrogate units, not directly in meters. Both clipping and reliability weighting are reasoning safeguards, not physical error bounds.

Monotonicity and missing evidence. Monotonicity concerns accomplishment of ordered events, not the robot's distance from its destination: physical backtracking can increase the remaining within-event work without undoing an event. All candidates start from a belief grounded in the same real history. Within an imagined branch, the protocol permits forward belief changes only on affirmative evidence. In particular, repeated NO-EVIDENCE observations do not induce a forward drift. Evidence-gap transitions infer that intermediate constraints were satisfied without being rendered; they cannot override explicit evidence that an earlier required event remains unmet. Re-observation establishes the next real-state belief independently of speculative branch completion.

Agent implementation. The full protocol asks the agent to report the ordered events, current progress belief, and, for each candidate, a summary of temporal evidence, reliability, endpoint belief, within-event gain, advancement gain, and guarded preference. These are judgments by the frozen multimodal agent; no fixed emission table or transition matrix is fitted or numerically executed. The adaptive protocol first selects DIRECT or SCORED reasoning. In the latter mode, it uses a concise event-based comparison and the same qualitative safeguards, without requiring a numerical posterior or every term of the full report. Thus the adaptive ablation changes both invocation of structured reasoning and the verbosity of the prescribed procedure. It does not extend the world model's rollout horizon.

A.7 GEOMETRIC CALIBRATION AND FUSION DETAILS

Robust calibration. For each rollout we pool matched predicted and reference depths across frames, with a cap on samples from each frame so that the dense conditioning view does not dominate. The regression design for a sample $(\hat{D}_k(u), D_k^{\text{ref}}(u))$ is $(\hat{D}_k(u), 1, k\hat{D}_k(u), k)$. Gross mismatches above $1.5\,\text{m}$ are excluded and their rejection fraction is retained for confidence estimation. We perform three least-squares iterations with residual trimming, shrink the fitted slopes by 0.5, and clip scale to $[0.3, 3]$ and shift to $[-2, 2]\,\text{m}$. The joint fit requires at least 800 matched samples across multiple frames. Its fallback fits only the conditioning view, with a median-ratio scale or identity correction if the affine fit is degenerate.

Anchored confidence. We omit the candidate superscript b below for clarity. A frame supplies a confidence anchor when at least 100 sampled correspondences survive the robust fit. Let m_k be their median absolute residual after the applied scale and shift, including slope shrinkage and parameter clipping. Let s_k be the fraction of valid correspondences surviving the initial $1.5\,\text{m}$ gross-mismatch gate, before subsampling and residual trimming. This factor reduces confidence when most correspondences are rejected, even if the remaining pairs agree. With conditioning-frame photometric factor f_{rgb}, the confidence is

$$c_k = \text{clip}_{[0.1,1]}\Big[s_k f_{\text{rgb}}\ \text{clip}_{[0.15,1]}(1.2 - m_k/(1\,\text{m}))\Big]. \tag{12}$$

The photometric factor measures how well the world model reconstructs its conditioning observation. We resize the predicted and measured frame-0 RGB images to 64×64 and compute their mean absolute difference ϵ_0 over pixels and channels, with intensities in $[0, 255]$: $f_{\mathrm{rgb}} = \mathrm{clip}_{[0.3,1]}(1.35 - \epsilon_0/80)$. This factor is shared across the rollout. Anchored confidence is recomputed from real correspondences and can recover when the trajectory re-enters observed space. Unanchored frames use Eq. (8), with $c_{\min} = 0.1$.

Fitting motion-dependent decay. The two coefficients in the main text are fitted offline using generated navigation rollouts paired with recorded reference depth. The fitting data exclude the scenes sampled for the offline geometry evaluation. Calibration uses only the initial measured depth and its reprojections along the recorded poses; future reference depths are reserved for quality labels, not supplied to the calibration procedure.

Let D_k^{gt} denote the recorded reference depth and $\mathcal{V}_k$ the pixels with finite calibrated depth and reference depth between 0.3 and $9.5\,\mathrm{m}$. The frame-quality label is

$$z_k = \frac{1}{|\mathcal{V}_k|} \sum_{u \in \mathcal{V}_k} \mathbf{1}\Big[|\tilde{D}_k(u) - D_k^{\mathrm{gt}}(u)| < \max\big(0.2\,\mathrm{m}, 0.075 D_k^{\mathrm{gt}}(u)\big)\Big]. \tag{13}$$

Here the calibrated depth includes the filtering applied by the depth pipeline; frames with fewer than 500 valid pixels have no quality label. For selecting training pairs, the offline procedure treats a frame with at least 500 finite reprojected reference pixels as anchored. This is a coverage-based proxy, distinct from the post-fit correspondence criterion used at deployment. Each unanchored frame k is paired with its most recent anchor j, provided both quality labels are valid.

For each pair, we compute cumulative translation $L_{j:k}$ in meters and rotation $\Theta_{j:k}$ in radians from consecutive camera poses. The target is the negative log of relative quality retained since the anchor:

$$y_{j,k} = -\log\left[\mathrm{clip}_{[\epsilon,1]}\left(\frac{z_k}{\max(z_j, 10^{-6})}\right)\right], \qquad \epsilon = 0.05. \tag{14}$$

Clipping caps extreme degradation and maps quality improvements to zero decay. We fit the coefficients over the set of valid training pairs $\mathcal{P}$ by nonnegative Huber regression:

$$(\lambda_L, \lambda_\theta) = \operatorname*{arg\,min}_{\lambda_L, \lambda_\theta \geq 0} \sum_{(j,k) \in \mathcal{P}} \ell_{0.5}(\lambda_L L_{j:k} + \lambda_\theta \Theta_{j:k} - y_{j,k}), \tag{15}$$

where $\ell_{0.5}$ is the Huber loss with transition at 0.5. The fitted values are $\lambda_L = 0.424389\,\mathrm{m}^{-1}$ and $\lambda_\theta = 0.619169\,\mathrm{rad}^{-1}$. They are held fixed at deployment; future reference depth is not required.

For implementation completeness, a failed drift fit falls back to frame-0 calibration and propagates its confidence as $\mathrm{clip}_{[0.1,1]}(c_0\gamma^k)$, with $\gamma = 0.92$. The unanchored-frame decay also reverts to γ^{k-j} when fitted motion parameters or poses are unavailable. These fallback rules are separate from the motion-aware weighting used when the fit and required inputs are available.

Local and cross-frame filtering. The filtering stage applies a 3×3 median filter to calibrated depth and removes pixels with a depth-gradient magnitude above $0.6\,\mathrm{m}$ when the corresponding mean-channel RGB gradient is below 6 on the $[0, 255]$ intensity scale. Gradients use the sum of absolute horizontal and vertical forward differences. This is a local discontinuity check, not a test of whether generated appearance is factually correct.

For cross-frame consistency, each neighboring depth frame $j \in \{k-1, k+1\}$ is reprojected into frame k using the candidate camera poses. Let $w_{j\to k}(u)$ be its projected depth and let $\tau_k(u) = \max(0.20\,\mathrm{m}, 0.075\tilde{D}_k(u))$. A finite projection is eligible for comparison when $w_{j\to k}(u) < \tilde{D}_k(u) + \tau_k(u)$; agreement additionally requires $|w_{j\to k}(u) - \tilde{D}_k(u)| < \tau_k(u)$. A pixel is removed only if at least one neighbor is eligible and none agrees. A projection substantially behind the target surface is treated as an occlusion case rather than a contradiction. Pixels with no eligible neighbor retain their depth, so lack of overlapping evidence alone does not remove newly revealed geometry. These checks reject inconsistency; they do not certify the correctness of mutually consistent predictions.

