COWORLD REASONER: AGENTIC MULTIMODAL REASONING WITH THE EVOLVING VIDEO WORLD MODEL-DRIVEN HARNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

As Vision-Language Models (VLMs) advance, they are increasingly expected to perform spatiotemporal reasoning over motion, actions, and view changes beyond what the input images show. Many multimodal reasoning tasks require reasoning about unobserved scenes, transitions, and relations. Video World Models (VWMs) provide a natural unified engine for visual imagination by generating task-relevant visual content as additional evidence. However, existing approaches typically invoke a VWM through fixed pipelines centered on a predefined form of visual imagination, such as novel-view synthesis. To support flexible visual imaginations for diverse reasoning tasks, we introduce **CoWorld Reasoner**, an agentic framework that uses a VWM as a unified visual reasoning engine. Its Video World Model-Driven Harness exposes multiple visual operations through tools and uses skills to guide the VLM in deciding whether to imagine, selecting an operation, constructing a generation request, and interpreting the resulting visual evidence. We further introduce VLM-VWM Harness Co-Evolution, which jointly revises tools, skills, and execution code based on execution feedback while keeping both model weights fixed. Across seven benchmarks spanning spatial intelligence, temporal and causal reasoning, and perceptual reasoning, CoWorld Reasoner outperforms direct reasoning with the same VLM. The strongest variant improves on all seven benchmarks, including relative accuracy gains of 11.8% on MMSI and 13.0% on ERQA, and outperforms all five evolving baselines using the same backbone across five benchmarks. Code and data are included in the Supplementary Material.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated strong instruction-following and reasoning abilities, but they operate primarily over text. Vision-Language Models (VLMs) extend these capabilities to visual inputs, enabling multimodal understanding and reasoning. Recent VLMs such as Qwen3.8 (Qwen Team, 2026) and Gemini 3.8 (Google DeepMind, 2026) can effectively interpret images and analyze charts, and they are increasingly used in real-world multimodal applications.

Beyond static image understanding, VLMs are increasingly expected to perform spatiotemporal reasoning over motion, actions, and view changes. For example, a model may need to predict whether two objects would overlap if one were moved, reconstruct what happened between two observations, or infer what another viewpoint would reveal. Solving such problems requires going beyond observed visual evidence to reason about unobserved transitions and relations. One promising approach is visual imagination: explicitly generating task-relevant visual predictions of what is not observed and using them as additional evidence for reasoning. This raises a fundamental question: how can a VLM generate and exploit the right visual imagination for the reasoning problem at hand?

Video World Models (VWMs), such as MiniMax H3 (MiniMax, 2026) and LTX-Video (HaCohen et al., 2024), provide a natural engine for visual imagination by modeling visual dynamics from large-scale video data. Given input images and an appropriate visual operation, they can generate video depicting unobserved views, intermediate states, future events, or intervention outcomes. Recent approaches, including MindJourney (Yang et al., 2025) and AVIC (Yu et al., 2026), use a

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.

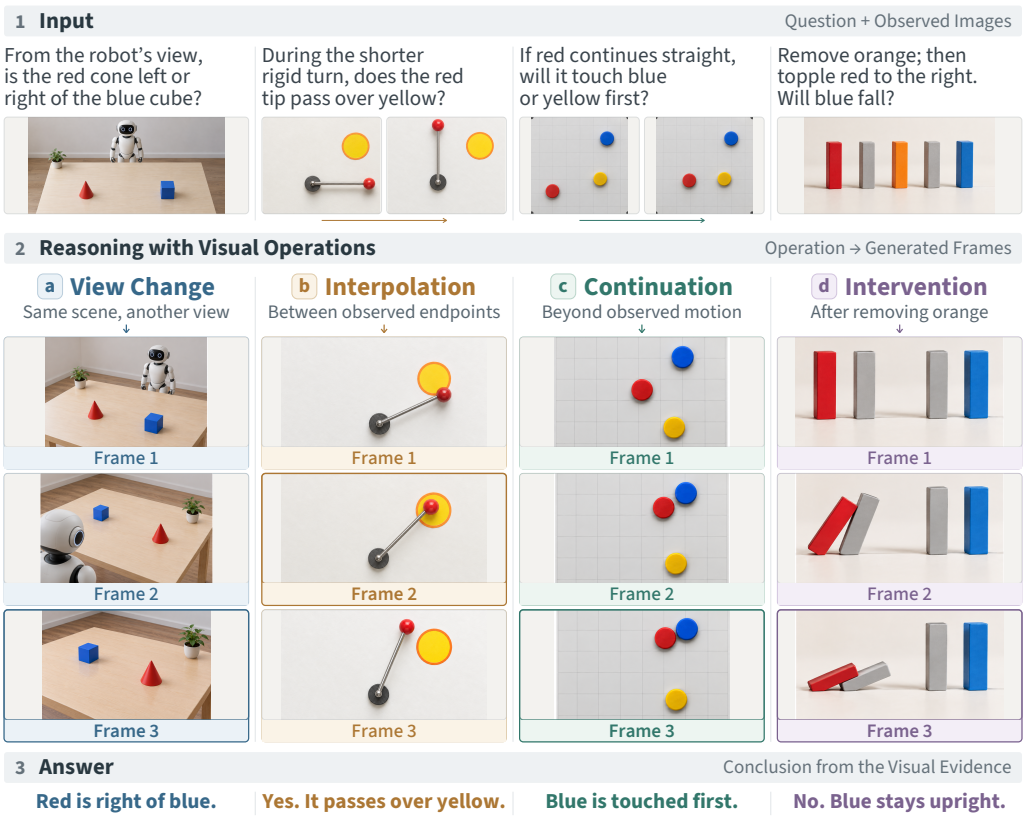


Figure 1: Illustration of how different multimodal reasoning tasks require different visual operations: (a) view change for spatial perspective taking; (b) interpolation to infer hidden motion between observed endpoints; (c) continuation to predict first contact from observed motion; and (d) intervention to reason about the outcome of a controlled change.

