WHEN TO PERCEIVE, REMEMBER, IMAGINE, OR ACT: AUTONOMOUS DECISION FOR EMBODIED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must decide what to do from incomplete and potentially conflicting visual evidence. Existing approaches offer complementary mechanisms for visual reasoning, memory retrieval, and predicting action outcomes. Some recent methods also adaptively choose when to gather imagined evidence, but coordinating perception, memory, imagination, and execution remains a broader decision problem. Without such coordination, an agent may overlook relevant details, ignore past experience, or rely on imagined outcomes that are inconsistent with reality. To address this problem, we propose **AutoPRIA**, a framework that *Autonomously* organizes embodied decision-making around four fundamental operations: *Perceive* finer details of the present, *Remember* the past, *Imagine* the future outcome of a candidate action, and *Act* to obtain a real observation from the environment. To the best of our knowledge, AutoPRIA is the first framework to bring all four operations together in one embodied agent, so that the agent decides what to know together with what to do. Under this framework, we build an agent with tool-augmented perception, dual memory, action-conditioned imagination with a fixed world model, and real actions whose observations refresh the information state and check earlier predictions. The policy is trained with supervised initialization followed by reinforcement refinement. Across 4 embodied tasks (active recognition, image-goal navigation, active embodied question answering, and robotic manipulation), AutoPRIA outperforms the Qwen3.5-9B baseline with the Wan2.2 world model on every task, with an average relative improvement of 17.7%, while executing fewer actions on AR, IGNav, and RM.

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.

1 INTRODUCTION

Recent advances in multimodal foundation models and generalist robot policies have accelerated the development of embodied agents that connect visual understanding and language with physical action (Driess et al., 2023; Bai et al., 2025; Physical Intelligence et al., 2025; Yang et al., 2025a). Such agents promise to support navigation, object recognition, and manipulation in homes and other everyday environments. Turning these capabilities into useful autonomy requires reliable decision-making under partial observability: an agent must act on incomplete evidence, and its actions determine both task progress and what it can observe next. Before committing to an action or an answer, the agent must therefore also decide what evidence to gather, making the choice of how to acquire information integral to the choice of what to do, as shown in Figure 1.

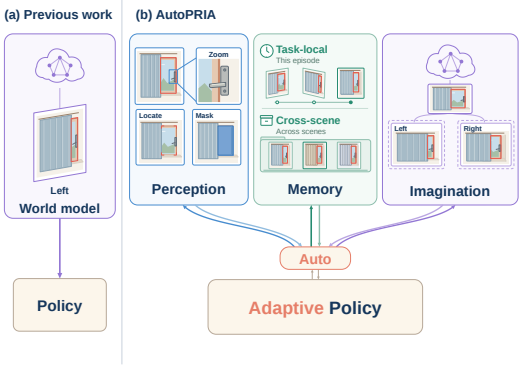

Figure 1: Representative prior work (left) and the proposed AutoPRIA (right), illustrated with schematic AR views.

Existing systems combine different subsets of these capabilities. For example, GR00T N1 couples a vision-language backbone with a diffusion action module to generate action chunks conditioned on visual observations, language instructions, and robot state (NVIDIA et al., 2025b). Visual world models provide an additional route: anticipating what an agent might observe after acting (Lu et al., 2025; Yang et al., 2025b). World-in-World already incorporates this route into a closed loop, in which a policy proposes candidate plans, a world model predicts their outcomes, and a revision policy selects a plan for execution (Zhang et al., 2025). However, access to reasoning tools and predictive rollouts does not by itself ensure a correct decision. Recent approaches learn when to invoke imagination (Yu et al., 2026; Zhu et al., 2026), while coordinating it with active perception, memory retrieval, and task execution remains a broader design challenge. Without such coordination, an agent may overlook relevant details, fail to connect current observations with past experience, or rely on imagined outcomes that are inconsistent with reality.

To address this problem, we propose **AutoPRIA** (Autonomous Perceive, Remember, Imagine, and Act), a unified framework that autonomously organizes embodied decision-making around four fundamental operations spanning the present, the past, the anticipated future, and the environment itself. *Perception* extracts task-relevant details from the present scene. *Memory* retrieves observations and experience from the past or other scenes. *Imagination* generates possible future observations conditioned on candidate actions. *Action* executes the selected decision in the environment, yielding new observations that inform subsequent reasoning. These operations expand the information available for a decision, and their results accumulate in a single evolving information state. The agent autonomously decides at every step which operation to invoke, including when to stop gathering evidence and commit to an action or an answer. To the best of our knowledge, AutoPRIA is the first framework for embodied agents that unifies evidence acquisition, whether from the present, the past, or the anticipated future, with task execution in a single decision process, so that the agent decides what to know together with what to do.

Under the proposed framework, we build an embodied agent in which all four operations are driven by one policy (Figure 2). Perception is realized by visual and spatial tools such as zooming, localization, and additional camera views. Memory combines a task-local store of the current episode with a cross-scene bank of experience retrieved by visual similarity. Imagination queries a fixed action-conditioned world model. Action executes in the environment and returns a real observation that refreshes the information state and reveals whether the imagined outcome matched reality. The policy is trained with supervised initialization followed by reinforcement refinement using task, action, and evidence-use rewards (Shao et al., 2024; Guo et al., 2025). We evaluate AutoPRIA on four embodied benchmarks covering Active Recognition (AR), Image-Goal Navigation (IGNav), Active Embodied Question Answering (A-EQA), and robotic manipulation (RM). Experiments show that AutoPRIA outperforms the Qwen3.5-9B baseline with the Wan2.2 world model on all of them, resulting in an average relative improvement of 17.7% on the primary metric of each task. We further conduct ablation studies and behavior analyses to demonstrate the necessity of each component and to understand how the agent makes decisions. To summarize, our contributions are as follows.

- We formulate evidence acquisition and task execution as a unified decision problem for embodied agents, covering perception, memory, imagination, and action.
- To address this problem, we propose **AutoPRIA**, a unified framework that autonomously organizes embodied decision-making around four fundamental operations spanning the present, the past, the anticipated future, and the environment itself.
- Under this framework, we build an embodied agent that realizes the four operations with tool-augmented perception, dual memory, action-conditioned imagination, and real actions against reality, all driven by one policy trained with supervision followed by reinforcement refinement.
- Extensive experiments on four embodied benchmarks demonstrate the superiority of AutoPRIA, which outperforms the Qwen3.5-9B baseline with the Wan2.2 world model on all four tasks while executing fewer actions on AR, IGNav, and RM.

