JANUS: SPATIAL BELIEF AND EVIDENCE IMAGINATION FOR EMBODIED ACTIVE PERCEPTION

Anonymous authors
Paper under double-blind review

"The longer you can look back, the farther you can look forward."
—Winston S. Churchill[1]

ABSTRACT

Active perception is an essential capability for embodied agents, providing the ability to selectively acquire the visual evidence required for downstream decision-making. Recent approaches leverage multimodal large language models (MLLMs) to decide where to observe next. However, without persistent memory and explicit spatial representations, these approaches struggle to maintain spatial consistency across observations and identify informative regions for further exploration. In this work, we seek to unlock the memory and associative reasoning capabilities of MLLMs, investigating whether an agent can efficiently accomplish visual perception tasks through a sequence of selective egocentric observations. To this end, we introduce **JANUS**, a framework that couples persistent spatial belief with imagination-guided next-view decisions. Specifically, we develop a *Hierarchical Spatial Belief* module that organizes accumulated real observations into a hierarchical representation over scenes, regions, objects, and their attributes, while preserving confidence estimates and links to the supporting visual evidence. We further introduce a *Belief-Conditioned Evidence Imagination* module that conditions on the current spatial belief to anticipate informative candidate observations, using predicted latent evidence to reactivate relevant past observations and guide next-view selection. Together, these modules connect accumulated observations with future information needs, enabling belief-guided decisions about where to observe next. Experiments on challenging perception benchmarks demonstrate the effectiveness of JANUS for embodied active perception. The code will be released.

1 INTRODUCTION

Embodied agents often need to complete tasks with only partial visual access to their surroundings. Relevant objects may lie outside the current field of view, while important details may be occluded or visible only from particular viewpoints. Effective task execution therefore requires not only interpreting available observations but also acquiring the visual evidence that remains missing. This motivates active perception, in which agents select observations to gather task-relevant information rather than passively process a fixed set of images Sakamoto et al. (2026). Multimodal large language models (MLLMs) increasingly control what to inspect: ACTIVE-o3 selects informative image regions, Thinking in 360° controls egocentric views, and MemoryEQA connects exploration to answering (Zhu et al., 2025; Yu et al., 2026; Zhai et al., 2025). Yet visual recognition does not guarantee effective sequential observation. Viewpoint-control benchmarks expose repeated actions and difficulty acquiring missing evidence; cross-view reasoning also remains challenging (Sakamoto et al., 2026; Xue et al., 2026). MemoryEQA and 3D-Mem show that retaining scene experience can help subsequent decisions (Zhai et al., 2025; Yang et al., 2025).

[1]Winston S. Churchill was a British statesman, writer, and twice Prime Minister of the United Kingdom.

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.

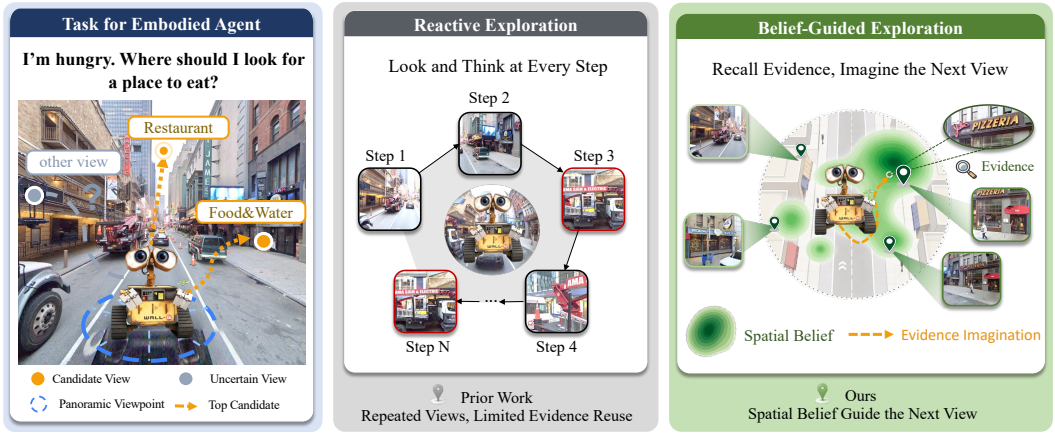

Figure 1: **Spatial beliefs guide where to look next. Left:** Given a task, an embodied agent must select informative views from a panoramic environment, such as locating visual evidence of a place to eat. **Middle:** Reactive exploration reasons over successive observations but can revisit the same regions with limited evidence reuse. **Right:** JANUS maintains a hierarchical spatial belief grounded in real observations, recalls task-relevant evidence, and anticipates what candidate views may reveal to guide the next observation. New visual evidence supports task completion and belief revision. Green regions illustrate spatial beliefs, while orange dashed arrows indicate imagination-guided view selection.

Figure 1 illustrates the distinction: revisiting a shop is useful when it resolves an unreadable sign, but redundant when it adds no relevant evidence. An agent therefore needs to connect *what it observed, what it currently believes, and what a new view could resolve*. Building on memory-guided planning and scene imagination, including ImagineNav++ (Wang et al., 2026), we introduce **JANUS** to study this connection through source-grounded spatial belief and latent evidence imagination.

Specifically, JANUS separates historical evidence from the agent's current interpretation of that evidence. Its *Hierarchical Spatial Belief* module organizes scene, region, object, and attribute claims together with their confidence and supporting real observations. Knowing that a shop exists need not imply knowing whether it is a pharmacy. Its *Belief-Conditioned Evidence Imagination* module anticipates candidate visual features, reactivates relevant observations, and guides further inspection. This creates a two-way interaction between recall and prediction: the current belief supplies context for imagination, and imagination supplies a visual query for retrieving related real experiences. Real observations then confirm, weaken, or correct the interpretation without rewriting the historical image evidence.