VWM to generate additional visual evidence for spatial reasoning, including synthesized views for cross-view reasoning. However, these approaches typically invoke a VWM through fixed pipelines centered on a predefined form of visual imagination, such as novel-view synthesis. This limits their applicability to diverse reasoning problems, which may require fundamentally different forms of visual imagination: changing viewpoints for perspective taking, interpolating between observations to reconstruct hidden transitions, continuing observed dynamics to anticipate future events, or intervening on the scene to reason about intervention outcomes, as illustrated in Figure 1. A visual reasoner should therefore actively determine what to imagine according to the reasoning problem and how to exploit the generated visual evidence to solve it.

To enable task-dependent visual imagination, we introduce **CoWorld Reasoner**, an agentic framework that uses a VWM as a unified visual reasoning engine. Unlike an agent confined to the observed images, it selects among VWM-driven tools to generate unseen views, intermediate states, continuations, or intervention outcomes according to the question. Skills guide when to invoke these tools, how to construct requests, and how to interpret the generated video; a verifier and execution code complete the harness. A predefined harness may fail to elicit the visual evidence required by different reasoning tasks or to guide its interpretation. Refinement must coordinate evidence generation and interpretation, because what the VLM can infer depends on what the VWM can reliably generate. We therefore introduce VLM-VWM Harness Co-Evolution, which iteratively refines the interface between the VLM and VWM by jointly revising tools, skills, and execution code based on execution feedback. With both model weights fixed, co-evolution improves how the two models interact and use task-relevant visual evidence.

We evaluated CoWorld Reasoner on seven benchmarks spanning spatial intelligence (Yang et al., 2026a; Ray et al., 2025; Zhang et al., 2026c; Wasi et al., 2026), temporal and causal reasoning (Gemini Robotics Team et al., 2025; Ma et al., 2026), and perception (Lin et al., 2026). Across

VLM/VWM variants, CoWorld Reasoner improves aggregate accuracy over direct reasoning with the same VLM. The strongest variant (GPT-5.6 and MiniMax H3) improves all seven benchmarks, with relative gains of 11.8% on MMSI and 13.0% on ERQA. It also outperforms all five evolving baselines using the same backbone on five benchmarks. Further analysis shows that gains require both VWM evidence and co-evolved control of its request and interpretation.

Our main contributions are summarized below:

1. We leverage a VWM as a unified visual reasoning engine for visual imagination, enabling the agent to generate task-relevant visual evidence through diverse visual operations.
2. We design a Video World Model-Driven Harness that exposes diverse visual operations through standardized tools and encodes reusable reasoning workflows as skills, enabling the agent to decide whether visual imagination is needed, which operation to invoke, and how to interpret the generated visual evidence.
3. We introduce VLM-VWM Harness Co-Evolution, which iteratively refines the interface between the VLM and VWM by jointly revising tools, skills, and execution code based on execution feedback.
4. We evaluate CoWorld Reasoner with multiple VLMs and VWMs across seven benchmarks spanning spatial, temporal, causal, and perceptual reasoning, showing gains over direct reasoning and strong performance against evolving baselines, while analyzing the contributions of the harness and its co-evolution.

2 Related Work

2.1 Multimodal Reasoning

Recent work uses visual evidence to support multimodal reasoning, particularly spatial tasks. MVoT (Li et al., 2025) and ThinkMorph (Gu et al., 2025) incorporate generated images into intermediate reasoning steps, while Omni-R1 (Cheng et al., 2026) uses image generation to unify diverse visual operations. Thinking with Video (Tong et al., 2026) uses video generation for visual reasoning. For spatial reasoning, SenseNova-SI (Cai et al., 2026) and OpenSpatial (Liu et al., 2026) use large-scale training datasets, while World2VLM (Zhang et al., 2026b) uses views generated by world models as supervision. At inference time, MindJourney (Yang et al., 2025) searches camera movements to explore a scene, and SpatialDreamer (Cao et al., 2025) uses a learned policy to explore generated views. AVIC (Yu et al., 2026) controls world-model use and the amount of visual content generated.

2.2 World Models

World models capture the structure and dynamics of environments to support prediction and simulation. Video generation models such as Wan (Team Wan, 2025), Seedance 2.0 (Team Seedance et al., 2026), and MiniMax H3 (MiniMax, 2026) learn visual dynamics from video data, and Cosmos-Predict2.5 (NVIDIA, 2025) focuses on world simulation. Marble (World Labs, 2025a) generates explicit 3D environments that users can explore and edit. Interactive models such as RTFM (World Labs, 2025b) generate video in real time in response to user actions. Our approach treats a VWM as a unified visual reasoning engine and exposes it through multiple VWM-driven tools.

2.3 Self-Evolving Agents

Recent approaches improve agents by iteratively revising prompts, skills, and code. GEPA (Agrawal et al., 2026) optimizes prompts from execution traces, while SkillOpt (Yang et al., 2026b) refines reusable skills from execution feedback. DGM (Zhang et al., 2025) revises the agent's own code, while Meta-Harness (Lee et al., 2026) optimizes harness implementations. CoEvoSkills (Zhang et al., 2026a) jointly evolves skills and a verifier that guides their refinement. Our co-evolution supervisor jointly revises the VWM-driven harness to improve both the generation of visual evidence and the agent's interpretation of it.