Scan-based BEV fusion. The `scan` configuration converts the surviving depth in each frame into a virtual 2D laser scan. We back-project depth samples at a pixel stride of 3, retaining depths between 0.3 and 6 m, and transform them using the candidate poses. The local BEV has 0.10 m cells and extent ± 4.8 m. Points 0.25–1.70 m above the floor are obstacle candidates; points within 0.20 m of the floor provide floor evidence.

Obstacle points are grouped into 0.75° azimuth bins around each frame's camera position. In a bin containing n points, we start at the 10th-percentile range and collect points up to 0.30 m farther away. The window must contain at least $\min(3, n)$ points; its median range defines the robust first hit. A beam contributes free-space evidence up to one grid cell before this hit and obstacle evidence at the hit, but writes nothing behind it. For bins without an accepted obstacle hit, floor observations support free space only out to their 90th-percentile range. No evidence is extrapolated beyond that extent.

Ray samples are spaced by 0.05 m. Obstacle hits contribute weight c_k, while each free-space sample contributes $c_k(0.05/0.10)$ to account for sampling density. Evidence accumulates across beams and frames. A cell is marked free at accumulated free-space weight 1.0 and occupied at obstacle weight 1.2, with occupied taking precedence if both thresholds are reached. Cells below both thresholds remain unknown. These are weighted evidence thresholds, not counts of distinct supporting frames. Space behind a hit receives no evidence from that beam, but may be resolved by other views or real observations.

The persistent real map is accumulated separately from measured depth; the `scan` fusion above applies to the temporary imagined layer. At rendering time, observed cells override imagined cells, and the two sources are visually distinguished. Predicted geometry is never written back into persistent real memory.

A.8 Case studies

This section expands Section 4.3. By following the base agent and +APG† through the same episodes, we identify four recurring effects of imagination on the agent's behaviour, and each builds on the one before it.

1. **Imagination supplies what the panorama cannot show** (Cases 1 to 5). A rollout reveals where a doorway leads, whether a path crosses a room, what lies behind a door and on which side an object passes, provided that the agent chooses to imagine the right candidates.
2. **A rollout is only as useful as the agent's judgement of it** (Cases 6 to 8 and Case B). Stronger agents notice rollouts that stall or run into a wall, whereas weaker ones may report evidence that a rollout does not contain, and even a correct choice can rest on a misread rollout.
3. **Event order keeps imagination aligned with the instruction** (Cases 9 to 12). APG rejects shortcuts that skip unfinished events and prevents false arrivals, although an agent may still stop on a predicted arrival, which the protocol forbids.
4. **Imagination costs time** (Case 13). Every request adds a decision round, and extra deliberation at the goal can turn a success into a time-out. Appendix A.4 quantifies this cost.

Because these effects push success in opposite directions, the average effect of imagination alone is small (Table 3).

Setting. We study +APG†, which combines imagination with adaptive APG but not UGG, because its written reasoning shows how rollouts are used, whereas UGG adds maps but no text. Configurations are named as in Table 3, namely Base, +Imag., +APG and +APG†, and +APG$_{\text{lean}}$ denotes a shortened protocol that we ran as an additional control but do not report in Table 3. Opus 5 and GPT-5.5 were run only as +APG†. Each figure shows the panorama at the round where the configurations diverge, with the waypoint numbers drawn by the proposer, together with the imagined rollouts, which start from the current view of each waypoint and end at the predicted arrival. Agent text is quoted verbatim with omissions marked [...], and the agent calls rollouts filmstrips because the prompt does. The result tables report the navigation error (NE), nDTW, path length, the numbers of waypoint moves and imagination requests, and how the episode ended, either by the agent's stop,

by the 30-move budget or by the 40-minute wall-clock limit, with a dash marking values that were not recorded.

What the prompt asks for. The +APG† agent sees the same panorama, candidates and history as the base agent and additionally receives the imagination tool, a note attached to every returned rollout and the adaptive APG protocol (Appendix A.6), whose relevant parts are reproduced verbatim below.

> *Tool description.* imagine(waypoints=[...]): OPTIONAL. A learned world model predicts, frame by frame, what you would see walking to each waypoint you name. Use it when the choice is genuinely uncertain — a junction, the instruction and the view not obviously lining up, two plausible candidates. [...] Skip it when the right move is already obvious.
>
> *Rollout note.* One filmstrip per waypoint follows [...] frame 0 is the labelled panorama view you already saw, the last frame is the predicted arrival. These are PREDICTIONS — trust layout, geometry and whether the path stays open more than fine detail or colour.
>
> *Adaptive protocol (excerpt).* MODE: DIRECT — one candidate is clearly right [...] MODE: SCORED — the candidates genuinely compete, or the instruction and what you see do not line up. 1. EVENTS [...] 2. WHERE YOU STAND [...] A landmark that a prediction fails to show means NO EVIDENCE, never failure. 3. READ EACH FILMSTRIP against the next unsatisfied event [...] melted or duplicated structure counts for less. 4. COMPARE END STATES, never frame numbers [...] 5. GUARDS. A short walk cannot legitimately gain much ground. A candidate showing the destination while skipping an earlier required event does not win. And never stop() because a PREDICTION looks like the goal — stopping is decided from what you actually see.

Table 6 separates the behaviours that the prompt requests from those it leaves to the agent. For each positive case we ask whether the decisive step followed a rule or went beyond the rules, and for each negative case whether a rule was ignored or was itself misleading.

Table 6: Which behaviours of the +APG† agent are prompted.

Behaviour	Source in the prompt	Status
Imagine at junctions or when candidates are hard to separate	tool description	guided
Declare DIRECT or SCORED on the first line	protocol	required
Decompose into ordered events and compare candidates by progress	protocol steps 1 to 4	guided
Reject a candidate that shows the goal but skips an earlier event	protocol step 5	guided
Discount blurred or melted rollouts	step 3 and rollout note	guided
Choose which candidates to imagine	"when the choice is genuinely uncertain"	agent's judgement
Revise an earlier preference after a rollout contradicts it	not mentioned	agent's judgement
After imagining, conclude that the goal is not yet reached	only stopping on a prediction is forbidden	agent's judgement
Stop on a predicted goal or report content absent from a rollout	forbidden and contrary to the rollout note	failure mode

Board-level counts. Table 7 reports how often these behaviours occur across the open-weight +APG† runs. The counts rely on keyword rules fixed before counting and are therefore lower bounds, and the mode declaration is read from the first three non-empty lines of each round. Smaller agents imagine more often, with 18.4, 7.8 and 5.8 requests per episode for the 4B, 9B and 27B agents. Applying the same rules to the 27B configurations without the adaptive protocol provides a baseline, in which rule R1 fires 0, 0 and 6 times for Base, +Imag. and +APG_{lean}, rule R2 0, 0 and 7 times, and rule R3 0, 11 and 18 times. Rejections for skipping events therefore stem from the wording of the protocol rather than from the agent.

Overview. Table 8 groups the thirteen cases by the remark they support or qualify.

Table 7: Behaviour of the open-weight +APG† agents (Qwen3.5-4B, Qwen3.5-9B and Qwen3.8-27B) on R2R-CE. Rule R1 counts rejections for skipping events (*skip*, *jump ahead*, *shortcut*), R2 statements that a rollout is blurred or unreliable (*blurry*, *artifact*, *melted*, *distorted*, *unreliable*, *hallucinat**, *incoherent*), R3 moves after imagining that contrast candidates (*instead*, *rather than*, *changed my mind*), and R4 stop rounds that mention a prediction (*predict**, *filmstrip*, *preview*, *world model*, *imagin**).