Our contributions are threefold:

- **A spatial-belief formulation.** We distinguish a source-grounded observation bank from a revisable hierarchy of spatial and attribute claims, making evidence provenance and uncertainty explicit rather than equating stored observations with accepted facts.
- **Evidence imagination with associative recall.** We couple latent future-view prediction to real-evidence retrieval and observation selection, while prohibiting predicted content from becoming factual support. We specify the full bidirectional design and identify the approximations used by the implemented controllers.
- **An empirical assessment across tasks and actors.** We report paired QA and active-search measurements, geolocation results, and mechanism diagnostics, including regressions and incomplete support for the proposed imagination mechanism.

2 Related Work

Spatial and episodic memory. Neural Map and Cognitive Mapping and Planning organize experience spatially (Parisotto & Salakhutdinov, 2018; Gupta et al., 2017), whereas Scene Memory

Transformer and Episodic Transformer retrieve historical observations for control (Fang et al., 2019; Pashevich et al., 2021). VLMaps and ConceptGraphs connect language with spatial features and object-centric relations (Huang et al., 2023; Gu et al., 2024); DUET combines coarse and fine reasoning (Chen et al., 2022), and MrSteve retrieves what–where–when experiences (Park et al., 2025). JANUS distinguishes stored evidence from its revisable interpretation, linking confidence-bearing claims to source views. Its current confidence scores remain heuristic rather than calibrated belief posteriors.

Imagination and observation planning. Potential functions and occupancy anticipation guide exploration toward unseen evidence (Ramakrishnan et al., 2022; 2020). Latent world models support planning without generating every future pixel (Hafner et al., 2019; 2020), while PathDreamer and ImagineNav predict prospective observations for navigation (Koh et al., 2021; Zhao et al., 2025). ImagineNav++ combines sparse-to-dense keyframe memory with imagined views (Wang et al., 2026). JANUS instead uses predicted visual latents as both action cues and queries over real evidence; predictions cannot become factual support. The full candidate-wise recall–reranking loop remains an extension, and current experiments do not establish superiority over these prior systems. Appendix B provides the extended literature discussion.

3 Problem Formulation

We study *embodied active perception*: given an instruction q, an agent actively acquires limited-field-of-view observations to produce a task-specific prediction. The agent remains at a fixed position and controls its camera pose $p_t = (\theta_t, \phi_t, z_t)$, representing yaw, pitch, and zoom, with zoom enabled only when supported by the protocol. In our benchmarks, an egocentric view $o_t = \mathcal{R}(\mathcal{E}, p_t)$ is rendered from a real-world panorama $\mathcal{E}$, which is never directly exposed to the agent.

Given its observation history $h_t = \{(p_i, o_i)\}_{i=0}^{t}$, the policy either acquires another view or submits a prediction:

$$a_t \sim \pi(\cdot \mid q, h_t), \qquad a_t \in \mathcal{A}_{\text{obs}} \cup \{\text{Submit}(y)\}. \tag{1}$$

For task k, the objective is to maximize its evaluation score under the prescribed interaction budget:

$$\pi_k^\star = \arg\max_{\pi} \mathbb{E}_{\mathcal{D}_k, \pi}[S_k(y, y^\star)], \qquad C(\tau) \le B_k, \tag{2}$$

where $y^\star$ is the hidden reference annotation and $C(\tau)$ counts budgeted interactions. The *memory-first* protocol provides task-blind observations before revealing q, whereas the *online* protocol reveals q from the outset without an additional pre-scan. We instantiate six tasks, illustrated in Appendix A:

Humanoid Object Search (HOS). Locate a specified object and submit a yaw–pitch direction $(\hat{\theta}, \hat{\phi})$ that centers the target, evaluated by the official object-search scorer.

Humanoid Path Search (HPS). Identify a feasible route toward a requested destination and submit its yaw direction $\hat{\theta}_{\text{path}}$, evaluated by the official path-search scorer. Physical traversal is not required.

Embodied Geolocation (Geo). Integrate geographic cues across acquired views to predict coordinates and location labels, evaluated by geographic-label accuracy, distance-based metrics, and geolocation score.

Foundational Perception (FP). Answer questions about observable architecture, infrastructure, signage, terrain, and vegetation by inspecting relevant visual evidence.

Spatial Awareness (SA). Estimate positions, directions, depths, distances, and spatial relations, including cross-view reasoning. Predictions retain the reference frame specified by the original question.