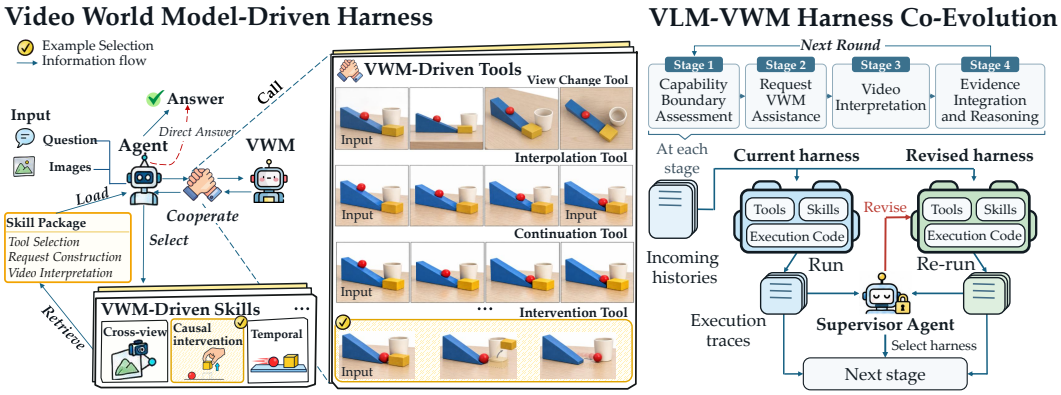


Figure 2: Overview of CoWorld Reasoner. Left: VWM-driven tools provide visual operations, while skills encode reusable workflows for the agent's tool selection, request construction, and video interpretation. Right: at each of four reasoning stages, a supervisor reviews execution traces and revises the relevant tools, skills, and execution code. The agent reruns the stage from the same incoming interaction histories. The selected harness and histories are passed to the next stage. These four stages repeat over multiple evolution rounds.

3 METHOD

CoWorld Reasoner uses a VWM as a unified visual reasoning engine. It has two parts: a Video World Model-Driven Harness for requesting and interpreting visual evidence, and VLM-VWM Harness Co-Evolution for improving the harness through execution feedback. Figure 2 illustrates both parts, and Algorithm 1 specifies the co-evolution procedure.

We consider a question q with an ordered sequence of n observed images $\mathbf{I} = (I_1, \ldots, I_n)$. A VLM M powers the agent, and a VWM W provides the underlying visual generation capabilities for the VWM-driven tools in an editable harness H. The answering process can be formulated as

$$\hat{a} = \mathrm{CoWorldReasoner}(q, \mathbf{I}; M, W, H), \tag{1}$$

where $\mathrm{CoWorldReasoner}$ denotes harness-guided execution and $\hat{a}$ is the predicted answer.

3.1 VIDEO WORLD MODEL-DRIVEN HARNESS

Harness components. The harness comprises L VWM-driven tools T_ℓ, J VWM-driven skills S_j, a verifier V, and an execution layer E:

$$H = (\mathcal{T}, \mathcal{S}, V, E), \qquad \mathcal{T} = \{T_\ell\}_{\ell=1}^{L}, \quad \mathcal{S} = \{S_j\}_{j=1}^{J}. \tag{2}$$

Each tool T_ℓ exposes a visual operation through a standardized interface. Our experiments use four types of visual operations: view changes, temporal interpolation, sequence continuation, and intervention. Appendix A.4 summarizes these operations and the corresponding reasoning skills. A request r specifies which source images to use and in what order, the intended operation, and the conditions to preserve. The tool executes this request through the shared VWM and returns the generated frame sequence $\widetilde{\mathbf{I}}$:

$$\widetilde{\mathbf{I}} = T_\ell(\mathbf{I}, r; W). \tag{3}$$

A skill S_j guides cross-view, temporal, or causal reasoning through a reusable workflow for selecting visual operations, constructing requests, and interpreting the returned frames. For example, a cross-view skill guides the agent to request a connecting view while preserving scene anchors, then compare the generated frames with the original observations to resolve the spatial relation. The verifier V describes visible events in generated frames and reports artifacts and uncertainty. The execution layer E contains the code that provides execution instructions, coordinates component calls, and adds their results to the agent's interaction history.

Agent execution. Given $(q, \mathbf{I})$, the agent assesses whether the observed images resolve the question and loads a relevant skill when further guidance is useful. Let h denote the current interaction history

Algorithm 1 VLM-VWM Harness Co-Evolution.

Require: VLM: M; VWM: W; Initial harness: $H_1^{(0)}$; Development set: $\mathcal{D}_{\text{dev}}$; Evolution rounds: R; Revision budget per stage: B

for $t = 1, \ldots, R$ **do**
  $\tau_0^* \leftarrow \textsc{InitHistories}(\mathcal{D}_{\text{dev}})$
  for $k = 1, \ldots, 4$ **do**
    $\mathcal{C}_k \leftarrow \emptyset$
    for $e = 0, \ldots, B$ **do**
      $\tau_k^{(e)} \leftarrow \textsc{RunStage}(\mathcal{D}_{\text{dev}}, H_k^{(e)}, \tau_{k-1}^*, k)$
      $P_k^{(e)} \leftarrow \textsc{Summarize}(\tau_k^{(e)}, H_k^{(e)}, k;$ available scorer feedback$)$
      Add $(H_k^{(e)}, \tau_k^{(e)}, P_k^{(e)})$ to $\mathcal{C}_k$
      if $e = B$ **then break**
      $\Delta_k^{(e)} \leftarrow \text{Review}_M(H_k^{(e)}, P_k^{(e)}; k)$
      if the supervisor ends revision **then break**
      $H_k^{(e+1)} \leftarrow \textsc{Revise}(H_k^{(e)}, \Delta_k^{(e)})$
    $(H_k^*, \tau_k^*) \leftarrow \text{Select}_M(\mathcal{C}_k; k)$
    if $k < 4$ **then** $H_{k+1}^{(0)} \leftarrow H_k^*$
  if $t < R$ **then** $H_1^{(0)} \leftarrow H_4^*$
return H_4^*

and s the loaded skill. The same VLM first decides on skill loading, then on the tool and request:

$$\begin{aligned} s &= \text{Load}_M(q, \mathbf{I}, h; H), \\ (\ell, r) &= \text{Request}_M(q, \mathbf{I}, h, s; H). \end{aligned} \tag{4}$$

Here $s \in \mathcal{S} \cup \{\emptyset\}$, with $s = \emptyset$ when skill loading is skipped. When generation is needed, the agent selects a tool index $\ell \in \{1, \ldots, L\}$ and request r to obtain the frames in Eq. 3. If generation is skipped, $(\ell, r) = (\emptyset, \emptyset)$ and $\widetilde{\mathbf{I}} = \emptyset$.