2 Related Work

Embodied Agents. Embodied agents connect semantic goals with perception and action under partial observability. SayCan grounds language-based plans in the feasibility of robotic skills (Ahn et al., 2022), PaLM-E incorporates continuous sensory observations into a multimodal language

model (Driess et al., 2023), and vision-language-action models such as RT-2 and $\pi_{0.5}$ connect pre-trained semantic knowledge with robot control (Brohan et al., 2023; Physical Intelligence et al., 2025). EmbodiedBench evaluates such multimodal language models as agents on both high-level and low-level embodied tasks (Yang et al., 2025a). These approaches address planning, control, and evaluation, while other work develops explicit mechanisms for gathering decision evidence. For the present, active perception acquires more informative observations, as in Embodied Amodal Recognition, which learns to move to better recognize heavily occluded objects (Yang et al., 2019), and ACTIVE-o3, which learns where to look with reinforcement learning (Zhu et al., 2025). Tool-augmented reasoning interleaves reasoning with external actions (Yao et al., 2023), and DeepEyes, OpenThinkIMG, and Mini-o3 teach multimodal models to crop, zoom, or call visual tools over multiple turns (Zheng et al., 2025; Su et al., 2025; Lai et al., 2026), mostly with the group-relative policy optimization recipe popularized by DeepSeek-R1 (Shao et al., 2024; Guo et al., 2025). For the past, KARMA maintains long-term and short-term memory (Wang et al., 2025b), and 3D-Mem builds a snapshot-based scene memory for exploration and question answering (Yang et al., 2025c). World models, reviewed below, add predictions of the anticipated future. These lines of work emphasize different information sources and mechanisms for acquiring them. AutoPRIA instead unifies perceiving the present, remembering the past, imagining the future, and acting in the environment within one agent process over an evolving information state, so that the policy decides what to know together with what to do.

World Models. World models predict the consequences of actions and provide a basis for planning or policy learning. Dreamer learns behavior through trajectories imagined in a latent dynamics model (Hafner et al., 2019). Visual world models make predicted outcomes accessible as images or videos: Pathdreamer synthesizes observations at unvisited viewpoints for indoor navigation (Koh et al., 2021), and Navigation World Models generate action-conditioned visual trajectories that support planning and candidate ranking (Bar et al., 2025). UniPi expresses planning as text-guided video generation and extracts control actions from the resulting visual plans (Du et al., 2023). Large video generators such as Stable Video Diffusion, HunyuanVideo, Wan, and Cosmos are increasingly adopted as general-purpose world simulators (Blattmann et al., 2023; Kong et al., 2024; Wan et al., 2025; NVIDIA et al., 2025a). World-in-World unifies diverse visual generators through an action interface and a proposal–simulation–revision loop, evaluating their utility across four embodied tasks (Zhang et al., 2025). Its findings establish that visual generation quality alone does not determine closed-loop task performance. Recent work also lets vision-language agents consult world models at test time. GenEx imagines exploration to update the agent's beliefs (Lu et al., 2025), MindJourney renders the views along camera paths planned by a vision-language model (Yang et al., 2025b), and VAGEN trains multi-turn agents to reason explicitly about states and transitions (Wang et al., 2025a). Closest to our setting, recent studies decide whether and how much to imagine (Yu et al., 2026) or train a policy with reinforcement learning to decide when to query a world simulator (Zhu et al., 2026). Building on this predictive planning capability, AutoPRIA treats imagined futures as one source of decision evidence alongside active perception and dual memory. The policy learns to use these sources while the world model remains fixed; subsequent real observations can challenge predictions for the actions actually executed.

3 METHOD

An embodied agent needs to decide which evidence to consult and which action or answer that evidence supports. AutoPRIA expresses these choices through four operations: perception, memory, imagination, and action (Figure 2). It builds on World-in-World's closed-loop planning interface (Zhang et al., 2025) and DeepEyes' tool-interleaved visual reasoning (Zheng et al., 2025). The central connection is an evolving information state: each operation supplies evidence for the next decision, or commits a task output.

3.1 AUTONOMOUS DECISION-MAKING IN AN AUGMENTED STATE SPACE

Let g denote the task goal, o_t the current observation bundle, and h_t the history of environmental observations, actions, and judgments. Environmental step t counts executed action primitives; decision step n counts completed operations, including evidence queries. The policy observes the

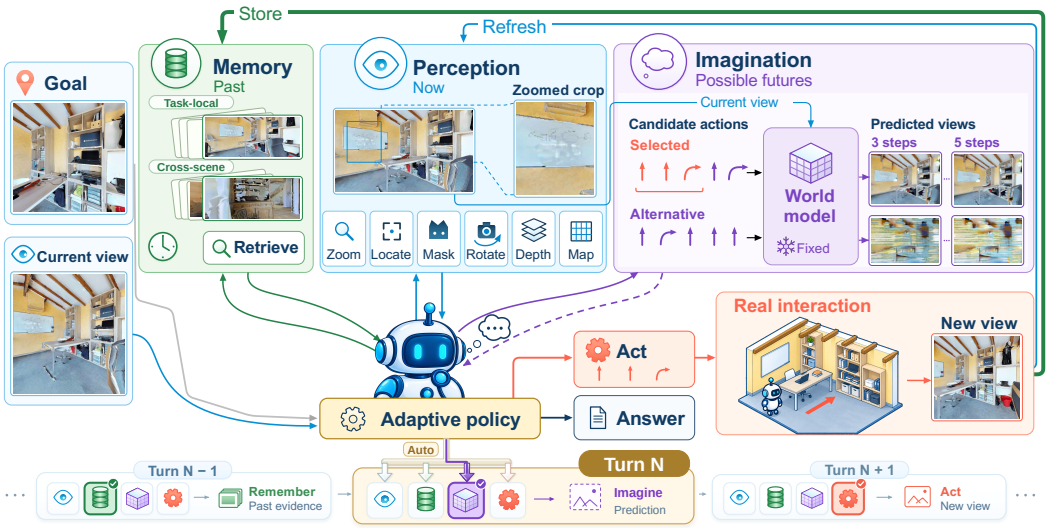

Figure 2: **Adaptive evidence selection for embodied decisions.** The policy can retrieve memory, inspect current evidence, imagine action-conditioned futures, execute an environmental action, or provide a terminal answer.