	4B	9B	27B
Decision rounds per episode	15.9	13.7	15.1
Imagination requests per episode	18.4	7.8	5.8
Rounds with a request	85%	49%	37%
Episodes with a request	99	88	88
Rounds with recorded text	3354	2026	1943
First line SCORED	23%	33%	9%
First line DIRECT	17%	30%	31%
No mode declared	60%	37%	60%
R1 skipped event (rounds)	79	82	34
R2 unreliable rollout (rounds)	5	4	5
R3 contrastive move (rounds)	45	12	19
R4 stop cites prediction (fail / succ.)	22 / 8	8 / 3	2 / 0
Reached goal without stopping	20	10	26
Wall-clock time-outs	1	0	10
Move budget exhausted	4	7	8

Table 8: The four remarks, the cases that support them and those that cut the other way.

Remark	For	Against	Board-level evidence
1. Imagination supplies what the panorama cannot show	1, 2, 3, 4, 5	–	In most of the episodes that only +APG† solves, the agent requested a rollout at the round where the configurations diverge.
2. A rollout is only as useful as the agent's judgement of it	6, 7	8	Agents rarely call a rollout blurred or unreliable, in only 5, 4 and 5 rounds for the 4B, 9B and 27B agents, and three further episodes are solved with a justification that does not match the rollout (Case B).
3. Event order keeps imagination aligned with the instruction	9, 10, 11	12	Rejecting a candidate for skipping an earlier event occurs in 34 rounds for the 27B +APG† agent and never for its base agent, while a stop that cites a prediction in an episode that fails occurs in 22, 8 and 2 episodes for the 4B, 9B and 27B agents.
4. Imagination costs time	–	13	In four episodes that the 27B base agent solves, +APG† is cut off by the wall-clock limit while already within 3 m of the goal.

A.8.1 Remark 1. Imagination supplies what the panorama cannot show (Cases 1 to 5)

In most of the episodes that only +APG† solves, the agent requested a rollout at the round where the configurations diverge.

Case 1. Two doorways that look alike lead to different places. In R2R-CE episode 568 with Qwen3.8-27B, the configurations diverge at round 1 (Figure 6), and the instruction reads "Exit the room and then go down the hallway until you get a frame on the wall to your right with a bear and balloons that says "welcome" and wait there."

The agent starts in a bedroom with two ways out, an arch beside the bed (waypoints 1 and 2) and an opening next to a dark arched door (waypoints 3 to 5), and the panorama does not reveal which of them leads to the hallway. The base agent picks waypoint 2, and the +imagination agent imagines only waypoints 1 and 2 before making the same choice. +APG† also imagines waypoint 3, whose

rollout shows the walls closing in on both sides as it enters a corridor, so it moves there, follows the corridor for 14 moves and stops 0.59 m from the goal with the "welcome" frame in view. The other two agents end up in the en-suite bathroom and never come within 10 m of the goal.

> +APG†, round 1. *"MODE: DIRECT — Waypoint 3 clearly shows the exit from the bedroom into a hallway — the predicted walk-through shows me passing the dark arched door and entering a corridor."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.8-27B	Base	fail	14.52	0.01	39.0	29	0	stop
Qwen3.8-27B	+Imag.	fail	12.74	0.01	37.7	29	–	move budget
Qwen3.8-27B	**+APG†**	**success**	**0.59**	**0.74**	**19.4**	**14**	**3**	**stop**

Relation to Remark 1. Both imagining agents use the same world model, so the difference lies in which candidates they choose to imagine, and the rollout that settles this episode is one the +imagination agent never requests.

Case 1 · Choosing the exit R2R-CE episode 568 · Qwen3.8-27B · diverging round 1

"Exit the room and then go down the hallway until you get a frame on the wall to your right with a bear and balloons that says "welcome" and wait there."

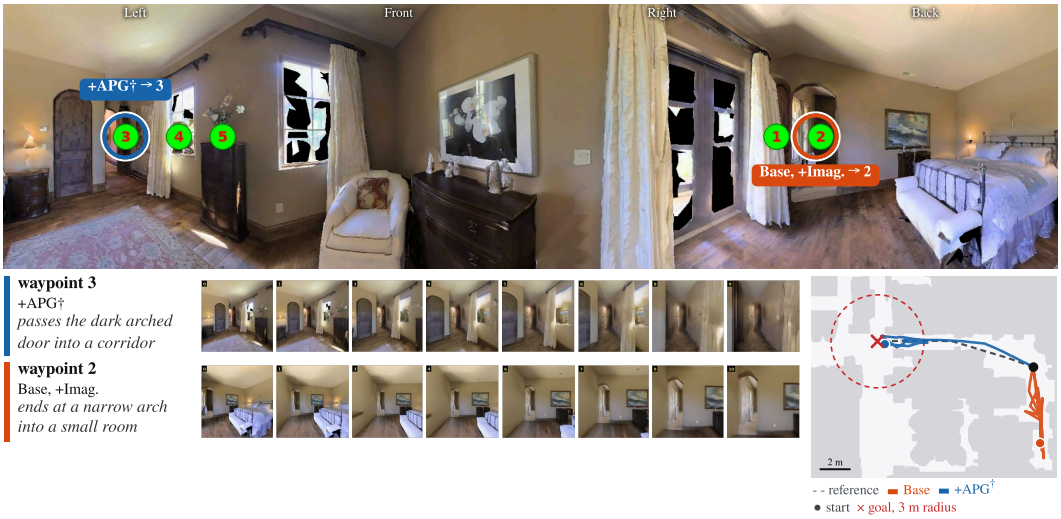


Figure 6: **Case 1.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 568, Qwen3.8-27B), with each rollout labelled by the configurations that chose it. The map shows the trajectories of the base agent and +APG† with the 3 m success radius.

Case 2. The first move decides whether a long route crosses the living room. In R2R-CE episode 70 with Qwen3.8-27B, the configurations diverge at round 1 (Figure 7), and the instruction reads "Go through living room, through the door on the to the right, through the den, through the dining to, to the outdoor foyer. Stop before going outside."

The instruction passes through five spaces, and at the start waypoints 2 to 5 all point roughly into the living room, differing only in angle. The base agent and the +imagination agent, which imagines waypoints 3 and 4, both pick waypoint 3. +APG† imagines waypoints 2 to 4 and sees that the rollout of waypoint 2 passes the sofa and coffee table while that of waypoint 3 hugs a wall towards the kitchen. It chooses waypoint 2, keeps imagining at later junctions with 13 requests over 16 moves, and stops 0.92 m from the goal. The base agent's first move in fact brings it geodesically closer to the goal (14.7 m against 15.5 m), yet it never gets closer than 9.3 m afterwards.

> +APG†, round 1. *"MODE: DIRECT — The instruction says to 'go through living room' first. [...] Waypoint 2's prediction clearly shows me moving through the living room past the sofa, which is the correct first step."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.8-27B	Base	fail	16.05	0.07	31.9	27	0	stop
Qwen3.8-27B	+Imag.	fail	9.30	0.13	33.4	28	–	time-out
Qwen3.8-27B	**+APG†**	**success**	**0.92**	**0.55**	**21.4**	**16**	**13**	**stop**

Relation to Remark 1. Going through the living room constrains the path rather than the endpoint, so a rule that rewarded getting closer to the goal at every step would have preferred the base agent's first move.