When generated frames are available, the agent compares them with the original observations. It checks whether the requested transformation was realized and interprets the evidence relevant to q. For closer inspection, it may request a description v from V, which analyzes only the generated frames using q and r as context. If verification is skipped, $v = \emptyset$. Let h^+ denote the interaction history updated with the loaded guidance, calls, and returned evidence. The agent predicts the answer $\hat{a}$ from the original observations and the available generated evidence:

$$\begin{aligned} v &= \begin{cases} V(q, r, \widetilde{\mathbf{I}}), & \text{if verification is requested,} \\ \emptyset, & \text{otherwise} \end{cases} \\ \hat{a} &= M\left(q, \mathbf{I}, \widetilde{\mathbf{I}}, v; h^+\right). \end{aligned} \tag{5}$$

3.2 VLM-VWM Harness Co-Evolution

VLM-VWM Harness Co-Evolution jointly refines the components that govern visual generation and evidence use, while keeping the VLM and VWM weights fixed. These components are coupled: changes to request construction can alter the visual cues available in generated frames, which in turn affect the guidance needed to interpret and integrate them.

The procedure alternates between forward execution and feedback-guided revision. At each of four stages, the agent first runs the current harness on the development set. A supervisor powered by the same VLM M uses the resulting feedback to revise the harness and rerun that stage from the same incoming interaction histories. This cycle can repeat before the supervisor selects the best candidate and passes its harness and histories to the next stage. We repeat these four stages for R evolution rounds.

Progressive execution. During *Capability Boundary Assessment* ($k = 1$), the agent judges whether the VWM can supply useful missing evidence and may load a relevant skill. In *Request VWM Assistance* ($k = 2$), it constructs and executes a generation request or skips generation. During

Video Interpretation ($k = 3$), it reviews the original observations and interprets any generated frames alongside them, optionally requesting verification. Finally, in *Evidence Integration and Reasoning* ($k = 4$), it combines the original observations with generated evidence to infer the final answer. The stage and candidate notation below is local to the current round. At stage k, each candidate $H_k^{(e)}$ after e revisions is run from the same selected interaction histories τ_{k-1}^*, preserving earlier decisions and evidence. Execution on the development set $\mathcal{D}_{\text{dev}}$ produces updated histories $\tau_k^{(e)}$:

$$\tau_k^{(e)} = \text{RunStage}\Big(\mathcal{D}_{\text{dev}}, H_k^{(e)}, \tau_{k-1}^*, k\Big). \tag{6}$$

Here $e = 0$ denotes the stage's initial candidate, and τ_0^* contains the initialized task histories.

Feedback-guided revision and selection. Each run produces stage feedback $P_k^{(e)}$: stage-specific metrics, sampled execution traces, and, when an answer is available, correctness feedback from an external answer scorer. During capability boundary assessment, for example, the supervisor examines skill-loading rates and sampled load/skip decisions to refine guidance on when and which skills to load. The supervisor reads additional execution traces, generated frames, and harness source files as needed to link recurring problems to the relevant harness components.

The proposed changes $\Delta_k^{(e)}$ can affect tool interfaces, skill workflows, verifier and stage instructions, and execution code:

$$\begin{aligned} \Delta_k^{(e)} &= \text{Review}_M\Big(H_k^{(e)}, P_k^{(e)}; k\Big), \\ H_k^{(e+1)} &= \text{Revise}\Big(H_k^{(e)}, \Delta_k^{(e)}\Big). \end{aligned} \tag{7}$$

Each revised candidate is evaluated via Eq. 6. Appendix A.5 details the stage-specific metrics and diagnostic sampling procedure.

Each stage allows up to B revisions, as specified in Appendix A.3; the supervisor may stop earlier. Let $\mathcal{C}_k$ collect the initial candidate and all completed revisions, each stored with its histories and stage feedback. The supervisor selects the best completed candidate by comparing stage-specific metrics and sampled execution traces, weighing targeted improvements against regressions across tasks. This selection returns the harness H_k^* and histories τ_k^*:

$$\begin{aligned} (H_k^*, \tau_k^*) &= \text{Select}_M(\mathcal{C}_k; k), \\ H_{k+1}^{(0)} &= H_k^*, \qquad k < 4. \end{aligned} \tag{8}$$

At each hand-off, the next stage resumes from the selected harness and its corresponding histories. After Stage 4, the selected harness H_4^* initializes Stage 1 of the next round, with task histories reinitialized from the original questions and images. Later revisions can thus affect earlier-stage behavior in subsequent rounds. After R rounds, the final selected harness is used for inference.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluated CoWorld Reasoner on the seven benchmarks in Table 1. MMSI (Yang et al., 2026a) assesses reasoning about positions and motion across images; SAT (Ray et al., 2025) covers ego and object motion, action outcomes, goals, and perspective taking; and Spin (Zhang et al., 2026c) tests perspective taking under translation and rotation. SpatiaLab (Wasi et al., 2026) uses Multiple-Choice (MCQ) and Open-Ended (Open) questions about real scenes. ERQA (Gemini Robotics Team et al., 2025) combines spatial, physical, and world knowledge, while CausalSpatial (Ma et al., 2026) tests reasoning about collisions, occlusions, and trajectories under object interventions. PerceptionBench (Lin et al., 2026) covers counting, attributes, and text reading.