augmented state

$$s_n = (g, o_t, c_n), \qquad t = t(n), \tag{1}$$

where c_n is a compact multimodal context assembled from h_t and returned evidence. It retains relevant tool results, memories, and predictions with their sources. This state represents the information available to the agent under partial observability, including prior judgments that may require revision. At each nonterminal state, the policy selects an operation and its arguments:

$$(u_n, \alpha_n) \sim \pi_\theta(\cdot \mid s_n), \quad u_n \in \mathcal{U}, \quad \mathcal{U} = \{\textsc{Perceive}, \textsc{Remember}, \textsc{Imagine}, \textsc{Act}\}. \tag{2}$$

The arguments specify a tool and its parameters, a memory query, candidate action sequences, or a task output. ACT encompasses environmental execution and terminal answer/stop decisions. Thus, committing an answer and executing a movement are two forms of the same operation category, with different transitions.

Let r_n be the returned operation record, containing the request, result, and source identity; for execution, it also records the actions actually performed. Write $c_n^+ = \mathrm{Update}(c_n, r_n)$ for incorporating this record into the context, and $\mathcal{U}_{\mathrm{info}} = \mathcal{U} \setminus \{\textsc{Act}\}$. The decision process has three transitions:

$$s_{n+1} = \begin{cases} (g, o_t, c_n^+), & u_n \in \mathcal{U}_{\mathrm{info}}, \\ (g, o_{t+\ell}, c_n^+), & u_n = \textsc{Act},\ \text{execution}, \\ s_{\mathrm{term}}(y_n), & u_n = \textsc{Act},\ \text{answer/stop}. \end{cases} \tag{3}$$

Here ℓ is the executed prefix length and y_n is the task output accepted by the termination rule. Execution updates $h_{t+\ell}$ and task-local memory with the new observation before the next decision; a terminal output ends the episode. The environmental index stays fixed for an information operation and advances by ℓ for execution. Consequently, the same physical viewpoint can support several distinct decision states as evidence accumulates. A crop, for instance, can be followed by another inspection when the target remains ambiguous or by immediate action when the evidence already suffices, because the formulation imposes no fixed ordering or mandatory traversal of components.

3.2 A Unified Framework for Autonomous Decision-Making

Tool-Augmented Active Perception. A current view may contain the relevant object while leaving its identity or spatial relation ambiguous. Perception makes such evidence accessible through a selected tool. For $\alpha_n = (k_n, \eta_n)$, with tool identity k_n and arguments η_n, the executor returns

$$r_n = \mathcal{T}_{k_n}(o_t; \eta_n). \tag{4}$$

The return is appended in the existing multimodal format, with its tool and source view identified. The next response can therefore reason about the requested detail using the actual result. Unavailable queries return an explicit status rather than fabricated evidence.

Zoom, Locate, and Mask expose local appearance, relevant regions, and object extent, helping the policy examine competing interpretations of a target. Rotate, Depth, and Map provide additional directional or spatial evidence. Rotate obtains a new camera yaw view, and Depth accesses measured depth. Map reads recorded poses and actions from h_t to relate the current view to prior exploration. Returned views retain their identities so a side view, an earlier crop, and the current front observation can be distinguished in reasoning. Appendix A specifies argument validation and sensing contracts.

Dual Memory. The current view alone does not convey which regions have been examined or how the target appeared from earlier viewpoints. Both memories use a fixed CLIP encoder f (Radford et al., 2021), and for a view v in which the target box is visible, $\tilde{v}$ denotes the crop around that box. Task-local memory M_n^{ep} stores the real views of the current episode in which the target box is visible, together with the embeddings $f(\tilde{v})$ of their crops, and resets between episodes. New experience updates the store automatically, and REMEMBER reads at most two earlier views into c_n, namely the view with the largest target box and the remaining view least similar in crop embedding to both the current view and that first view. Earlier answers are not carried over, so the policy reassesses the earlier evidence instead of treating a previous judgment as ground truth. A new view can therefore support revising a judgment that previously appeared plausible.

Cross-scene memory supplies complementary appearance-based experience from a fixed bank $\mathcal{B}$ of annotated views from training scenes. Each entry $b \in \mathcal{B}$ stores a view v_b, the crop $\tilde{v}_b$ around its annotated target, and the target's category. Retrieval scores every entry against the current view v_n and its target crop $\tilde{v}_n$ by

$$s_n(b) = 0.8 \cos\big(f(\tilde{v}_n), f(\tilde{v}_b)\big) + 0.2 \cos\big(f(v_n), f(v_b)\big), \tag{5}$$

where cos denotes cosine similarity, so the target appearance dominates and the surrounding view adds context. It returns, in decreasing order of $s_n(b)$, at most $K = 2$ entries with $s_n(b) \geq \tau$, keeping at most one entry per training scene, where the threshold $\tau = 0.897$ is calibrated on non-benchmark scenes (Appendix B). Retrieval runs only when the target box is visible in the current view, and it returns no entry when none reaches τ. Each retrieved entry enters the context as a thumbnail of its view with the target box marked, together with its category, and is labeled as an example from a different scene rather than an answer, which suggests candidate categories and visually similar alternatives to examine against the current evidence. In recognition, such look-alike examples expose alternative interpretations of the current target. If the distinction depends on a small visual detail, the next operation may request a crop; memory then informs what to inspect as well as what answer the evidence supports. Task-local memory describes the current trajectory, while the bank offers analogous appearances from other scenes. Evaluation scenes are excluded from the bank, and stored labels refer only to historical examples. Retrieval provides neither the current target label nor unobserved geometry of the evaluated scene.

Action-Conditioned Imagination. When a decision depends on what an action may reveal, imagination supplies a conditional prediction. Following World-in-World, the policy proposes candidate action sequences $\mathbf{A}_t^{(m)} = (a_t^{(m)}, \ldots, a_{t+H-1}^{(m)})$ of horizon H as the arguments of IMAGINE. For each candidate, a fixed world model produces

$$\hat{o}_{t+1:t+H}^{(m)} = W_\phi\Big(o_t, \mathcal{C}(\mathbf{A}_t^{(m)})\Big), \tag{6}$$

where $\mathcal{C}$ maps the action sequence to the model's conditioning interface. Each rollout retains its originating view and conditioning actions, allowing the policy to compare task-relevant consequences such as target visibility. Observed and remembered evidence provide context for interpreting these hypotheses before selecting the next operation. Imagination can therefore be followed by further inspection or retrieval as well as execution. Memory lookup and future generation remain separate mechanisms: memories augment the policy context, while W_ϕ predicts from the current observation and action controls.