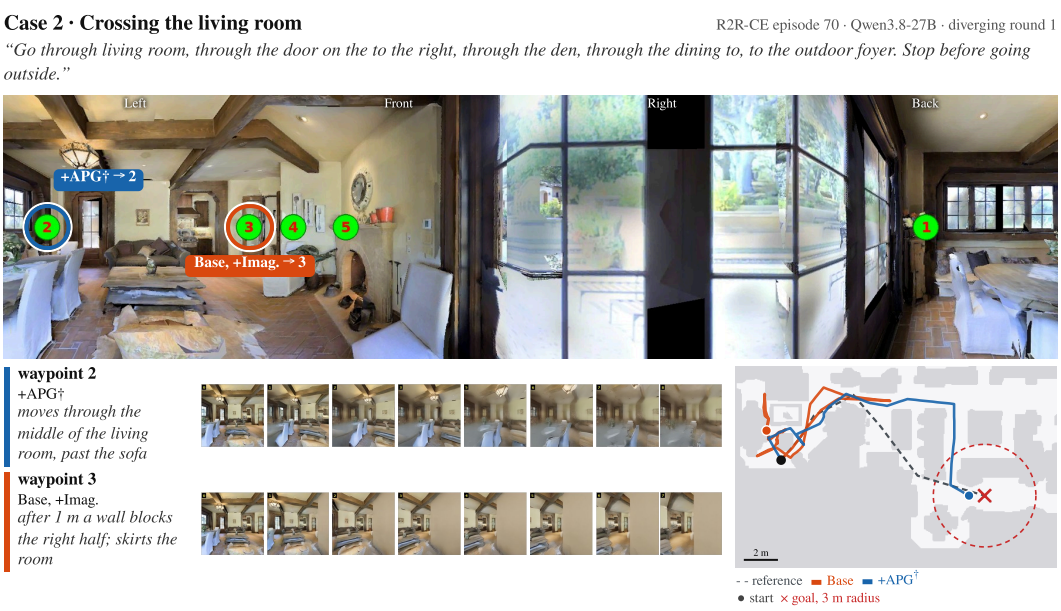


Figure 7: **Case 2.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 70, Qwen3.8-27B), with each rollout labelled by the configurations that chose it. The map shows the trajectories of the base agent and +APG† with the 3 m success radius.

Case 3. A small agent misreads a carpet as a tiled floor. In R2R-CE episode 586 with Qwen3.5-4B, the configurations diverge at round 1 (Figure 8), and the instruction reads "Go into the room with the tile floor and wait."

The agent stands in a hall with a parquet floor. The front door (waypoint 1) opens onto a bedroom with a pink carpet, and a light floor is faintly visible behind the back door (waypoint 2). The base agent reads the pink carpet as "a white tiled floor", walks through the front door and ends 18.7 m from the goal. Both imagining agents request the rollout of waypoint 2 before moving and cite the "tiled or stone floor" it shows as their reason to choose it. In the same episode the 9B and 27B base agents, Opus 5 and GPT-5.5 all succeed, while the 9B +APG† agent fails at 5.2 m.

> +APG†, round 1. *"MODE: DIRECT — The predicted path to waypoint 2 leads directly through a doorway into what appears to be a tiled hallway/room, which matches the goal."*

> Base, round 1. *"[...] the room ahead (Front view, waypoint 1) shows a doorway into a different room with what appears to be a white tiled floor. This is the target room."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.5-4B	Base	fail	18.72	0.00	38.8	20	0	stop
Qwen3.5-4B	+Imag.	success	1.60	0.94	3.8	2	3	stop
Qwen3.5-4B	**+APG†**	**success**	**0.82**	**0.62**	**11.6**	**5**	**4**	**stop**

Relation to Remark 1. The failure comes from a small model misreading detail in the panorama, and the rollout helps by bringing the relevant floor into close view.



Figure 8: **Case 3.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 586, Qwen3.5-4B), with each rollout labelled by the configurations that chose it.

Case 4. Passing on the left of a couch is visible only while moving. In R2R-CE episode 1148 with Qwen3.8-27B, the configurations diverge at round 3 (Figure 9), and the instruction reads "walk towards fireplace, walk on the left side of the brown couch, turn right after couch, turn left after four person table and stop in front of barstools."

At round 3 an orange sectional couch lies straight ahead with windows and a fireplace to its left, and waypoints 1 and 4, straight ahead and 18° to the left, both point towards it. The base agent reasons that waypoint 4 lies "between the fireplace and the couch" and drifts off course, ending 6.45 m from the goal with an nDTW of 0.03. Both imagining agents compare waypoints 1 and 4 and choose waypoint 1 because the couch stays on their right while they pass it.

> +APG†, round 3. *"MODE: DIRECT — [...] waypoint 1 (Front) takes me along the left side of it, as confirmed by the predicted walk-through showing the couch on my right as I pass it."*

> Base, round 3. *"Waypoint 4 is positioned between the fireplace and the couch, which is exactly where I need to go to walk on the left side of the couch."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.8-27B	Base	fail	6.45	0.03	32.1	24	0	stop
Qwen3.8-27B	+Imag.	success	1.51	0.50	17.7	11	3	stop
Qwen3.8-27B	**+APG†**	**success**	**1.38**	**0.53**	**19.5**	**12**	**5**	**stop**
Opus 5	+APG†	success	0.35	0.95	8.8	6	2	stop

Relation to Remark 1. Passing on the left of an object is defined by the direction of travel. In the panorama both candidates lie left of the couch only in projection, whereas the rollout shows on which side the couch actually passes. The gain here comes from imagination itself, since the agent without APG also chooses correctly and +APG† makes a direct decision.

Case 5. The rollout reveals which candidate exposes the next event. In R2R-CE episode 150 with Qwen3.5-4B, the configurations diverge at round 1 (Figure 10), and the instruction reads "Turn around and walk to the end of the bed. At the end of the bed turn right and walk to the hallway. Once at the entrance way, walk through the hallway to the next room."

The agent starts beside the bed, and its four walkable candidates (waypoints 2 to 5) all lie behind it within a few tens of degrees and between 0.25 and 1.25 m away. Lacking any other cue, the base agent picks the nearest candidate. +APG† imagines all four, switches to the structured comparison and chooses waypoint 5, whose rollout reaches the end of the bed with an arched doorway in view on the right, which is exactly the exit that the next event of the instruction, turning right towards the hallway, requires.

> +APG†, round 1. *"MODE: SCORED [...] the filmstrip shows a clear path to the bed's end, with the hallway opening up behind it, which is exactly what I need to perform the 'turn right' next."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.5-4B	Base	fail	5.50	0.52	23.2	23	0	stop
Qwen3.5-4B	+Imag.	success	2.14	0.09	42.6	27	61	stop
Qwen3.5-4B	**+APG†**	**success**	**2.79**	**0.47**	**24.6**	**18**	**22**	**stop**

Case 4 · Left of the couch R2R-CE episode 1148 · Qwen3.8-27B · diverging round 3

"walk towards fireplace, walk on the left side of the brown couch, turn right after couch, turn left after four person table and stop in front of barstools."