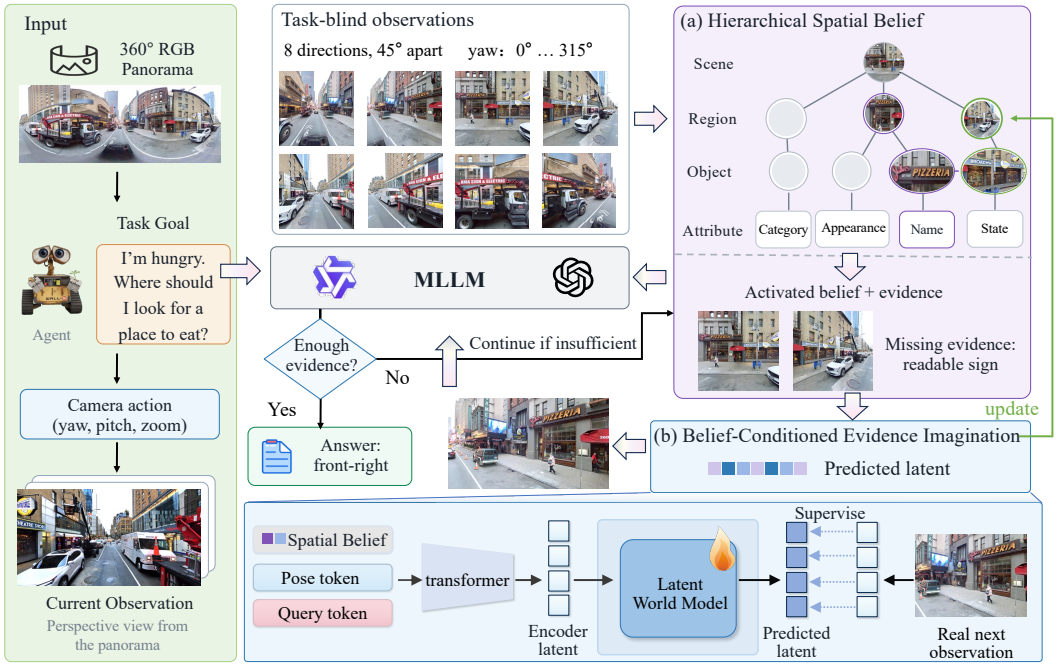

Figure 2: **Overview of JANUS.** Given a task and a 360° RGB panorama, the agent controls yaw, pitch, and zoom to acquire perspective observations. Eight task-blind views, spaced 45° apart, initialize its spatial representation. **(a) Hierarchical Spatial Belief (HSB)** organizes scene, region, object, and attribute information with supporting visual evidence. Task-conditioned activation recalls relevant observations and identifies missing evidence. **(b) Belief-Conditioned Evidence Imagination (BEI)** predicts candidate-view features to guide evidence retrieval and next-view selection. A frozen MLLM examines real observations and recalled evidence to answer or continue exploration, while newly acquired evidence supports belief revision. **Bottom: BEI pre-training.** The latent world model encodes spatial-memory features, pose tokens, and a candidate-action query to predict the visual representation of a future view, supervised by features extracted from the corresponding real observation. Imagined features guide exploration; only real visual evidence supports factual belief updates and answers.

Common Sense (CS). Combine visual evidence with contextual and functional knowledge to answer single-view, dual-view, and cross-view questions; these categories follow the benchmark's original organization.

For FP, SA, and CS, we preserve the original questions, answer formats, and scorers while enabling active observation. Answers are accompanied by supporting real-view identifiers and camera poses; grounding is assessed separately from answer correctness.

4 METHODOLOGY

4.1 OVERVIEW

As illustrated in Figure 2, JANUS connects past observations, current spatial understanding, and future evidence acquisition through two complementary modules: *Hierarchical Spatial Belief* (HSB) and *Belief-Conditioned Evidence Imagination* (BEI). HSB organizes observations into confidence-bearing spatial claims, while BEI anticipates candidate observations to guide evidence retrieval and camera control.

In the memory-first setting, eight task-blind views, spaced 45° apart, initialize the spatial representation. Given a subsequent instruction q, JANUS activates relevant beliefs and their supporting observations, identifies missing evidence, and selects an informative camera action. A frozen MLLM

examines the acquired and recalled evidence to either submit an answer or continue exploration. New observations update the spatial belief, forming the iterative loop shown in Figure 2.

4.2 HIERARCHICAL SPATIAL BELIEF

Observation-grounded construction. Figure 2(a) organizes spatial knowledge into four levels: *Scene–Region–Object–Attribute*. Scene nodes summarize the environment, region nodes organize its directional context, object nodes represent landmarks and entities, and attribute records describe their category, appearance, name, and state.

We distinguish the observation bank E_t from its current interpretation B_t:

$$E_t = \{(I_i, p_i, \mathbf{z}_i, t_i)\}_{i \le t}, \qquad B_t = (V_t, \mathcal{R}_t, \mathcal{A}_t), \tag{3}$$

where each observation retains its image I_i, camera pose p_i, visual feature $\mathbf{z}_i$, and timestamp t_i. The belief comprises hierarchical nodes V_t, relational links $\mathcal{R}_t$, and attribute claims $\mathcal{A}_t$. Each claim stores its confidence and supporting-view references. Hierarchical links preserve spatial context, while lateral associations connect related entities and observations.

Task-conditioned activation. Given q, JANUS retrieves relevant nodes and expands their hierarchical and associative context:

$$(\mathcal{G}_t, \mathcal{D}_t) = \mathrm{Activate}(q, B_t, E_t), \tag{4}$$

where $\mathcal{G}_t$ contains the activated belief subgraph and supporting observations, and $\mathcal{D}_t$ describes the evidence still needed to resolve the task. Activation includes both cross-level context, such as an object's parent region, and related same-level nodes.

For the request "I'm hungry. Where should I look for a place to eat ? " in Figure 2, a storefront may be relevant even when its function is uncertain. HSB recalls the storefront and its surrounding context while identifying a readable sign as the missing evidence.

Observation-grounded revision. After a new observation e_{t+1} is acquired, JANUS extends the evidence bank and revises the associated beliefs:

$$E_{t+1} = E_t \cup \{e_{t+1}\}, \qquad B_{t+1} = \mathrm{Revise}(B_t, e_{t+1}). \tag{5}$$

New details add attributes, supporting observations strengthen existing claims, and contradictory observations weaken or correct previous interpretations. Revision operates on individual claims: reading a restaurant sign refines the storefront's identity without replacing its previously observed location or appearance.

4.3 BELIEF-CONDITIONED EVIDENCE IMAGINATION

BEI uses the activated spatial context to anticipate which observation could resolve the current evidence gap. As shown in Figure 2(b), it predicts visual representations rather than synthesizing future images.