Action and Feedback. Execution supplies the real observation that determines the next state in Equation 3 and enters task-local memory. When an aligned prediction is available, the policy can

inspect its agreement with reality over the executed prefix $\ell \leq H$ and matching camera view. For instance, a real view that fails to reveal a predicted target provides evidence to reconsider that expectation. Only the corresponding executed prediction receives this check; rejected candidates and unexecuted continuations remain hypothetical. The feedback changes the evidence available for subsequent decisions while the world model remains fixed, without a separate calibration loss. Terminal ACT submits an answer or stop, corresponding to the separate Answer exit in Figure 2.

3.3 EVIDENCE-GROUNDED TWO-STAGE POLICY LEARNING

Stage I: supervised initialization. SFT establishes valid evidence requests and decisions conditioned on their returns. Training examples serialize the task, visual context, and available history, together with demonstrated calls, executed returns, and subsequent actions or answers where available. Call supervision teaches the tool and arguments from the preceding context. Decision supervision then teaches what to conclude after the requested evidence is supplied. Direct-action and direct-answer examples retain the immediate ACT path. These existing response formats provide supervision without additional four-class routing labels.

Memory-conditioned examples use the history representation encountered at inference. Prior judgments may support or conflict with the current target, which remains the supervised answer. Recognition examples also preserve the online image ordering and category-output interface. For example z, let $\mathcal{I}(z)$ contain assistant target-token positions, $\xi_{z,j}$ their preceding multimodal contexts, and $w_{z,j}$ their weights. On a minibatch $\mathcal{D}_b$, with token index j distinct from decision index n, we minimize

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\frac{\sum_{z\in\mathcal{D}_b}\sum_{j\in\mathcal{I}(z)} w_{z,j}\log \pi_\theta(z_j \mid \xi_{z,j})}{\sum_{z\in\mathcal{D}_b}\sum_{j\in\mathcal{I}(z)} w_{z,j}}. \tag{7}$$

Prompt, padding, and tool-return tokens are masked. Executed images and measurements remain inputs to the decision, not targets to be generated.

Stage II: GRPO refinement. GRPO (Shao et al., 2024; Guo et al., 2025) adjusts the supervised policy using task-related scores of sampled responses. Using the DeepEyes codebase (Zheng et al., 2025) built on veRL (Sheng et al., 2025), $\pi_{\theta_{\mathrm{old}}}$ samples G responses or tool-interleaved continuations for the same decision context x. Permitted calls receive backend results or verified artifacts matched to the request. For valid continuation i, the task-specific reward is

$$R_i = \lambda_{\mathrm{ans}} R_i^{\mathrm{ans}} + \lambda_{\mathrm{act}} R_i^{\mathrm{act}} + \lambda_{\mathrm{ev}} R_i^{\mathrm{ev}} - \lambda_{\mathrm{pen}} P_i. \tag{8}$$

Answer credit measures agreement with the training target; action credit measures reference-sequence and prefix agreement. Evidence credit checks valid calls and references to their returns, providing a proxy for evidence use. P_i collects configured penalties for unsupported, omitted, or unnecessary calls; flat call penalties can be disabled. Malformed outputs receive the prescribed invalid-output reward. These state-level signals remain distinct from complete episode outcomes.

The relative advantage is $A_i = (R_i - \bar{R})/(s_R + \varepsilon)$, with group mean $\bar{R}$, standard deviation s_R, and stabilizer ε. It guides the clipped policy update on assistant-generated tokens; tool-return tokens are excluded. Both training stages update the configured policy parameters while keeping the world model and retrieval encoder fixed. Appendix B lists the reward weights, trainable parameters, and other training settings.

4 EXPERIMENTS

We evaluate on four embodied tasks from World-in-World (Zhang et al., 2025). In Active Recognition (AR), the agent must identify the category of a target object marked by a bounding box in its initial view, and it may execute up to ten actions to obtain more informative viewpoints before answering. The 551 AR episodes span 29 Matterport3D scenes (Chang et al., 2017). In Image-Goal Navigation (IGNav), the agent navigates through HM3D scenes (Ramakrishnan et al., 2021) to the location where a goal image was captured, over 144 episodes. Active Embodied Question Answering (A-EQA) contains 184 open-ended questions about HM3D scenes from OpenEQA (Majumdar et al., 2024), which the agent answers after exploring. An LLM judge compares each answer

Table 1: Active Recognition (AR) and Image-Goal Navigation (IGNav) performance across various models. Higher success rate (**SR**, %) is better. **Mean Traj.** reports mean trajectory length.

Model Details						AR		IGNav	
Policy Backbone	Model Type	Method	Control Type	Input Type	#Param.	SR ↑	Mean Traj. ↓	SR ↑	Mean Traj. ↓
Qwen2.5-VL 72B	Base Policy	VLM (w/o WM)	–	RGB	72B	50.27	6.24	35.42	47.5
	+ Image Gen.	PathDreamer	Viewpoint	RGB-D; Pano	0.69B	56.99	5.28	36.80	47.3
	+ Image Gen.	SE3DS	Viewpoint	RGB-D; Pano	1.1B	57.53	5.29	36.11	47.0
	+ Video Gen.	NWM	Trajectory	RGB	1B	57.35	5.68	40.28	47.1
	+ Video Gen. Zero-Shot	SVD	Image	RGB	1.5B	57.71	5.29	40.28	46.4
		LTX-Video	Text	RGB	2B	56.08	5.37	36.81	47.5
		Hunyuan	Text	RGB	13B	57.71	5.21	36.11	46.8
		Wan2.1	Text	RGB	14B	58.26	5.24	38.19	48.2
		Wan2.2	Text	RGB	5B	55.35	5.73	38.88	46.5
		Cosmos-P2	Text	RGB	2B	55.35	5.71	36.81	47.6
Qwen3.5 9B	Base Policy	VLM (w/o WM)	–	RGB	9B	57.53	6.985	32.64	53.86
	+ Video Gen. Zero-Shot	Hunyuan	Text	RGB	13B	56.81	5.352	38.89	53.951
		Wan2.1	Text	RGB	14B	54.45	6.024	42.36	53.319
		Wan2.2	Text	RGB	5B	54.81	5.650	31.25	54.96
	AutoPRIA + Wan2.2		Action	RGB; Pano	5B	**62.98**	**4.737**	**41.67**	**50.27**