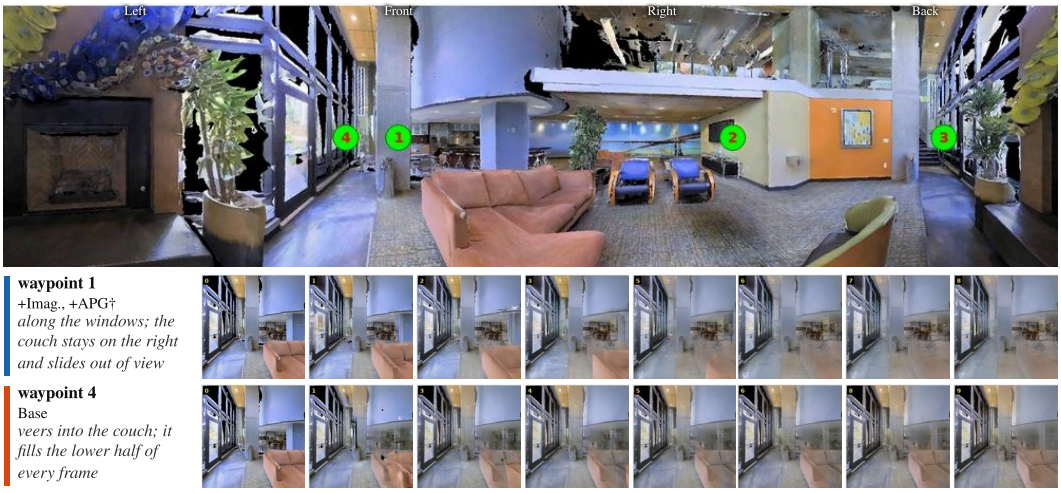


Figure 9: **Case 4.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 1148, Qwen3.8-27B), with each rollout labelled by the configurations that chose it.

Relation to Remark 1. The agent uses the event decomposition to compare candidates by what they make possible next, information that only the later frames of a rollout provide. The +imagination agent also succeeds, but only after 61 imagination requests.

Case 5 · The end of the bed R2R-CE episode 150 · Qwen3.5-4B · diverging round 1

"Turn around and walk to the end of the bed. At the end of the bed turn right and walk to the hallway. Once at the entrance way, walk through the hallway to the next room."

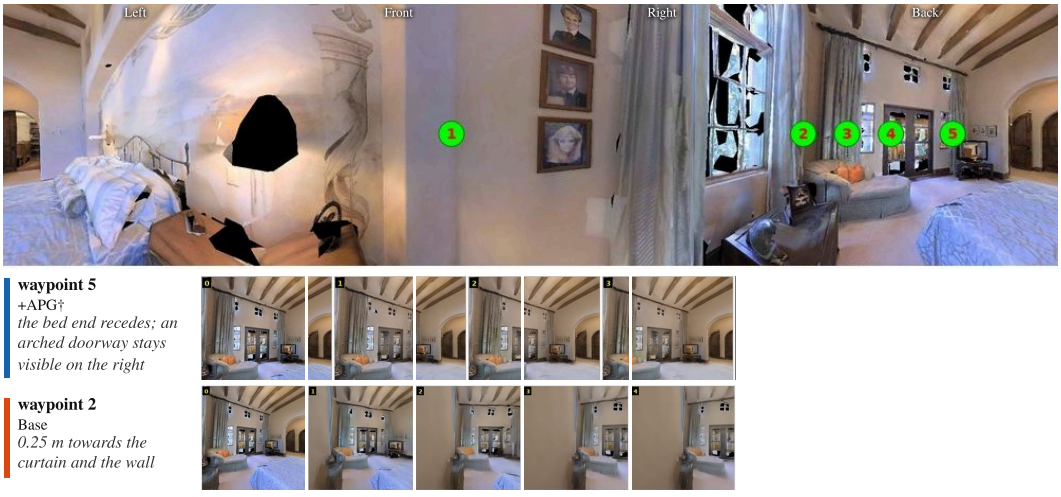


Figure 10: **Case 5.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 150, Qwen3.5-4B), with each rollout labelled by the configurations that chose it.

A.8.2 Remark 2. A rollout is only as useful as the agent's judgement of it (Cases 6 to 8)

Agents rarely call a rollout blurred or unreliable, in only 5, 4 and 5 rounds for the 4B, 9B and 27B agents, and three further episodes are solved with a justification that does not match the rollout (Case B).

Case 6. GPT-5.5 revises its choice after a rollout runs into a wall. In R2R-CE episode 1084 with GPT-5.5, the configurations diverge at round 3 (Figure 11), and the instruction reads "Exit the bathroom and enter the hall. Take a left in the hall and then enter another hall. Take a right at this hall, then take a left to enter another bathroom."

The proprietary agents have no base-agent run, so the evidence here comes from the agent's own reasoning before and after the only imagination request of the episode. Before imagining, GPT-5.5 reads waypoint 2 as the right-hand corner that the instruction asks for. After imagining waypoints 1 and 2 it switches to waypoint 1, because the rollout of waypoint 2 runs into a wall while that of waypoint 1 follows an open passage. Four moves later it stops after confirming from the real panorama that it is inside the second bathroom, whereas all ten other runs of this episode end with an nDTW of at most 0.07.

> GPT-5.5, round 3, before imagining. *"MODE: SCORED — I'm executing "take a right at this hall." Waypoint 2 appears to turn around the right-hand corner; waypoint 1 enters a room straight ahead, so I'll compare those candidates before moving."*

> GPT-5.5, round 3, after imagining. *"MODE: SCORED — [...] The first two events are complete. Waypoint 1's filmstrip follows a coherent open passage toward the next turn, while waypoint 2 runs into a wall/closed structure."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
GPT-5.5	**+APG†**	**success**	**0.58**	**0.79**	**6.8**	**6**	**1**	**stop**
Opus 5	+APG†	fail	3.52	0.01	38.2	29	20	stop
Qwen3.8-27B	+APG†	fail	12.61	0.00	32.5	27	10	stop
Qwen3.8-27B	Base	fail	3.97	0.01	26.0	21	0	stop

Relation to Remark 2. This is one of the few records that contain a judgement both before and after imagining in the same round. GPT-5.5 rarely imagines, with 0.2 requests per episode on R2R-CE, and here it does so exactly once, in a round it had marked for a structured comparison.

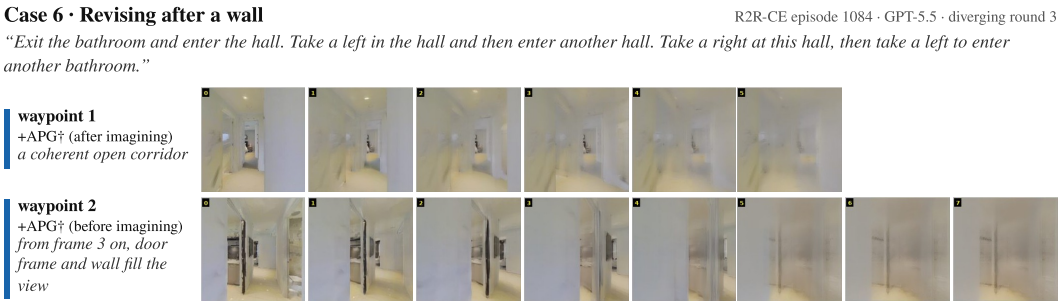


Figure 11: **Case 6.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 1084, GPT-5.5), with each rollout labelled by the configurations that chose it.

Case 7. GPT-5.5 recognises a rollout that barely moves. In R2R-CE episode 546 with GPT-5.5, the configurations diverge at round 4 (Figure 12), and the instruction reads "Turn right, and then make a right again entering the room that is on the right of the painting placed on the wall. Walk inside the room, and stop next to the mirror sliding closet doors."

This is the same episode as Case 8. At round 4, waypoints 4 and 5 both lie in a corridor lined with paintings. GPT-5.5 first intends to move to waypoint 4 but then switches to the structured comparison and imagines both candidates. The rollout of waypoint 4 consists of almost identical corridor frames, while that of waypoint 5 advances to the glass doors, so it picks waypoint 5 and stops 0.74 m from the goal. The 27B +APG† agent also reaches 0.74 m but runs out of wall-clock time, as described in Case 13.