Candidate-observation prediction. For each candidate camera configuration $a = (\theta, \phi, z)$, we encode the spatial belief, pose, and candidate-observation query as tokens:

$$\begin{aligned} \mathbf{h}_{t,a} &= T_\theta\big(\mathbf{b}_t, \mathbf{p}(a), \mathbf{v}(a)\big), \\ \widehat{\mathbf{z}}_{t,a} &= F_\theta(\mathbf{h}_{t,a}), \end{aligned} \tag{6}$$

where $\mathbf{b}_t$ summarizes the available spatial context, $\mathbf{p}(a)$ encodes the candidate pose, and $\mathbf{v}(a)$ is the observation-query token used to read out the candidate representation. The Transformer produces an encoded latent, which the latent world model maps to the predicted next-view feature $\widehat{\mathbf{z}}_{t,a}$. The observation-query token is distinct from the language instruction q, which conditions evidence activation and candidate evaluation.

Imagination-guided recall and view selection. Predicted features serve two complementary roles. First, their compatibility with the task and missing evidence helps rank candidate views. Second, the selected prediction acts as a visual query over the observation bank:

$$\begin{aligned} a^\star &= \arg\max_a U(q, \mathcal{D}_t, \mathcal{G}_t, \widehat{\mathbf{z}}_{t,a}, a), \\ \mathcal{C}_t &= \mathrm{TopK}_{e_i \in E_t}\, r(\mathbf{z}_i, \mathbf{z}_q, \widehat{\mathbf{z}}_{t,a^\star}), \end{aligned} \tag{7}$$

where U evaluates candidate usefulness and r combines task-to-observation and predicted-to-observation similarity. This creates two distinct activations: the instruction retrieves semantically relevant spatial context, while the imagined visual query recalls experiences resembling the anticipated observation.

In the example, a predicted closer storefront view can reactivate an earlier image containing part of its sign. The retrieved evidence may already resolve the question; otherwise, it informs the next inspection.

Evidence-driven decision. Following the central decision branch in Figure 2, the frozen MLLM examines the current observation, activated beliefs, and recalled evidence:

$$d_t = \mathrm{MLLM}(q, o_t, \mathcal{G}_t, \mathcal{C}_t, a^\star). \tag{8}$$

If the available evidence is sufficient, the agent submits a task-specific answer. Otherwise, it executes a camera action, acquires another view, and updates HSB. For the illustrated task, inspecting the storefront may support a final directional answer such as "front-right."

Latent world-model pre-training. The bottom panel of Figure 2 illustrates training with real candidate observations. Given a partially observed spatial context and a candidate pose, the model predicts the feature of the corresponding held-out view:

$$\mathbf{z}_a^{\mathrm{real}} = \mathrm{Enc}_v\big(\mathcal{R}(\mathcal{E}, a)\big), \qquad \mathcal{L}_{\mathrm{pred}} = 1 - \cos\big(\widehat{\mathbf{z}}_{t,a}, \mathbf{z}_a^{\mathrm{real}}\big). \tag{9}$$

Training encourages the predicted latent to capture the visual information revealed by the candidate action. The visual encoder and MLLM remain frozen, while the latent prediction module is optimized.

5 Experiments

5.1 Experimental Setup

Benchmark and Metrics. We evaluate JANUS on ERGeoBench (Xue et al., 2026) and H* Bench (Yu et al., 2026). ERGeoBench evaluates four capabilities, including Foundational Perception (FP), Spatial Awareness (SA), Common Sense (CS), and embodied geo-localization. We split ERGeoBench by unique panorama ID into training, validation, and test sets with a 70%/10%/20% ratio, ensuring that all tasks and samples associated with the same panorama remain in the same split. H* Bench evaluates Humanoid Object Search (HOS) and Humanoid Path Search (HPS); we follow its released training and evaluation splits. For evaluation, we follow the official metrics of the two benchmarks. On ERGeoBench, we report accuracy for FP, SA, and CS, and use the Geo-localization Score (GLS) as the primary metric for geographic localization. On H* Bench, we report the official Success Rate (%) for HOS and HPS, where success is determined by whether the predicted search direction satisfies the benchmark-defined target criterion.

Baselines. We evaluate JANUS across a diverse set of open- and closed-source MLLMs. The open-source models include Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL2.5-8B, and InternVL3-8B, while the closed-source models include Qwen-VL Max, GPT-4o, and the Gemini series. On H* Bench, we additionally compare with the SFT-based HVS-8B Yu et al. (2026) as a task-specific reference.

Implementation Details. The agent receives only egocentric images rather than the full equirectangular panorama. On ERGeoBench, JANUS first observes eight task-blind horizontal views at 45° intervals, followed by at most two task-dependent inspections for QA or seven for geolocation.

Table 1: Results on H* Bench. Success rates (%) are reported overall and by difficulty. **Bold** indicate the better result within each paired comparison. * denotes models fine-tuned with supervised fine-tuning (SFT).