Baselines. We compare with five evolving agents adapted to the same GPT-5.6 inputs and scoring protocol (HSI, ReasoningBank, Meta-Harness, ExpeL, and GEPA) and use direct GPT-5.6 reasoning as the reference. The comparison tables also include published results for fine-tuned embodied models, specialist spatial models, and vision-based agents; mechanisms, configurations, and source-protocol notes are provided in Appendix A.

Table 1: Multimodal reasoning accuracy (%, ↑). Base Models and Fine-tuned Models use published results; dashes denote benchmark scores unreported in these sources. Avg. requires complete benchmark coverage. SL and PB denote SpatiaLab and PerceptionBench.

	Spatial Intelligence					Temp. & Causal		Perception	
Method	**MMSI**	**SAT**	**Spin**	**SL-MCQ**	**SL-Open**	**ERQA**	**Causal**	**PB**	**Avg.**
Base Models									
GPT-5-mini	34.10	75.40	—	54.29	**40.93**	—	45.44	—	—
GPT-5	41.90	80.10	78.50	—	—	—	54.17	—	—
Gemini-2.5-Pro	36.90	76.70	75.20	50.07	33.61	55.80	50.55	—	—
Qwen3-VL-8B	31.10	52.70	67.40	—	—	40.10	43.53	—	—
Fine-tuned Models									
MiMo-Embodied	30.20	78.67	—	—	—	46.75	—	—	—
RoboBrain2.0	28.50	86.67	—	—	—	46.00	—	—	—
Evolving Agents									
GPT-5.6	38.35	88.33	82.24	56.64	36.21	51.75	60.72	50.84	58.29
+ *HSI*	35.05	87.67	73.46	55.71	29.57	57.25	58.06	48.92	54.39
+ *ReasoningBank*	36.80	86.67	81.69	56.57	30.29	59.00	60.65	50.07	57.30
+ *Meta-Harness*	37.42	88.33	83.31	55.86	30.57	**60.00**	60.07	49.83	57.63
+ *ExpeL*	37.42	89.00	82.36	56.64	29.57	57.00	61.22	49.12	57.25
+ *GEPA*	38.87	87.67	79.92	54.86	28.21	59.75	**62.51**	47.61	56.25
+ CoWorld Reasoner	**42.89**	**89.67**	**83.76**	**57.36**	37.64	58.50	60.79	**51.35**	**59.69**
Gain (pp)	+4.54	+1.33	+1.51	+0.71	+1.43	+6.75	+0.07	+0.51	+1.40

Evaluation Protocol. We use 120 development examples, excluded from testing, to evolve CoWorld Reasoner and the five evolving baselines; 30 examples come from each of MMSI, Spin, CausalSpatial, and PerceptionBench. Accuracy is $100 \times N_{\text{correct}}/N$; differences are in percentage points, and averages are item-weighted. We measure error recovery as accuracy on the subset of test items that the GPT-5.6 direct reasoning baseline answers incorrectly. Appendix A details evaluation sizes and scoring protocols.

Implementation. We used GPT-5.6 for the task agent, co-evolution supervisor, and external answer scorer, and MiniMax H3 Max Turbo (MiniMax, 2026) as the VWM. We tested the final co-evolved harness without further updates. Appendix A.3 specifies model and generation settings; Appendix A.5 details candidate selection.

4.2 Main Results

Overall Performance. Table 1 shows that CoWorld Reasoner, using its final co-evolved harness, improves over the GPT-5.6 direct reasoning baseline on all seven benchmarks, especially on MMSI and ERQA, with gains of 4.54 and 6.75 percentage points, respectively. It leads all five evolving baselines on five benchmarks and achieves the highest aggregate accuracy. All five evolving baselines also score below the GPT-5.6 direct reasoning baseline overall. The advantage over the best evolving baseline reaches 7.07 points on open-ended SpatiaLab. Gains also extend to SAT, SpatiaLab, and ERQA, which provide no development examples; these results show that the final harness transfers across benchmarks. The skills organize reasoning around shared spatial and causal operations, so the harness can reuse these workflows across benchmarks. Appendix A.6 further compares development and test performance.

Spatial Intelligence. As shown in Table 2, gains span all four MMSI categories and reach 6.56 points for positional relationships. Overall accuracy approaches SenseNova-SI's reported result. SAT perspective taking reaches 84.85%, the highest among the compared methods; goal reasoning also improves, while object motion declines. Spin likewise improves in view changes, mental rotation, and perspective taking, as detailed in Appendix A.7. The cross-view skill guides the agent to track object identity and interpret relations in the question's reference frame.

Table 2: Spatial intelligence accuracy on MMSI and SAT (%, ↑). PR, Mot., and MSR denote positional relationship, motion, and multi-step reasoning. Dashes denote scores unreported in the source publications. Best results are bolded.