> GPT-5.5, round 4, after imagining. *"MODE: SCORED — [...] Waypoint 5 ends farther inside the visible room than waypoint 4, and its coherent filmstrip advances toward the glass/mirrored doors; waypoint 4 barely moves along the same corridor."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
GPT-5.5	**+APG†**	**success**	**0.74**	**0.73**	**11.8**	**10**	**1**	**stop**
Opus 5	+APG†	success	1.77	0.54	13.0	10	6	stop
Qwen3.5-9B	+APG†	fail	8.99	0.33	4.7	4	3	stop
Qwen3.5-9B	Base	success	1.28	0.38	15.7	12	0	stop
Qwen3.8-27B	+APG†	fail	0.74	0.49	14.3	11	8	time-out

Relation to Remark 2. The protocol asks the agent to discount melted or duplicated structure but never states that a static rollout means no progress, and GPT-5.5 infers this from the rollout itself. The value of imagination therefore depends on judging the prediction, not only on understanding the image.

Case 7 · A rollout that barely moves R2R-CE episode 546 · GPT-5.5 · diverging round 4

"Turn right, and then make a right again entering the room that is on the right of the painting placed on the wall. Walk inside the room, and stop next to the mirror sliding closet doors."



Figure 12: **Case 7.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 546, GPT-5.5), with each rollout labelled by the configurations that chose it.

Case 8. A 9B agent reports an arrival that its rollout never shows. In R2R-CE episode 546 with Qwen3.5-9B, the configurations diverge at round 2 (Figure 13), and the instruction reads "Turn right, and then make a right again entering the room that is on the right of the painting placed on the wall. Walk inside the room, and stop next to the mirror sliding closet doors."

At round 2 the 9B +APG† agent requests a single rollout, for waypoint 4 at 1.25 m. All six frames show the same window, half a painting and the corner of a table, and the camera never passes through a doorway. The agent nevertheless attributes the final event of the instruction to this prediction and moves to waypoint 4, stopping 8.99 m from the goal after four moves, whereas the base agent succeeds at 1.28 m.

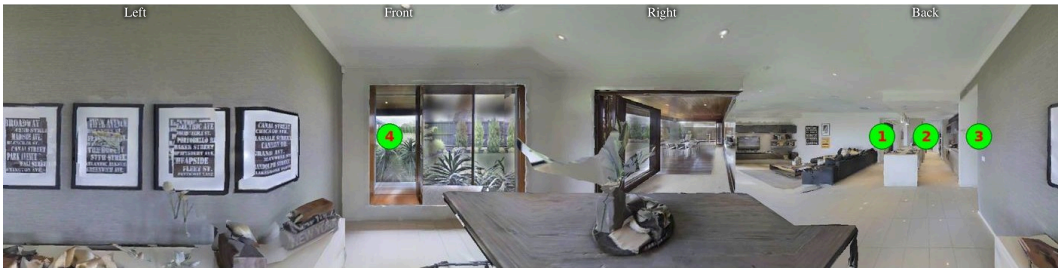
> Qwen3.5-9B +APG†, round 2. *"MODE: DIRECT [...] The predicted walk-through for waypoint 4 shows the robot entering that room, moving forward, and eventually reaching the sliding closet doors with the mirror."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.5-9B	Base	success	1.28	0.38	15.7	12	0	stop
Qwen3.5-9B	+Imag.	fail	8.39	0.05	15.8	16	20	stop
Qwen3.5-9B	$+\text{APG}_{\text{lean}}$	fail	3.50	0.36	14.0	10	7	stop
Qwen3.5-9B	+APG	fail	6.29	0.12	10.3	7	6	stop
Qwen3.5-9B	**+APG†**	**fail**	**8.99**	**0.33**	**4.7**	**4**	**3**	**stop**

Counterpoint to Remark 2. The endpoint of the instruction is copied into the description of the rollout, which then merely confirms a choice already made. The report reads as better grounded than the base agent's, and only the frames reveal that the evidence is absent. This is the reverse of the omission rule in APG, since content missing from the prediction is reported as evidence.

Case 8 · Evidence the rollout does not contain R2R-CE episode 546 · Qwen3.5-9B · diverging round 2

"Turn right, and then make a right again entering the room that is on the right of the painting placed on the wall. Walk inside the room, and stop next to the mirror sliding closet doors."



waypoint 4
Qwen3.5-9B +APG†
six frames of the same window; no doorway, room or closet

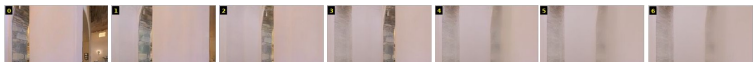


Figure 13: **Case 8.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 546, Qwen3.5-9B), with each rollout labelled by the configurations that chose it.

Case B. A correct choice with a mismatched justification. Not every correct choice rests on a correct reading of the rollout. Episode 191 is solved only by +APG† (0.53 m against 7.21 m), which imagines waypoints 1, 3 and 4 at round 5 and picks waypoint 3, stating that its rollout "clearly shows me passing through an archway into a bigger room". In fact the rollout of waypoint 3 approaches a wall and then blurs, and the archways it describes appear in the rollout of waypoint 4 (Figure 14). Waypoint 3 turns out to be correct, but the episode cannot count as evidence that imagination drove the decision. The same pattern appears in episode 677 (9B), whose structured report describes two nearly static rollouts as entering a side room and continuing down a corridor, and in episode 1077 (27B), whose justification describes the rollout of a candidate it did not choose. Together with Case 8, this shows that a sentence such as "the prediction shows" is not by itself evidence that the rollout was actually read.

Case B · correct choice, mismatched justification R2R-CE episode 191 · Qwen3.8-27B · round 5

waypoint 3
+APG† (chosen)
0.75 m towards a wall; later frames blur



waypoint 4
Base
through a series of arches into the dining room



Figure 14: **Case B** (R2R-CE episode 191, Qwen3.8-27B). +APG† chooses waypoint 3 but describes archways that appear only in the rollout of waypoint 4.

A.8.3 Remark 3. Event order keeps imagination aligned with the instruction (Cases 9 to 12)

Rejecting a candidate for skipping an earlier event occurs in 34 rounds for the 27B +APG† agent and never for its base agent, while a stop that cites a prediction in an episode that fails occurs in 22, 8 and 2 episodes for the 4B, 9B and 27B agents.

Case 9. APG refuses a shortcut to a visible goal. In R2R-CE episode 620 with Qwen3.8-27B, the configurations diverge at round 1 (Figure 15), and the instruction reads "Walk forward along the

railing to the balcony. Go past the cabinet on your left. Stop on the rug before you pass the first bed."

In a loft the Right view faces a large bed with a rug in front of it, while the Back view shows the passage along the railing, and the instruction lists three events in order. The base agent heads for the rug in front of the bed and stops after 1.75 m, and the +imagination agent imagines waypoints 1 and 5 and stops without moving, claiming to have "come along the railing past the cabinet". +APG† lists the three events, notes that none of them is complete and rejects waypoint 1 because it would skip the first two. It then walks along the railing and past the cabinet and stops on the rug 1.83 m from the goal.

> +APG†, round 1. *"MODE: SCORED [...] Events: 1) Walk along railing to balcony, 2) Pass cabinet on left, 3) Stop on rug before first bed. [...] None of the events are complete yet. Waypoint 1: Takes me directly toward the bed (skipping the 'along the railing' and 'past the cabinet' events). This would jump ahead too far."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.8-27B	Base	fail	6.28	0.41	1.8	1	0	stop
Qwen3.8-27B	+Imag.	fail	5.84	0.42	0.0	0	–	stop
Qwen3.8-27B	**+APG†**	**success**	**1.83**	**0.81**	**9.0**	**7**	**4**	**stop**

Relation to Remark 3. The chosen rollout is nearly static, so the decision rests on two rules of the protocol, namely that a candidate showing the destination while skipping an unfinished event cannot win and that stopping is decided from real observations rather than from a prediction that looks like the goal.