	Humanoid Object Search				Humanoid Path Search				
Model	Overall	Easy	Medium	Hard	Overall	Easy	Medium	Hard	Extreme
Closed-Source Models									
Qwen-VL Max	33.75	71.43	47.06	19.76	23.75	25.00	30.77	11.43	24.00
Qwen-VL Max **+ JANUS**	**35.83**	**73.21**	47.06	**22.16**	23.75	25.00	30.77	11.43	24.00
Improvement	+2.08	+1.78	0.00	+2.40	0.00	0.00	0.00	0.00	0.00
GPT-4o	19.75	18.17	17.35	20.92	23.69	**26.80**	22.59	26.17	**13.89**
GPT-4o **+ JANUS**	**35.83**	**62.50**	**23.53**	**28.14**	**26.88**	20.83	**40.38**	**28.57**	8.00
Improvement	+16.08	+44.33	+6.18	+7.22	+3.19	-5.97	+17.79	+2.40	-5.89
Open-Source Models									
Qwen2.5-VL-7B	11.38	23.42	**9.10**	7.02	6.31	**9.00**	5.92	5.84	1.85
Qwen2.5-VL-7B **+ JANUS**	**21.67**	**30.36**	5.88	**20.36**	**13.75**	6.25	**23.08**	**8.57**	**16.00**
Improvement	+10.29	+6.94	-3.22	+13.34	+7.44	-2.75	+17.16	+2.73	+14.15
Qwen3-VL-8B	21.17	57.72	26.14	6.65	14.44	**17.60**	12.72	14.72	**10.19**
Qwen3-VL-8B **+ JANUS**	**22.92**	**60.71**	**35.29**	**8.98**	**17.50**	16.67	**21.15**	**20.00**	8.00
Improvement	+1.75	+2.99	+9.15	+2.33	+3.06	-0.93	+8.43	+5.28	-2.19
InternVL2.5-8B	1.67	5.36	5.88	0.00	1.88	2.08	3.85	0.00	0.00
InternVL2.5-8B **+ JANUS**	**17.50**	**16.07**	**29.41**	**16.77**	**10.63**	**6.25**	**21.15**	0.00	**12.00**
Improvement	+15.83	+10.71	+23.53	+16.77	+8.75	+4.17	+17.30	0.00	+12.00
InternVL3-8B	12.08	41.07	11.76	2.40	11.25	12.50	21.15	0.00	4.00
InternVL3-8B **+ JANUS**	**23.33**	**44.64**	11.76	**17.37**	**15.00**	**14.58**	**26.92**	**2.86**	**8.00**
Improvement	+11.25	+3.57	0.00	+14.97	+3.75	+2.08	+5.77	+2.86	+4.00
HVS-8B* Yu et al. (2026)	60.29	**72.03**	**48.30**	57.12	32.87	43.80	28.07	**35.05**	13.43
Qwen3-VL-8B* **+ JANUS**	**61.25**	71.43	41.18	**59.88**	**33.13**	**45.83**	**28.85**	28.57	**24.00**
Improvement	+0.96	-0.60	-7.12	+2.76	+0.26	+2.03	+0.78	-6.48	+10.57

H* uses no additional pre-scan and resets memory between episodes. QA uses a single answer router trained on Qwen2.5-VL-7B training candidates and calibrated on disjoint validation panoramas, with its weights and thresholds transferred unchanged across models. The latent world model is trained on the released H* training scenes and transferred to ERGeoBench without adaptation. ERGeoBench renders 640×640 images, and the actor uses a temperature of 0.1 with up to 1,200 output tokens per request.

5.2 Main Results

Tables 1 and 2 evaluate JANUS from complementary aspects of active perception. Overall, the results show that coupling persistent spatial belief with imagination-guided observation selection improves the agent's ability to preserve spatial context, identify missing evidence, and acquire informative views across different model families.

As shown in Table 1, JANUS yields substantial gains on embodied object and path search, where successful decisions directly depend on selecting informative viewing directions. Qwen2.5-VL-7B improves by 10.29% on HOS and 7.44% on HPS, while InternVL2.5-8B and GPT-4o achieve HOS gains of 15.83% and 16.08%, respectively. These improvements are consistent with the role of BEI, which anticipates task-relevant visual evidence and uses it to guide next-view selection. The SFT-based Qwen3-VL-8B with JANUS further reaches 61.25% HOS and 33.13% HPS, compared with 60.29% and 32.87% for HVS-8B.

Table 2 further shows the benefit of maintaining a structured spatial belief across observations. JANUS improves QA All from 61.35% to 62.37% on Qwen2.5-VL-7B, from 66.88% to 68.84% on Qwen3-VL-8B, and from 60.12% to 61.82% on InternVL2.5-8B. Notably, Spatial Awareness improves by 1.67% and 1.62% on Qwen2.5-VL-7B and InternVL2.5-8B, respectively, which is consistent with HSB's explicit organization of observations into spatially linked scene, region, object, and attribute representations. For embodied geolocation, JANUS raises GLS from 62.07 to 65.26 on Qwen2.5-VL-7B, and also improves Qwen3-VL-8B and GPT-4o from 47.61 to 48.85 and 66.11 to 68.23, respectively.

Table 2: Results on ERGeoBench. **Bold** indicates the better result within each paired comparison. FP, SA, and CS denote Foundational Perception, Spatial Awareness, and Common Sense, respectively, while QA All denotes the overall question-answering accuracy.

	Question Answering				Embodied Geolocation			
Model	FP	SA	CS	QA All	Street	City	Country	GLS
Closed-Source Models								
Qwen-VL Max	89.28	54.87	57.65	81.89	3.03	44.44	91.92	60.95
Qwen-VL Max **+ JANUS**	**90.60**	**56.48**	**60.03**	**83.35**	**5.86**	**46.77**	**93.70**	**62.00**
Improvement	+1.32	+1.61	+2.38	+1.46	+2.83	+2.33	+1.78	+1.05
GPT-4o	88.66	55.94	59.89	75.41	7.07	52.53	94.95	66.11
GPT-4o **+ JANUS**	**90.92**	**58.36**	**60.41**	**77.54**	**7.52**	**55.89**	**98.87**	**68.23**
Improvement	+2.26	+2.42	+0.52	+2.13	+0.45	+3.36	+3.92	+2.12
Open-Source Models								
Qwen2.5-VL-7B	74.44	40.32	51.31	61.35	1.01	52.53	93.94	62.07
Qwen2.5-VL-7B **+ JANUS**	**75.14**	**41.99**	**52.23**	**62.37**	**2.02**	**52.96**	**95.61**	**65.26**
Improvement	+0.70	+1.67	+0.92	+1.02	+1.01	+0.43	+1.67	+3.19
Qwen3-VL-8B	80.12	44.90	58.92	66.88	1.01	36.36	94.95	47.61
Qwen3-VL-8B **+ JANUS**	**82.73**	**45.55**	**61.24**	**68.84**	**3.97**	**39.31**	**97.51**	**48.85**
Improvement	+2.61	+0.65	+2.32	+1.96	+2.96	+2.95	+2.56	+1.24
InternVL2.5-8B	83.50	19.72	50.76	60.12	1.01	27.27	79.80	38.20
InternVL2.5-8B **+ JANUS**	**85.42**	**21.34**	**51.50**	**61.82**	**1.77**	**28.77**	**83.84**	**39.88**
Improvement	+1.92	+1.62	+0.74	+1.70	+0.76	+1.50	+4.04	+1.68
InternVL3-8B	82.63	49.71	51.59	69.31	0.00	25.25	92.93	46.04
InternVL3-8B **+ JANUS**	**84.07**	**52.40**	**54.32**	**71.27**	**1.23**	**28.28**	**96.16**	**47.17**
Improvement	+1.44	+2.69	+2.73	+1.96	+1.23	+3.03	+3.23	+1.13