with the reference and rates it from 1 to 5, and the answer score maps these ratings linearly to a 0–100 scale. AR, IGNav, and A-EQA are simulated in Habitat (Savva et al., 2019). Robotic manipulation uses 200 RLBench episodes (James et al., 2020) across four tasks, namely inserting onto a square peg, pushing buttons, sliding a block to a color target, and stacking cups. We report success rate (SR) for AR, IGNav, and manipulation, and the answer score for A-EQA. Mean trajectory length counts executed actions for AR and IGNav, meters traveled for A-EQA, and high-level planning steps for manipulation. We follow World-in-World's evaluation protocol and add Qwen3.5-9B (Qwen Team, 2026) evaluations alongside Qwen2.5-VL-72B (Bai et al., 2025). Table 1 evaluates these policies alone and in the listed combinations with image generators such as PathDreamer and SE3DS (Koh et al., 2021; 2023) or video generators such as HunyuanVideo, Wan2.1, Wan2.2, and Cosmos-Predict2 (Bar et al., 2025; Blattmann et al., 2023; HaCohen et al., 2024; Kong et al., 2024; Wan et al., 2025; NVIDIA, 2025). AutoPRIA uses Qwen3.5-9B as its policy and the same fixed Wan2.2 world model.

4.1 Benchmark Results

Table 1 reports AR and IGNav results. On 551 AR episodes, AutoPRIA achieves 62.98% SR versus 54.81% for the raw Qwen3.5-9B policy with Wan2.2, an increase of 8.17 percentage points. Mean trajectory length also falls from 5.650 to 4.737 executed actions. On 144 IGNav episodes, SR increases from 31.25% to 41.67%, an increase of 10.42 percentage points, and mean trajectory length falls from 54.96 to 50.27 actions. On both tasks, AutoPRIA therefore completes more episodes while executing fewer actions. On 184 A-EQA questions, Table 2 reports an answer score of 40.63 for AutoPRIA and 36.28 for the Qwen3.5-9B baseline, a difference of 4.35 points on the graded 0–100 scale. Mean trajectory length increases from 6.13 to 8.02 meters. Table 3 reports robotic manipulation on RLBench. Over 200 episodes, raw Qwen3.5-9B with Wan2.2 succeeds on 66 episodes (33.0%), with a mean trajectory of 3.39 planning steps. AutoPRIA succeeds on 73 episodes (36.5%), with a mean trajectory of 2.96 planning steps.

4.2 Ablation and Analysis

Memory ablation. After two-stage training, we ablate memory on AR (Table 4), where the dual-memory configuration is the AutoPRIA setting in Table 1. Task-local memory yields the largest single gain, raising SR from 60.98% to 62.25% and reducing the mean trajectory from 5.154 to 4.684 actions, since it supplies earlier real views of the same target and lets the policy build on evidence it

Table 2: Active Embodied Question Answering (A-EQA) performance.

Model Details		A-EQA Performance	
Model Type	Method	Ans. Score ↑	Mean Traj. ↓
Base Policy	VLM (w/o WM)	45.7	20.4
+ Image Gen.	PathDreamer	46.0	20.4
+ Image Gen.	SE3DS	45.8	20.3
+ Video Gen.	NWM	47.1	20.5
+ Video Gen.	Wan2.1	45.7	20.1
	Wan2.2	46.3	20.3
	LTX-Video	46.6	20.8
	Cosmos-P2	46.6	21.0
	Hunyuan	46.8	20.4
	SVD	46.9	20.4
Qwen3.5 9B	+ Wan2.2	36.28	**6.13**
	AutoPRIA	**40.63**	8.02

Table 3: Robotic manipulation performance on RLBench.

Model Details		Manipulation Performance	
Model Type	Method	SR ↑	Mean Traj. ↓
Base Policy	VLM (w/o WM)	44.5	2.52
+ Video Gen.	SVD	44.0	2.47
	LTX-Video	44.5	2.46
	Hunyuan	44.5	2.44
	Wan2.1	44.0	2.51
	Cosmos-P2	44.0	2.50
Qwen3.5 9B	+ Wan2.2	33.0	3.39
	AutoPRIA	**36.5**	**2.96**

Table 4: Memory ablation on AR (551 episodes).

Task-local	Cross-scene	Correct	SR (%) ↑	Mean Traj. ↓
Off	Off	336	60.98	5.154
On	Off	343	62.25	4.684
Off	On	334	60.62	5.080
On	**On**	**347**	**62.98**	**4.737**

Table 5: Training-stage without memory ablation on AR. Mean Traj. counts executed actions.

Stage	Correct	SR (%) ↑	Mean Traj. ↓
SFT	324	58.80	4.214
SFT + GRPO	**336**	**60.98**	**5.154**

has already collected. Cross-scene memory alone does not improve SR (60.62%), likely because its retrieved examples only suggest candidate categories that must be checked against current evidence, but on top of task-local memory it adds four correct recognitions and reaches the best SR of 62.98%. The two memories are therefore complementary, and every memory configuration also shortens the mean trajectory relative to the policy without memory.

Two-stage training. Table 5 separates supervised initialization from GRPO refinement. This comparison is designed to fix the task split, inference interfaces, memory configuration, world model, and stopping rules while changing the policy checkpoint. SFT supplies evidence-conditioned decision supervision; GRPO subsequently optimizes task-related rewards. SFT achieves 58.80% SR with a mean trajectory of 4.214 actions; SFT + GRPO achieves 60.98% SR with 5.154 actions. GRPO thus raises SR by 2.18 percentage points, while the mean trajectory grows by 0.940 actions.