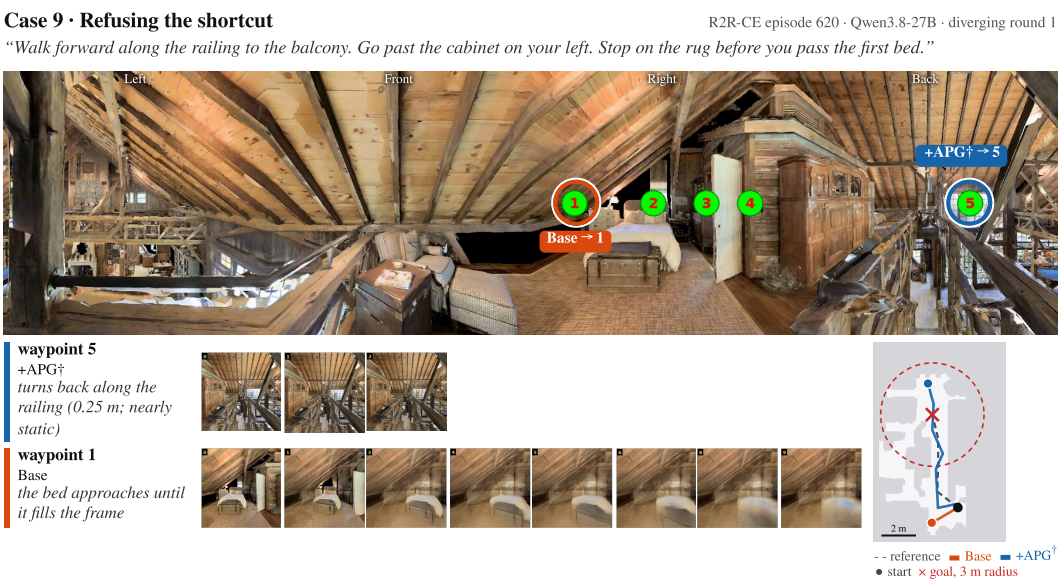


Figure 15: **Case 9.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 620, Qwen3.8-27B), with each rollout labelled by the configurations that chose it. The map shows the trajectories of the base agent and +APG† with the 3 m success radius.

Case 10. Only the agent that checks the next event climbs the stairs. In R2R-CE episode 371 with Qwen3.5-4B, the configurations diverge at round 5 (Figure 16), and the instruction reads "Walk through the dining room and kitchen air to the short flight of stairs. Walk up the short flight of stairs. Wait by the door at the top of the stairs."

The three 4B agents make identical choices for four rounds and stand at the foot of the stairs at round 5, with the steps and the door at their top in view. Both the base agent and +APG† first state that they have reached the top of the stairs. The base agent acts on this belief and stops one round later,

4.40 m from the goal, with a report that places it both in the dining room and at the top of the stairs. +APG† instead imagines waypoints 1 and 2, recognises walking up the stairs as the next unfinished event, climbs and stops 0.12 m from the goal.

+APG†, round 5, before imagining. *"You've arrived at the top of the stairs [...] Let me look at the current view to confirm I'm at the right place."*

+APG†, round 5, after imagining. *"MODE: SCORED [...] The next part of the instruction is to 'Walk up the short flight of stairs.' [...] Waypoint #2 is clearly the right choice to complete the next required event of walking up the stairs."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.5-4B	Base	fail	4.40	0.81	7.8	5	0	stop
Qwen3.5-4B	+Imag.	success	2.11	0.90	7.3	5	7	stop
Qwen3.5-4B	**+APG†**	**success**	**0.12**	**0.94**	**9.5**	**6**	**6**	**stop**

Relation to Remark 3. The successful and the failing agent hold the same false belief when they diverge, and checking the next unfinished event before stopping is what prevents the false arrival.

Case 10 · Not yet up the stairs R2R-CE episode 371 · Qwen3.5-4B · diverging round 5

"Walk through the dining room and kitchen air to the short flight of stairs. Walk up the short flight of stairs. Wait by the door at the top of the stairs."

waypoint 2
+APG†
climbs the stairs; the door at the top draws closer



Figure 16: **Case 10.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 371, Qwen3.5-4B), with each rollout labelled by the configurations that chose it.

Case 11. Only one of three rollouts actually leaves the bedroom. In R2R-CE episode 810 with GPT-5.5, the configurations diverge at round 6 (Figure 17), and the instruction reads "Leave the closet, and walk out of the bedroom. Stop once you exit the bedroom door."

Three exits are visible from the bedroom. At round 6 GPT-5.5 switches to the structured comparison to check which of them actually crosses the bedroom threshold and imagines waypoints 1, 3 and 4. Its report describes all three rollouts in agreement with the frames, and it picks waypoint 1 and stops 0.07 m from the goal. For the small models the episode shows a clear contrast, as every 4B and 9B configuration that requested at least one rollout succeeds and every one that did not fails, including the 9B +APG† agent, which never imagined. All five 27B configurations and Opus 5 succeed.

GPT-5.5, round 6, after imagining. *"MODE: SCORED EVENTS: leave the closet; cross the bedroom; exit through the bedroom door. The closet has been left, so 1 of 3 events is complete. Waypoint 1 coherently passes through the archway into the corridor, while 3 remains inside near another doorway and 4 leads toward the chair/window."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
GPT-5.5	**+APG†**	**success**	**0.07**	**0.91**	**9.5**	**7**	**1**	**stop**
Qwen3.5-4B	Base	fail	8.53	0.10	31.4	–	0	move budget
Qwen3.5-4B	+Imag.	success	0.44	0.95	6.7	6	5	stop
Qwen3.5-4B	$+\text{APG}_{\text{lean}}$	success	1.63	0.86	9.5	8	9	stop
Qwen3.5-4B	+APG†	success	1.63	0.37	32.2	28	36	stop
Qwen3.5-9B	Base	fail	3.82	0.13	33.6	30	0	move budget
Qwen3.5-9B	+Imag.	success	2.18	0.40	19.0	14	13	stop
Qwen3.5-9B	$+\text{APG}_{\text{lean}}$	success	1.64	0.82	11.0	9	7	stop
Qwen3.5-9B	+APG	success	2.01	0.90	9.2	7	2	stop
Qwen3.5-9B	+APG†	fail	6.48	0.47	2.0	3	0	stop

Relation to Remark 3. Whether the agent has left a room is a threshold crossing that the panorama can only suggest, and the rollout turns it into something that can be checked for each candidate. The

report follows the protocol by listing the events, marking which are complete and stating on which side of the threshold each candidate ends.