	MMSI					SAT					
Method	**PR**	**Attr.**	**Mot.**	**MSR**	**All**	**Ego**	**Obj**	**Act.**	**Goal**	**Pers.**	**All**
Base Models											
o1	—	—	—	—	—	78.30	82.60	73.00	73.50	69.70	74.60
GPT-4.1	30.65	30.77	34.67	28.79	30.90	95.70	73.90	78.30	88.20	39.40	74.00
GPT-5.4	—	—	—	—	37.50	—	—	—	—	—	—
Qwen3-VL-8B	33.72	26.15	28.00	29.80	31.10	—	—	—	—	—	52.70
Fine-tuned Models											
SenseNova-SI	**52.87**	40.77	32.67	27.78	**43.30**	—	—	—	—	—	—
ViewFusion	39.66	35.38	36.67	23.23	35.40	—	—	—	—	—	—
OpenSpatial	—	—	—	—	41.90	—	—	—	—	—	—
World2VLM	—	—	—	—	—	**100.00**	65.22	67.57	79.41	57.57	72.67
MiMo-Embodied	—	—	—	—	30.20	—	—	—	—	—	78.67
RoboBrain2.0	26.44	33.85	30.00	29.29	28.50	—	—	—	—	—	86.67
Spatial-Interactor	—	—	—	—	32.20	—	—	—	—	—	74.00
Agents with Vision Models											
AVIC	31.23	44.62	35.33	32.32	33.80	**100.00**	78.20	83.70	85.20	54.50	79.30
AVIC-R	—	—	—	—	—	91.30	86.90	83.70	85.20	57.50	80.00
MindJourney	—	—	—	—	—	95.70	82.60	83.80	88.20	75.80	84.70
SandboxVLM	—	—	—	—	—	**100.00**	82.60	79.40	**92.90**	70.00	84.10
Evolving Agents											
GPT-5.6	35.59	42.40	40.00	41.62	38.35	**100.00**	95.65	89.19	83.82	78.79	88.33
+ *HSI*	30.22	42.40	38.62	40.10	35.05	**100.00**	91.30	91.89	82.35	77.27	87.67
+ *ReasoningBank*	33.60	43.20	40.69	38.07	36.80	97.83	91.30	87.84	83.82	77.27	86.67
+ *Meta-Harness*	34.19	45.60	36.55	41.12	37.42	91.30	**97.83**	**93.24**	80.88	81.82	88.33
+ *ExpeL*	32.41	**48.80**	37.24	**43.15**	37.42	**100.00**	**97.83**	89.19	83.82	80.30	89.00
+ *GEPA*	36.18	44.00	39.31	42.13	38.87	93.48	93.48	89.19	85.29	80.30	87.67
+ CoWorld Reasoner	42.15	46.40	**43.45**	42.13	42.89	**100.00**	93.48	89.19	85.29	**84.85**	**89.67**
Gain (pp)	+6.56	+4.00	+3.45	+0.51	+4.54	+0.00	-2.17	+0.00	+1.47	+6.06	+1.33

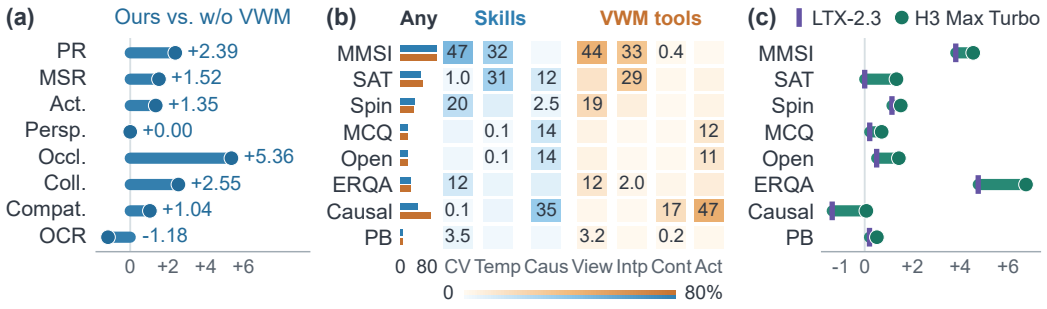


Figure 3: Analysis of VWM contributions. (a) Subtask gains over Ours (w/o VWM). (b) Skill-loading and tool-use rates by benchmark; tool calls are grouped by visual operation (% of items). (c) Benchmark gains over the GPT-5.6 direct reasoning baseline with two VWMs.

4.3 Analysis

Contribution of VWM-Driven Tools. To assess the contribution of generated evidence, we disabled VWM-driven tools and the verifier while retaining the evolved skills and four-stage execution. Figure 3a shows the largest tool gains on CausalSpatial Occlusion (Occl.) and Collision

(Coll.). Generated sequences expose the intermediate interactions needed to answer these questions. SAT Perspective Taking (Persp.) already improves without tools and gains no further benefit from generation. Gains without tools indicate that harness guidance can also improve reasoning over the original images.

VWM-Driven Skill and Tool Use. We measured skill-loading and tool-use rates to examine task-dependent selection. Figure 3b shows that on MMSI, the agent mainly loads Cross-View (CV) and Temporal (Temp) skills and calls View-Change (View) and Interpolation (Intp) tools to relate views and infer hidden motion. On CausalSpatial, it uses the Causal-Intervention (Caus) skill with Intervention (Act) and Sequence-Continuation (Cont) tools to predict intervention outcomes. Skills and tools are rarely used on PerceptionBench, where direct perception often suffices.

Effect of VWM Choice. We compared H3 Max Turbo with LTX-2.3 using GPT-5.6, retaining the evolved skills and four-stage execution. Both VWM variants improve overall accuracy and performance on MMSI, Spin, and SpatiaLab, as shown in Figure 3c. The H3 variant yields larger gains across the evaluated benchmarks and also improves SAT, while LTX-2.3 falls below GPT-5.6 on CausalSpatial. These results show that the harness can reuse spatial reasoning workflows across VWMs and benefit from a stronger visual generation engine.

Recovery from GPT-5.6 Errors. For each benchmark, we evaluated all harness variants and evolving baselines on the subset of test examples that the GPT-5.6 direct reasoning baseline answered incorrectly. CoWorld Reasoner leads on MMSI, SpatiaLab, CausalSpatial, and PerceptionBench (Figure 4). Without VWM tools, it exceeds the evolving baselines on MMSI and SpatiaLab; adding them improves CausalSpatial recovery, leaves SAT unchanged, and reduces ERQA recovery. These CausalSpatial gains mirror the occlusion and collision improvements in Figure 3a: the harness resolves additional GPT-5.6 errors even at similar overall accuracy.

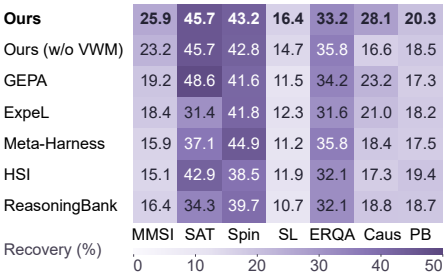


Figure 4: Recovery on errors made by the GPT-5.6 direct reasoning baseline. SL combines SpatiaLab MCQ and Open.