Why AR trajectories are shorter. Figures 3 and 4 compare the Qwen3.5-9B+Wan2.2 baseline with AutoPRIA on the same 551 AR episodes, where AutoPRIA answers 347 episodes correctly (62.98%) with 4.737 actions per episode, compared with 302 (54.81%) and 5.650 for the baseline. Paired outcomes reveal two changes in behavior. First, AutoPRIA answers with fewer actions in the outcome groups that cover 510 of the 551 episodes (Figure 3a), for example, from 8.05 to 5.14 actions in the 86 episodes that only it solves, and fewer of its episodes reach the ten-action limit (185 versus 252; Figure 4b). Second, it consults imagination more selectively (Figure 3b), receiving fewer successful imagination returns per 100 decisions in every group (84.3 versus 93.6 overall), and 21.7 returns per correct recognition compared with 32.3. Its SR gain also grows with target visibility, from 5.43 to 8.20 and 10.87 percentage points as shown in Figure 3c.

Termination behavior. Figure 4 shows that AutoPRIA reaches correct answers earlier than the baseline. From the first action onward, it has answered more episodes correctly than the baseline at every action budget, and within three actions it answers 39.4% of all episodes correctly, compared with 31.8% (Figure 4a). The normalized area under these curves is 41.6 versus 33.8, with a 95% scene-bootstrap interval of $[2.0, 12.3]$ for the difference. Within the 261 episodes recognized correctly by both systems, mean actions decrease from 3.847 to 3.330 (13.4%), so the shorter trajectories also hold among shared successes.

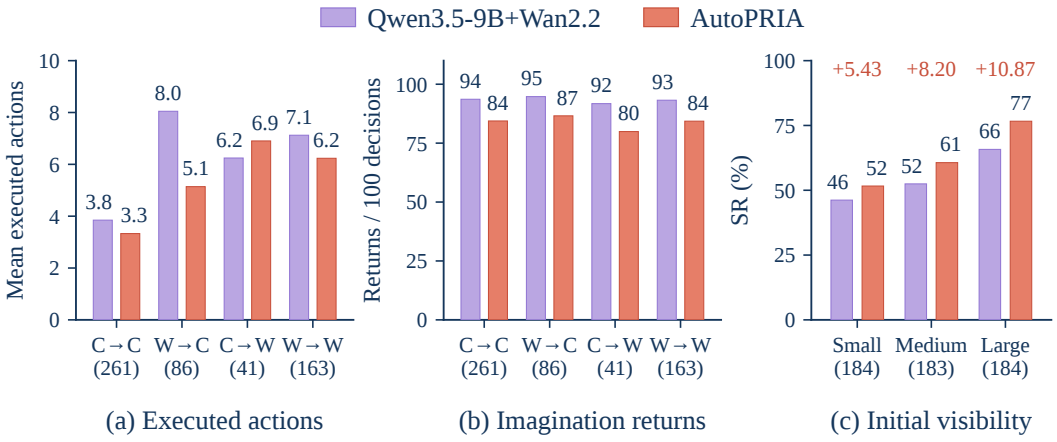

Figure 3: AR result analysis. In (a) and (b), episodes are grouped by correctness from Qwen3.5-9B+Wan2.2 to AutoPRIA, where C and W denote correct and wrong. (a) Mean executed actions. (b) Successful imagination-tool returns per 100 recorded planner/evaluator decisions. (c) SR by initial visible-pixel tercile, with shared boundaries at 1,407.33 and 4,173.00 pixels; annotations give SR differences in percentage points. Sample sizes appear in parentheses below the bars.

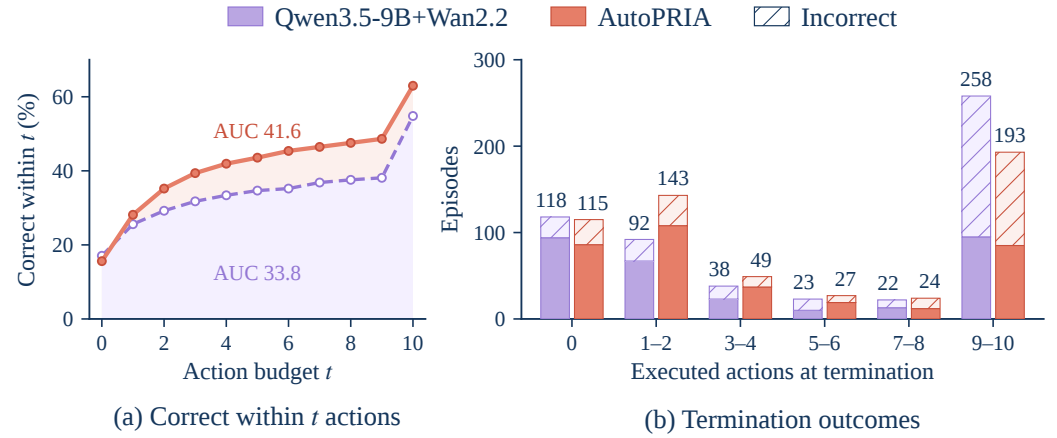

Figure 4: Termination on AR. (a) Percentage of all episodes answered correctly within t executed actions; AUC is the area under each curve over $t = 0, \ldots, 10$, normalized to 0–100. (b) Episodes grouped by the number of executed actions at termination (0, 1–2, …, 9–10); solid and hatched segments denote correct and incorrect outcomes, and labels give total episode counts.

5 Conclusion

We presented AutoPRIA, a unified framework that autonomously organizes embodied decision-making around four fundamental operations spanning the present, the past, the anticipated future, and the environment itself. At every step, the agent chooses which operation to invoke and when to stop gathering evidence, thereby deciding what to know together with what to do. We realize these operations with tool-augmented perception, dual memory, action-conditioned imagination with a fixed world model, and real actions whose observations check earlier predictions, all driven by one policy trained with supervision followed by reinforcement refinement. AutoPRIA outperforms the Qwen3.5-9B baseline with the Wan2.2 world model on all four tasks, with an average relative improvement of 17.7%, while executing fewer actions on AR, IGNav, and RM. Ablations confirm that task-local and cross-scene memory are complementary and that reinforcement refinement raises success, while behavior analyses reveal that the agent consults imagination more selectively and reaches correct answers earlier. Future work will deploy AutoPRIA on physical robots and use mismatches between imagined and observed outcomes to adapt the world model itself.