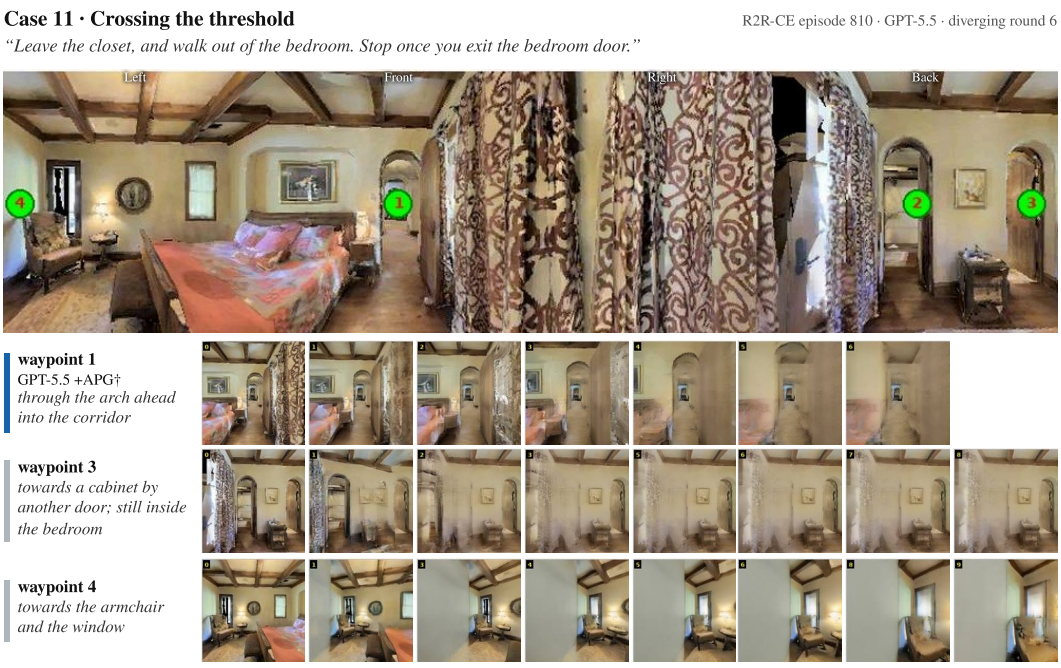


Figure 17: **Case 11.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 810, GPT-5.5), with each rollout labelled by the configurations that chose it.

Case 12. A 9B agent stops on an imagined arrival. In R2R-CE episode 765 with Qwen3.5-9B, the configurations diverge at round 3 (Figure 18), and the instruction reads "Turn right at the post and walk forward alongside the glass doors on the left. Wait in front of the pool table."

At round 3 the pool table is in the Front view several metres away, and +APG† imagines only waypoint 1, which lies 1.75 m towards the table. It then stops without moving, citing the arrival frame of the rollout, and ends 3.17 m from the goal, just 0.17 m outside the success radius. The other four 9B configurations all succeed in this episode.

> Qwen3.5-9B +APG†, round 3. *"MODE: DIRECT [...] The predicted filmstrip for waypoint 1 shows that I will arrive directly in front of the pool table, which is exactly where I need to be to wait. [...] I should stop now as I have reached the endpoint."*

> Protocol, step 5 (in every system prompt). *"And never stop() because a PREDICTION looks like the goal — stopping is decided from what you actually see."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End
Qwen3.5-9B	Base	success	1.43	0.89	5.0	3	0	stop
Qwen3.5-9B	+Imag.	success	2.30	0.19	26.1	16	21	stop
Qwen3.5-9B	+APG$_{\text{lean}}$	success	1.28	0.35	24.1	13	9	stop
Qwen3.5-9B	+APG	success	0.63	0.79	10.5	6	10	stop
Qwen3.5-9B	**+APG†**	**fail**	**3.17**	**0.74**	**3.3**	**2**	**1**	**stop**

Counterpoint to Remark 3. The agent violates an explicit rule while phrasing its decision in the protocol's own terms. This error can only arise with a world model, because an agent without imagination never sees an arrival it has not reached, and Table 7 gives a lower bound on how often it occurs (rule R4).

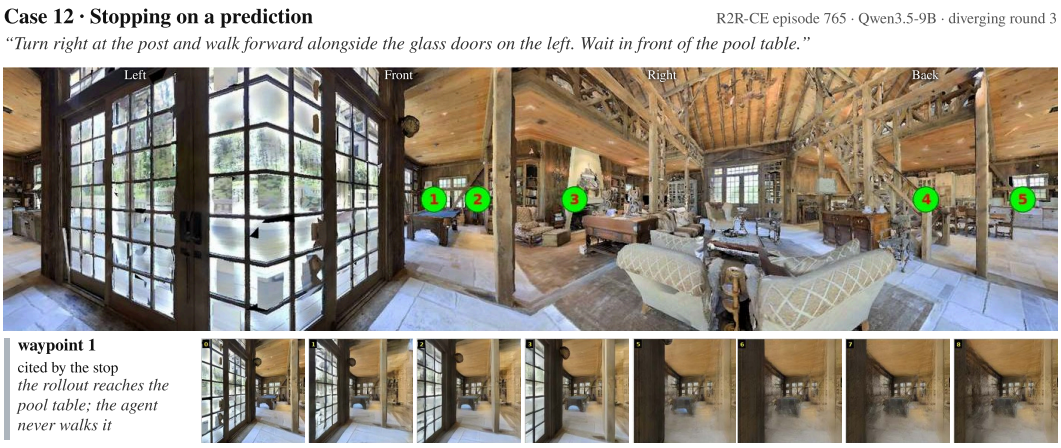


Figure 18: **Case 12.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 765, Qwen3.5-9B), with each rollout labelled by the configurations that chose it.

A.8.4 Remark 4. Imagination costs time (Case 13)

In four episodes that the 27B base agent solves, +APG† is cut off by the wall-clock limit while already within 3 m of the goal.

Case 13. Deliberating at the goal ends in a time-out. In R2R-CE episode 244 with Qwen3.8-27B, the configurations diverge at round 6 (Figure 19), and the instruction reads "Go through the door and turn left. Go to the left of the stairs. Stop in the doorway to the left of the white double doors."

+APG† and the base agent choose the same waypoints for five rounds and stand at the same pose at round 6, 0.22 m from the goal, where the base agent stops and succeeds after 224 s. +APG† imagines waypoints 1 and 2, which lead deeper into the bedroom, concludes that it is inside the bedroom rather than in the doorway and backs out into the hallway. Fifteen rounds and ten imagination requests later it is still moving between the stairs and the bedroom when the 40-minute wall-clock limit ends the episode at 2540 s, 2.17 m from the goal and inside the success radius but without having stopped.

> +APG†, round 5. *"[...] Waypoint 4 is further into the bedroom itself. I need to be *in the doorway*, not in the room."*
>
> +APG†, round 6. *"MODE: DIRECT Looking at my situation: I'm currently inside a bedroom. [...] I need to go back to that doorway (the threshold), not stay inside the bedroom."*

Agent	Configuration	Result	NE (m)	nDTW	Path (m)	Moves	Imag.	End	Time (s)
Qwen3.8-27B	Base	success	0.22	0.96	7.5	5	0	stop	224
Qwen3.8-27B	+Imag.	success	0.24	0.64	19.5	13	10	stop	1258
Qwen3.8-27B	+APG$_{\text{lean}}$	success	0.22	0.96	7.5	5	3	stop	410
Qwen3.8-27B	+APG	success	0.87	0.49	17.3	12	9	stop	2226
Qwen3.8-27B	**+APG†**	**fail**	**2.17**	**0.21**	**27.6**	**21**	**13**	**time-out**	**2540**

Counterpoint to Remark 4. The rollouts are read correctly, but the criterion the agent set at round 5, to stand in the doorway rather than in the room, is used at round 6 to overturn a correct position, so finer events give it another chance to second-guess itself at the goal. The cost is paid in wall-clock time, as a 27B round takes more than 30 s and each imagination request 8 to 15 s.

Case 13 · Deliberating at the goal R2R-CE episode 244 · Qwen3.8-27B · diverging round 6

"Go through the door and turn left. Go to the left of the stairs. Stop in the doorway to the left of the white double doors."



waypoint 1
imagined at the goal
deeper into the bedroom

Figure 19: **Case 13.** Panorama and imagined rollouts where the configurations diverge (R2R-CE episode 244, Qwen3.8-27B), with each rollout labelled by the configurations that chose it.