Table 3: **SFT backbone component ablation on H* Bench.** All rows use our frozen Qwen3-VL-8B SFT. HSB denotes the memory module and BEI the imagination module; the evaluated switches control real-memory intervention and future-action scoring. All four configurations are measured; bold marks the highest success rate in each task.

			Success (%) ↑	
Configuration	HSB	BEI	HOS	HPS
SFT actor	–	–	55.00	29.38
SFT + HSB	✓	–	57.50	31.88
SFT + BEI	–	✓	55.83	32.50
SFT + JANUS (HSB + BEI)	✓	✓	**61.25**	**33.13**

5.3 Ablation Study

As shown in Table 3, both HSB and BEI individually improve the SFT actor. HSB increases HOS/HPS from 55.00%/29.38% to 57.50%/31.88%, while BEI reaches 55.83%/32.50%. Combining both modules yields the best performance, achieving 61.25% on HOS and 33.13% on HPS. These results demonstrate that persistent spatial belief and imagination-guided view selection provide complementary benefits for active perception. In particular, HSB contributes more strongly to object search, whereas BEI provides a larger individual gain on path search, reflecting their different roles in evidence organization and future-view selection.

5.4 Qualitative Analysis

Figure 3 presents qualitative examples of JANUS across six representative tasks. The results show that JANUS adapts its evidence-acquisition strategy to different information needs: it zooms in for fine-grained perception, inspects relative spatial relationships, recalls previously observed views for commonsense reasoning, gathers additional scene cues for geolocation, and selects informative



Figure 3: **Qualitative results of JANUS across six tasks.** We illustrate fundamental perception, spatial awareness, commonsense reasoning, geolocation, humanoid object search, and humanoid path search. Each panel shows the task instruction, visual observations, intermediate observation or recall actions, and the final answer. The examples demonstrate how JANUS adaptively acquires and recalls task-relevant evidence through zooming, spatial inspection, cross-view recall, and informative view selection.

viewing directions for object and path search. These cases highlight the complementary roles of HSB and BEI, where HSB preserves and retrieves task-relevant spatial evidence, while BEI guides the acquisition of additional observations needed to resolve the current evidence gap.

6 Conclusion

We present **JANUS**, a spatial-belief formulation for active perception that separates acquired evidence from revisable spatial interpretations and couples these interpretations to latent evidence imagination. The current controllers approximate this design with source-linked memory, heuristic confidence, and imagined visual queries for real-evidence recall. The framework keeps predicted features separate from observed facts and leaves the multimodal actor frozen. Completed ERGeoBench tests show small overall QA gains across three backbones using a shared router, with significant aggregate gains for two of them; the H* development audit shows improvements with two Qwen actors and GPT-4o. Geolocation runs provide system measurements but await paired baselines. The controlled imagination ablation reveals a current limitation: transferring the H* world model to ERGeoBench can reduce downstream accuracy despite strong latent-prediction quality. Future work will implement and validate calibrated object-attribute beliefs, contradiction-aware relation revision, and candidate-wise recall–reranking, alongside task-utility calibration on disjoint development scenes, structural and cross-task ablations, and physically navigable environments.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of this manuscript, including correcting grammar, refining wording, and improving clarity and readability. Their use was limited to language editing and did not extend to research design, code implementation, data generation or processing, mathematical proofs, or interpretation of results. The authors take full responsibility for the accuracy, originality, and compliance of the final manuscript.

REFERENCES

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, 2020a. URL https://arxiv.org/abs/2007.00643.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural SLAM. In *International Conference on Learning Representations*, 2020b. URL https://arxiv.org/abs/2004.05155.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological SLAM for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020c. doi: 10.1109/CVPR42600.2020.01289.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01604.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.00063.

Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.15648.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso M. de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *IEEE International Conference on Robotics and Automation*, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.01603.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation*, 2023. doi: 10.1109/ICRA48891.2023.10160969.

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*, 2021. doi: 10.1109/ICCV48922.2021.01447.

Yiqing Liang, Boyuan Chen, and Shuran Song. SSCNav: Confidence-aware semantic scene completion for visual semantic navigation. In *IEEE International Conference on Robotics and Automation*, 2021. doi: 10.1109/ICRA48506.2021.9560925.

Joel Loo, Zhanxin Wu, and David Hsu. Open scene graphs for open world object-goal navigation. *arXiv preprint arXiv:2407.02473*, 2024. URL https://arxiv.org/abs/2407.02473.

Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1702.08360.

Junyeong Park, Junmo Cho, and Sungjin Ahn. Mrsteve: Instruction-following agents in minecraft with what-where-when memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2af7168a1f19e0ae61134f89eb238e57-Abstract-Conference.html.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.01564.

Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *European Conference on Computer Vision*, 2020. doi: 10.1007/978-3-030-58558-7_24.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01832.

Koya Sakamoto, Taiki Miyanishi[1], Daichi Azuma[1], Shuhei Kurita, Shu Morikuni[1], Naoya Chiba, Motoaki Kawanabe, and Yusuke Iwasawa[1]. for viewpoint-dependent active perception in 3d gaussian splatting. In *Computer Vision–ECCV 2026: 19th European Conference, Malmö, Sweden, September 8–12, 2026, Proceedings, Part XXXV*, pp. 60. Springer Nature, 2026.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.00900.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.01609.

Teng Wang, Xinxin Zhao, Wenzhe Cai, and Changyin Sun. Imaginenav++: Prompting vision-language models as embodied navigator through scene imagination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2022. URL https://arxiv.org/abs/2206.14176.

Kaiwen Xue, Tao Wei, Guoxin Zhang, Zhonghong Ou, Kaoyan Lu, Yu Feng, Yifan Zhu, and Haoran Luo. Ergeobench: A comprehensive benchmark for embodied reasoning and geo-localization in multimodal large language models. *arXiv preprint arXiv:2605.31251*, 2026.

Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3D-Mem: 3D scene memory for embodied exploration and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2411.17735.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, Saining Xie, and Yiming Li. Thinking in 360°: Humanoid visual search in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2511.20351.

Albert J. Zhai and Shenlong Wang. PEANUT: Predicting and navigating to unseen targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.01003.

Mingliang Zhai, Zhi Gao, Yuwei Wu, and Yunde Jia. Memory-centric embodied question answering. *arXiv preprint arXiv:2505.13948*, 2025.

Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01553.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb261df4322a8bd0a73093c4d8a0d02d-Abstract-Conference.html.

Minzhao Zhu, Binglei Zhao, and Tao Kong. Navigating to objects in unseen environments by distance prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022. doi: 10.1109/IROS47612.2022.9981766.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, et al. Active-o3: Empowering mllms with active perception via pure reinforcement learning. *arXiv preprint arXiv:2505.21457*, 2025.

A TASK OVERVIEW

We study six tasks spanning visual question answering, geolocation, and directional search. ERGeoBench includes Foundational Perception (FP), Spatial Awareness (SA), Common Sense (CS), and Embodied Geolocation (Geo). H* Bench includes Humanoid Object Search (HOS) and Humanoid Path Search (HPS).

Across these tasks, the agent remains at a fixed position and acquires limited-field-of-view observations by controlling its camera pose $p_t = (\theta_t, \phi_t, z_t)$, where θ_t, ϕ_t, and z_t denote yaw, pitch, and zoom, respectively. Zoom is enabled only when supported by the evaluation protocol. The real-world panorama is available only to the renderer and is never directly exposed to the agent.

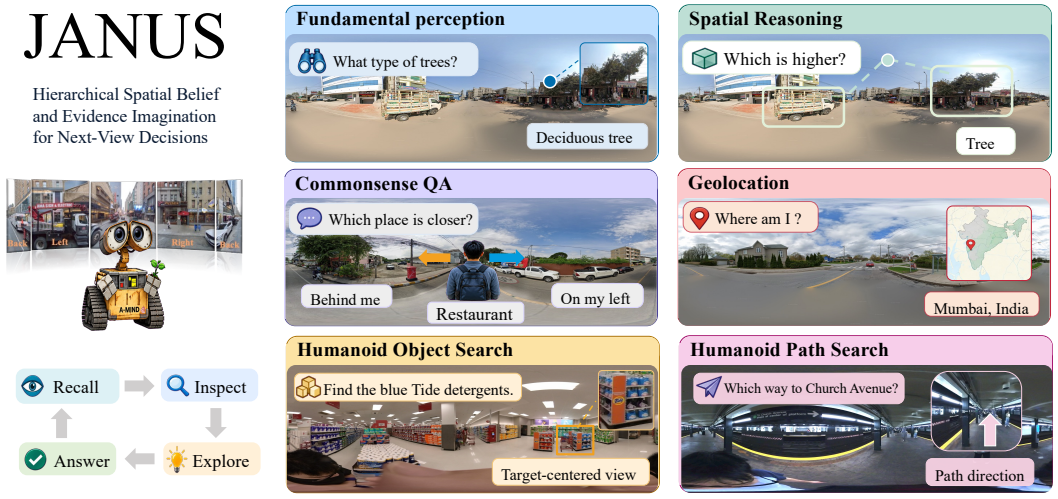

Figure 4: **Overview of the six active-perception tasks considered in JANUS.** ERGeoBench covers Foundational Perception, Spatial Awareness, Common Sense, and Embodied Geolocation, requiring agents to recognize scene attributes, reason about spatial relations, combine visual evidence with contextual knowledge, and infer geographic location, respectively. H* Bench evaluates Humanoid Object Search and Humanoid Path Search, requiring a target-centered yaw–pitch direction or the yaw direction of a feasible route. The left panel summarizes the recall–inspect–explore–answer loop. Panoramic backgrounds provide context for visualization only: the agent receives egocentric observations and controls yaw, pitch, and, where supported, zoom from a fixed position. Path search requires direction prediction rather than physical traversal.

Foundational Perception (FP). This task evaluates recognition of directly observable scene attributes, including architecture, infrastructure, signage, terrain, and vegetation. The agent inspects relevant visual regions and produces an answer in the benchmark's original format. The example in Figure 4 illustrates identifying a tree type from visible evidence.