Harness Transfer Across VLMs. We transferred the GPT-5.6 harness to Grok-4.5 and ran one additional co-evolution round before evaluation. Figure 5 shows that item-weighted accuracy rises from 32.86% to 39.78%, with the largest gains on PerceptionBench and Spin. MMSI and CausalSpatial also improve, while SAT declines slightly. The gains span tasks that use the original images directly and tasks that require reasoning about unseen views or outcomes. After adaptation, Grok-4.5 reuses the source harness's procedures for selecting, interpreting, and integrating visual evidence across benchmarks.

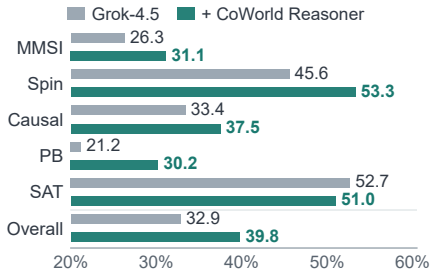


Figure 5: Grok-4.5 accuracy with direct reasoning and the transferred harness.

5 Conclusion and Future Work

We presented CoWorld Reasoner, which treats a VWM as a unified visual reasoning engine through a Video World Model-Driven Harness of tools, skills, and a verifier. VLM-VWM Harness Co-Evolution uses execution feedback to improve how the agent requests, evaluates, and integrates visual evidence. CoWorld Reasoner improves over direct reasoning with the same VLM on all seven benchmarks and outperforms all five evolving baselines on five of these benchmarks. Agentic tool use and harness co-evolution extend VWM-based reasoning beyond fixed novel-view synthesis pipelines. Future work will explore test-time training to co-evolve the harness, VLM, and VWM for visual generation and multimodal reasoning across diverse reasoning tasks.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing, identifying relevant literature, and reviewing implementation code.

ETHICS STATEMENT

We evaluated publicly released benchmark datasets without recruiting human participants or collecting new personal data. We are not aware of ethics concerns specific to this study.

REPRODUCIBILITY STATEMENT

The Method and Experiments sections describe the algorithm, data splits, and evaluation protocol. The Appendix specifies model and generation settings, scoring, and candidate selection. The Supplementary Material provides implementation code, development data, and run instructions; full benchmark data are available from their original public releases.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0e9e708b6f48e14fd0ac29e167413f76-Paper-Conference.pdf.

Zhongang Cai, Ruisi Wang, Chenyang Gu, et al. Scaling spatial intelligence with multimodal foundation models, 2026. URL https://arxiv.org/abs/2511.13719.

Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. SpatialDreamer: Incentivizing Spatial Reasoning via Active Mental Imagery. *arXiv preprint arXiv:2512.07733*, 2025. URL https://arxiv.org/abs/2512.07733v1. Version 1.

Dongjie Cheng, Yongqi Li, Zhixin Ma, Hongru Cai, Yupeng Hu, Wenjie Wang, Liqiang Nie, and Wenjie Li. Omni-R1: Towards the unified generative paradigm for multimodal reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38907–38924, 2026. doi: 10.18653/v1/2026.findings-acl.1937. URL https://aclanthology.org/2026.findings-acl.1937/.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, et al. Gemini Robotics: Bringing AI into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

Google DeepMind. Gemini 3.8 Flash. Model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. *arXiv preprint arXiv:2510.27492*, 2025. URL https://arxiv.org/abs/2510.27492.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-Harness: End-to-end optimization of model harnesses, 2026. URL https://arxiv.org/abs/2603.28052.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. URL https://arxiv.org/abs/2501.07542.

Kehan Li, Bohan Hou, Minghao Zhu, et al. RynnBrain 1.1: Towards more capable and generalizable embodied foundation model, 2026. URL https://arxiv.org/abs/2607.17977.

Zichao Lin, Yifeng Xie, Bowen Qu, et al. PerceptionBench: Evaluating atomic visual perception in multimodal large language models, 2026. URL https://arxiv.org/abs/2607.24957.

Jianhui Liu, Haoze Sun, Wenbo Li, et al. OpenSpatial: A principled data engine for empowering spatial intelligence, 2026. URL https://arxiv.org/abs/2604.07296.

Wenxin Ma, Chenlong Wang, Ruisheng Yuan, et al. CausalSpatial: A benchmark for object-centric causal spatial reasoning, 2026. URL https://arxiv.org/abs/2601.13304.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research, 2026. URL https://www.minimax.io/blog/minimax-h3.

NVIDIA. World Simulation with Video Foundation Models for Physical AI. *arXiv preprint arXiv:2511.00062*, 2025. URL https://arxiv.org/abs/2511.00062v2. Version 2.

Siru Ouyang, Jun Yan, I-Hung Hsu, et al. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jL7fwchScm.

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

Arijit Ray, Jiafei Duan, Ellis Brown, et al. SAT: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Team Seedance, De Chen, Liyang Chen, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314v2. Version 2.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, Xinchi Chen, Jun Zhao, Xuanjing Huang, and Xipeng Qiu. Thinking with Video: Video generation as a promising multimodal reasoning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41121–41129, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Tong_Thinking_with_Video_Video_Generation_as_a_Promising_Multimodal_Reasoning_CVPR_2026_paper.html.

Azmine Toushik Wasi, Wahid Faisal, Abdur Rahman, et al. SpatiaLab: Can vision-language models perform spatial reasoning in the wild?, 2026. URL https://arxiv.org/abs/2602.03916.

World Labs. Marble: A multimodal world model. World Labs Research and Insights, 2025a. URL https://www.worldlabs.ai/blog/marble-world-model.

World Labs. RTFM: A real-time frame model. World Labs Research and Insights, 2025b. URL https://www.worldlabs.ai/blog/rtfm.

Sihan Yang, Runsen Xu, Yiman Xie, et al. MMSI-Bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2505.23764v3.