AI USE STATEMENT

In this work, we used generative AI tools for [generating synthetic data sets, helping develop theoretical models or conceptual frameworks, designing research methods or experiments or providing feedback on them, implementing methods, assisting with translation]. We have not used generative AI tools for [supporting qualitative and thematic data analysis, interpreting results], and [formulating mathematical claims, providing key elements for proving mathematical claims, generating or refining hypotheses, cleaning and reformatting data sets] are not applicable to this work. We have reviewed all AI-assisted work. In particular, all LLM-generated code was verified and tested for correctness by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022. URL https://arxiv.org/abs/2204.01691.

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

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)*, pp. 15791–15801, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bar_Navigation_World_Models_CVPR_2025_paper.html.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. URL https://arxiv.org/abs/2311.15127.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision (3DV)*, pp. 667–676, 2017. URL https://arxiv.org/abs/1709.06158.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. URL https://arxiv.org/abs/2303.03378.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning Universal Policies via Text-Guided Video Generation. *arXiv preprint arXiv:2302.00111*, 2023. URL https://arxiv.org/abs/2302.00111.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. URL https://doi.org/10.1038/s41586-025-09422-z.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv preprint arXiv:1912.01603*, 2019. URL https://arxiv.org/abs/1912.01603.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020. URL https://arxiv.org/abs/1909.12271.

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, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Koh_Pathdreamer_A_World_Model_for_Indoor_Navigation_ICCV_2021_paper.html.

Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3D scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/abs/2204.02960.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.07969.

TaiMing Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. GenEx: Generating an explorable world. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8NlUL0Cv1L.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498, 2024. URL https://doi.org/10.1109/CVPR52733.2024.01560.

NVIDIA. Cosmos-Predict2: World simulation model for physical AI, 2025. URL https://github.com/nvidia-cosmos/cosmos-predict2.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a. URL https://arxiv.org/abs/2501.03575.

NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025b. URL https://arxiv.org/abs/2503.14734.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2109.08238.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://arxiv.org/abs/1904.01201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, 2025. URL https://arxiv.org/abs/2409.19256.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. OpenThinkIMG: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025. URL https://arxiv.org/abs/2505.08617.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Fei-Fei Li, Yejin Choi, and Manling Li. VAGEN: Reinforcing world model reasoning for multi-turn VLM

agents. In *Advances in Neural Information Processing Systems*, 2025a. URL https://arxiv.org/abs/2510.16907.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. KARMA: Augmenting embodied AI agents with long-and-short term memory systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025b. URL https://doi.org/10.1109/ICRA55743.2025.11128047.

Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J. Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yang_Embodied_Amodal_Recognition_Learning_to_Move_to_Perceive_Objects_ICCV_2019_paper.html.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631, 2025a. URL https://proceedings.mlr.press/v267/yang25f.html.

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*, 2025b. URL https://arxiv.org/abs/2507.12508.

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)*, pp. 17294–17303, 2025c. URL https://arxiv.org/abs/2411.17735.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *arXiv preprint arXiv:2602.08236*, 2026. URL https://arxiv.org/abs/2602.08236.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-World: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025. URL https://arxiv.org/abs/2510.18135.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. URL https://arxiv.org/abs/2505.14362.

Chenming Zhu, Jingli Lin, Yilin Long, Peizhou Cao, Tai Wang, Jiangmiao Pang, and Xihui Liu. Thinking with imagination: Agentic visual spatial reasoning with world simulators. *arXiv preprint arXiv:2606.06476*, 2026. URL https://arxiv.org/abs/2606.06476.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. ACTIVE-o3: Empowering MLLMs with active perception via pure reinforcement learning. *arXiv preprint arXiv:2505.21457*, 2025. URL https://arxiv.org/abs/2505.21457.

APPENDIX CONTENTS

A Perception and Evidence Interfaces **14**

B Implementation Details **15**

C Evaluation Protocols **15**

C.1 Task Settings 15

C.2 World-Model Settings 15

C.3 Reported Baselines 15

A PERCEPTION AND EVIDENCE INTERFACES

Table 6 records the AR interfaces. Requests contain a tool name and an arguments object. Image indices reference displayed views (0: front; 1: panorama), subject to each tool's restrictions. The executor validates keys, types, view availability, and parameter domains.

Table 6: Concrete AR operation arguments. The displayed names are concise figure labels.

Operation	Arguments	Contract
Zoom	image_index, bbox_2d	Four finite coordinates define a nonempty crop in the selected view.
Locate	query	A permitted target description selects grounding evidence for the current front view.
Mask	image_index, query	The AR interface uses the front view and a permitted target query.
Rotate	image_index, angle	The angle belongs to the adapter's allowed yaw set, and a new view is rendered.
Depth	image_index	The AR interface visualizes the front-view depth measurement.
Map	Empty object	Returns a visualization of recorded poses and executed actions.
Task-local memory	Current scene	Returns at most two earlier real views of the same target from this episode, the one with the largest target box and the one least similar to it and to the current view, without earlier answers.
Cross-scene memory	Current scene	When the target box is visible in the current view, bank entries are scored with Equation (5), and at most two annotated views from distinct training scenes that reach the calibrated threshold are returned with their categories.
Imagination	Candidate action sequences	Each sequence consists of admissible actions, and the world model predicts its future views from the current state.

The world-model tool rolls out the candidate action sequences proposed by the policy. Cached predictions remain bound to their originating episode state and observation.

Argument and sensing conventions. Boxes use $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ in pixels. Coordinates outside pixel bounds are rescaled from $[0, 1000]$ only if all four lie in that range; boxes are then clipped and checked. AR yaw queries permit $90°$, or $90°$, $180°$, and $270°$ under the expanded contract; these differ from movement increments. Depth uses sensor measurements and Map summarizes explored poses and actions. Sensor access and query limits are declared alongside movement budgets.

Evidence identity and memory. Tool results carry their source state and execution status, including explicit unavailable results. Task-local memory resets between episodes. Cross-scene retrieval uses a training-scene bank and a fixed CLIP ViT-B/32 image encoder. Cross-scene retrieval scores bank entries with Equation (5) and returns at most two entries from distinct training scenes above the calibrated threshold, and task-local memory supplies at most two earlier views of the current episode.

B Implementation Details

This section lists the settings of the evaluated AutoPRIA policy. The policy is initialized from Qwen3.5-9B and trained on AR with the two stages of Section 3.3. The same checkpoint is then applied to IGNav, A-EQA, and RLBench without further training, with task adapters and task-specific dual memory. The Wan2.2-5B world model and the CLIP ViT-B/32 retrieval encoder remain fixed throughout. Table 7 summarizes the training and inference settings.