Spatial Awareness (SA). This task evaluates reasoning about positions, directions, relative heights, depths, distances, and spatial relationships, including relationships that require observations from different views. Predictions retain the reference frame specified by the original question. The illustrated example compares the heights of two visible entities.

Common Sense (CS). This task requires combining observed evidence with contextual and functional knowledge to answer questions about the scene. The benchmark includes single-view, dual-view, and cross-view questions. Relevant evidence may be distributed across several directions, requiring the agent to recall earlier observations and inspect additional regions before answering.

Embodied Geolocation (Geo). The agent integrates geographic cues across acquired views to predict coordinates and location labels. Such cues may include readable signs, architectural styles,

road infrastructure, vegetation, and their surrounding context. Predictions are evaluated using geographic-label accuracy, distance-based metrics, and the benchmark geolocation score.

Humanoid Object Search (HOS). Given a target description, the agent searches the surrounding scene and submits a yaw–pitch direction $(\hat{\theta}, \hat{\phi})$ that centers the requested object. Success is determined by the official object-search scorer. The illustrated example requires locating the specified detergent among visually similar products.

Humanoid Path Search (HPS). Given a requested destination, the agent identifies a feasible route and submits its yaw direction $\hat{\theta}_{\mathrm{path}}$. The task may require inspecting directional signs, passageways, and barriers. Predictions are evaluated by the official path-search scorer; the agent does not physically traverse the predicted route.

Observation and answer conventions. The memory-first protocol provides task-blind observations before revealing the instruction, whereas the online protocol reveals the instruction from the outset without an additional pre-scan. For FP, SA, and CS, we preserve the original questions, answer formats, and scorers. Benchmark anchor views remain available when required to preserve image-relative question semantics. Supporting real-view identifiers and camera poses are recorded separately from the submitted answer, allowing grounding to be examined independently of answer correctness.

B Related Work

B.1 Memory for Embodied Agents

Learned and map-based memory. Embodied agents use memory to retain observations that are no longer visible and make them available for subsequent decisions. Neural Map writes latent features into a spatially addressable memory (Parisotto & Salakhutdinov, 2018), while Cognitive Mapping and Planning couples an explicit spatial representation with differentiable planning (Gupta et al., 2017). Scene Memory Transformer attends over stored observations (Fang et al., 2019), and Episodic Transformer models multimodal experience for embodied instruction following (Pashevich et al., 2021). Explicit mapping offers complementary representations: Active Neural SLAM combines mapping with exploration (Chaplot et al., 2020b), Goal-Oriented Semantic Exploration uses semantic information to guide object search (Chaplot et al., 2020a), and Neural Topological SLAM represents the environment through places and their connections (Chaplot et al., 2020c). These approaches motivate retaining spatially organized experience for sequential decisions. JANUS focuses on how acquired views support revisable scene interpretations and task-conditioned evidence retrieval.

Open-vocabulary and hierarchical memory. Language-aligned spatial representations allow agents to query their surroundings using semantic descriptions. VLMaps associates visual-language features with spatial map locations (Huang et al., 2023), while ConceptGraphs constructs object-centric 3D scene graphs for perception and planning (Gu et al., 2024). Open Scene Graphs further studies relational representations for open-world object-goal navigation (Loo et al., 2024). Hierarchical representations organize information at different decision scales: DUET combines fine-grained local reasoning with a coarse global graph (Chen et al., 2022). MrSteve retrieves what–where–when experiences for instruction following (Park et al., 2025), and ImagineNav++ combines hierarchical keyframe memory with prospective navigation (Wang et al., 2026). JANUS builds on these directions by connecting hierarchical organization, associative recall, and task-specific evidence requirements. Stored images retain their provenance, while their interpreted attributes can be revised after further observation. The current implementation uses heuristic confidence rather than a calibrated belief posterior.

B.2 Imagination for Embodied Agents

Predictive maps and unseen-target inference. Predicting information beyond the current observation can guide an agent toward useful regions. PONI predicts area and object potential functions to

select exploration goals (Ramakrishnan et al., 2022). Distance Prediction estimates distances to unseen objects (Zhu et al., 2022), while PEANUT predicts likely target locations to guide navigation (Zhai & Wang, 2023). Other methods anticipate the environment itself: Occupancy Anticipation predicts free and occupied space beyond the visible region (Ramakrishnan et al., 2020); L2M learns task-relevant semantic maps (Georgakis et al., 2022); SSCNav combines semantic scene completion with navigation (Liang et al., 2021); and Imagine Before Go generates local map predictions for object-goal navigation (Zhang et al., 2024). These methods motivate evaluating the potential value of an unobserved region before acquiring it. JANUS applies this principle to candidate camera views, using predicted visual features to guide evidence acquisition and recall.

World models and future-view imagination. Latent world models provide a mechanism for predicting future observations and evaluating actions. PlaNet learns latent dynamics for planning (Hafner et al., 2019), Dreamer learns behavior through latent imagination (Hafner et al., 2020), and DayDreamer extends world-model learning to physical robots (Wu et al., 2022). For visual navigation, PathDreamer predicts panoramic observations along candidate paths (Koh et al., 2021). Pose-guided view synthesis (Tseng et al., 2023) and ZeroNVS (Sargent et al., 2024) further explore generating observations from novel camera poses. ImagineNav uses imagined future views as options for vision-language navigation (Zhao et al., 2025), and ImagineNav++ integrates prospective-view reasoning with hierarchical visual memory (Wang et al., 2026).

JANUS uses a latent predictor conditioned on acquired spatial memory and a candidate camera pose. Its predictions guide view ranking and retrieval of real observations. The predictor does not generate factual OCR strings, object identities, or answers. Evidence type and expected clarity are planning interpretations rather than explicit neural outputs in the current implementation. A subsequent real observation is required before a predicted relation or attribute can update a persistent claim.