Training data. Both stages use one AR dataset rendered in eight HM3D scenes, none of which appears in any benchmark, and no example reuses benchmark episode parameters. It contains 1,696 training and 256 validation examples, and the two splits share no scene. Each example asks for the target category together with a four-action plan. Tool examples carry the verified return of the requested tool. Each tool example is also assigned an evidence tier by comparing the recognition of a frozen Qwen3.5-9B judge given the exact return, no return, and a matched control input. For Stage I, each tool example yields four supervised sequences, namely the tool request, the decision after the returned evidence, the answer, and the answer given the evidence, while each no-tool example yields the decision and the answer, giving 6,360 training sequences.

C Evaluation Protocols

AutoPRIA and the Qwen3.5-9B baseline share the episodes, observations, action spaces, budgets, and success criteria in Table 8.

C.1 Task Settings

In AR, the agent recognizes the object marked by a bounding box in its initial view and may execute up to ten actions before answering. Both systems answer once their recognition confidence reaches 0.95 or when the action budget is exhausted. In IGNav, the agent navigates to the location where a goal image was captured. Each decision selects a plan of up to five primitives and executes all but the last two, and an episode succeeds once the agent is within 0.5 m of the goal. In A-EQA, the agent explores an HM3D scene and answers an open-ended question from OpenEQA (Majumdar et al., 2024). The answer is compared with the reference by the LLM-Match scorer of OpenEQA, with GPT-4o-mini as the judge, and the ratings from 1 to 5 are averaged and mapped linearly to 0–100. In manipulation, a 7-DoF arm performs Push Buttons, Slide Block to Color Target, Insert onto Square Peg, and Stack Cups in RLBench (James et al., 2020). Each action specifies the end-effector position, orientation, and gripper state, and an episode succeeds when the task goal is reached within 15 low-level actions.

C.2 World-Model Settings

All tasks use the same fixed Wan2.2-5B model (Wan et al., 2025) without post-training, with the settings listed in Table 7. For AR and A-EQA, Wan2.2 runs four denoising steps and generates 17 frames under a 512×512 area budget, of which 14 are returned at 480×480. The Qwen3.5-9B baseline in Table 2 is evaluated with the same A-EQA setting. IGNav uses eight denoising steps and returns 8 of the 17 generated frames, and manipulation uses 50 denoising steps.

C.3 Reported Baselines

The Qwen2.5-VL-72B rows of Table 1 and all rows of Tables 2 and 3 outside the Qwen3.5-9B block are taken from World-in-World (Zhang et al., 2025). All world models in these rows are

Table 7: Training and inference settings of the evaluated AutoPRIA policy.

Setting	Value
Models and hardware	
Policy	Qwen3.5-9B in bf16, served with vLLM
World model	Wan2.2-5B, fixed
Retrieval encoder	CLIP ViT-B/32 on CPU, fixed
Training hardware	8 NVIDIA H100 GPUs (80 GB) for each stage
Stage I (SFT)	
Trainable parameters	last 8 transformer layers and output head
Optimizer	AdamW, learning rate 10^{-6}, weight decay 0.01, gradient-norm clipping 1.0
Batch and schedule	32 sequences per update (8 GPUs, 1 sequence per GPU, 4 accumulation steps), 2 epochs over 6,360 sequences, 398 optimizer updates (199 per epoch)
Sequence length	8,192 tokens
Token weights	$w_{z,j} = 1$
Stage II (GRPO)	
Initialization	Stage I checkpoint
Trainable parameters	last 4 transformer layers and final normalization
Optimizer	AdamW, constant learning rate 10^{-6}, weight decay 0.01, gradient-norm clipping 1.0
Batch and schedule	8 prompts per step, one update per step, 256 steps
Sampling	group size $G = 4$, temperature 0.8, top-p 0.95
Turns and lengths	at most 3 turns, 8,192 prompt tokens, 3,072 response tokens
Objective	clipping $\epsilon = 0.2$, $\beta = 0$, entropy coefficient 0.001, $\varepsilon = 10^{-6}$, token-mean aggregation
Reward weights	$\lambda_{\text{ans}} = 0.6$, $\lambda_{\text{act}} = 0.4$, $\lambda_{\text{ev}} = \lambda_{\text{pen}} = 1$, -1 for an invalid output, -0.05 per rejected tool turn
Dual memory (AR)	
Task-local memory	up to 2 earlier views of the current episode per query
Cross-scene bank	341 annotated views from 21 training scenes, none in the AR benchmark
Retrieval	top-K with $K = 2$, score $0.8\cos_{\text{crop}} + 0.2\cos_{\text{view}}$, threshold 0.897
Threshold calibration	75th percentile of retrieval scores on 222 calibration examples from 28 non-benchmark scenes
Imagination and decoding (AR)	
Candidate plans	up to 2 per decision, horizon $H = 4$
Wan2.2	4 denoising steps, 17 frames generated under a 512×512 area budget, 14 returned at 480×480
Episode limits	at most 10 actions, answering once recognition confidence reaches 0.95
Decoding	temperature 0.1, at most 3 tool turns, thinking disabled, seed 23
Wan2.2 on the other tasks	
IGNav	8 denoising steps, 17 frames generated and 8 returned at 480×480
A-EQA	same as AR
RLBench	50 denoising steps

Table 8: Evaluation protocols shared by AutoPRIA and the Qwen3.5-9B baseline.

Task	Data	Observations and actions	Budget	Success and score
AR	551 episodes in 29 Matterport3D scenes	Panorama and 90° front view, discrete navigation primitives	10 executed actions	Final category equals the label
IGNav	144 episodes in 87 HM3D scenes	Panorama and 90° front view, discrete navigation primitives	21 decisions of up to 3 primitives each	Agent within 0.5 m of the goal
A-EQA	184 questions in 54 HM3D scenes	Four 105° perspective views, discrete navigation primitives	250 primitive actions	Judge rating from 1 to 5, mapped to 0–100
Manipulation	200 episodes, 50 for each of 4 RLBench tasks	End-effector position, orientation, and gripper state of a 7-DoF arm	15 low-level actions	Task goal reached within the budget

used without World-in-World post-training. We evaluate the Qwen3.5-9B rows under the protocols above, replacing the World-in-World policy with Qwen3.5-9B while keeping its planning pipeline.