Yifan Yang et al. SkillOpt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026b. URL https://arxiv.org/abs/2605.23904.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. MindJourney: Test-time scaling with world models for spatial reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/file/9e3b203e72c4e058de26d02a92a81844-Paper-Conference.pdf.

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/2602.08236v2. Version 2.

Hanrong Zhang, Shicheng Fan, Henry Peng Zou, Yankai Chen, Zhenting Wang, Jiayu Zhou, Chengze Li, Wei-Chieh Huang, Yifei Yao, Kening Zheng, Xue Liu, Xiaoxiao Li, and Philip S. Yu. CoEvoSkills: Self-Evolving Agent Skills via Co-Evolutionary Verification. *arXiv preprint arXiv:2604.01687*, 2026a. URL https://arxiv.org/abs/2604.01687v3. Version 3.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025. URL https://arxiv.org/abs/2505.22954.

Wanyue Zhang, Wenxiang Wu, Wang Xu, et al. World2VLM: Distilling world model imagination into VLMs for dynamic spatial reasoning, 2026b. URL https://arxiv.org/abs/2604.26934.

Yuyou Zhang, Radu Corcodel, Chiori Hori, et al. SpinBench: Perspective and rotation as a lens on spatial reasoning in VLMs, 2026c. URL https://arxiv.org/abs/2509.25390.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM Agents Are Experiential Learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://doi.org/10.1609/aaai.v38i17.29936.

Tailin Zhou. Hierarchical self-improvement: A framework for task-specific evolvable agent harnesses, 2026. URL https://arxiv.org/abs/2608.08466.

A Appendix

A.1 Evaluation Data

Table 3 summarizes the development and evaluation sets. We used the 120 development examples to evolve the reusable components of our method. We adapted the five evolving baselines to the same visual inputs and scoring interface, then evaluated the selected variants on benchmarks excluding these development examples.

Table 3: Development examples and evaluation items.

	MMSI	**SAT**	**Spin**	**SpatiaLab**		**ERQA**	**CausalSpatial**	**PerceptionBench**
				MCQ	**Open**			
Development	30	0	30	0	0	0	30	30
Evaluation	970	300	2,709	1,400	1,400	400	1,395	2,970

SAT evaluates 150 questions under two option permutations, yielding 300 items. SpatiaLab MCQ and Open use the same 1,400 questions in two answer formats.

Baseline configurations. The evolving-agent comparison uses HSI (Zhou, 2026), ReasoningBank (Ouyang et al., 2026), Meta-Harness (Lee et al., 2026), ExpeL (Zhao et al., 2024), and GEPA (Agrawal et al., 2026), all adapted to GPT-5.6 with the same visual inputs and scoring interface. They operate directly on the original images and evolve prompts, memories, or code; none uses VWM-driven tools. Our harness co-evolves VWM-driven tools, skills, and execution code. The direct GPT-5.6 system answers from the original observations and is the reference baseline.

For the published comparison rows, Qwen3-VL uses the 8B-Instruct model; MiMo-Embodied and RoboBrain2.0 use their 7B and 32B models, respectively. Spatial-Interactor uses Qwen3-VL-8B with SFT and on-policy distillation. SandboxVLM uses GPT-5-mini; AVIC-R uses a Qwen2.5-VL-7B policy with GPT-4.1 for question answering. MMSI category scores for Qwen3-VL and

Table 4: Main mechanisms of the evolving-agent baselines.

Method	Main mechanism
HSI	Hierarchically revises task-specific agent harnesses.
ReasoningBank	Builds and retrieves reusable reasoning memories from agent experience.
Meta-Harness	Proposes and evaluates changes to executable harness code.
ExpeL	Distills insights from task experience and retrieves examples for later tasks.
GEPA	Reflectively revises prompts using execution and evaluation feedback.

RoboBrain2.0 are computed by weighting reported subcategory scores by sample count (Cai et al., 2026; Yang et al., 2026a). MiMo-Embodied's MMSI score and RoboBrain2.0's ERQA score are reported by Li et al. (2026). Avg. is reported for methods evaluated on all benchmarks and weighted by the evaluation sizes in Table 3.

A.2 Scoring and Usage Statistics

We computed accuracy over all items in each benchmark or category. We normalized multiple-choice answers before comparing them with the correct option. For PerceptionBench, we used deterministic option matching when an answer maps to a supplied choice; otherwise, a GPT-5.6 judge evaluates correctness under the benchmark's general-QA criterion. The same judge evaluates semantic equivalence for open-ended SpatiaLab questions. The judge operates in an evaluation context fully independent of the task agent.

Recovery measures accuracy on the subset of test items answered incorrectly by the GPT-5.6 direct reasoning baseline. When combining SpatiaLab MCQ and Open, we aggregated correct answers and item counts across both formats.

Skill-loading rate is the percentage of items on which the corresponding skill is loaded. Tool-use rate is the percentage of items with a VWM-driven tool call for the corresponding visual operation. An item can contribute to multiple columns; the Any column counts items with at least one skill load or tool call in the corresponding group.

A.3 Agent and Generation Configuration

The task agent, co-evolution supervisor, and external answer scorer each use the gpt-5.6-luna deployment in separate contexts. The GPT-5.6 direct reasoning baseline answers each item from the original observations in one stage. All five evolving baselines use this GPT-5.6 deployment. Table 5 lists agent and video generation settings.

The agent carries its interaction history through stages 1–4, following the order in Section 3.2. The verifier receives the generated frames. The "Ours (w/o VWM)" variant retains the VWM-driven skills and four-stage execution while disabling all VWM-driven tools and the verifier.

We used the MiniMax H3 Max Turbo image-to-video endpoint on FAL. A request supplies one source image or two images defining the start and end frames. Generation requests describe the visual transformation without prescribing the task answer.

Table 5: Agent and video generation settings.

Task agent		**Video generation**	
Reasoning effort	Medium	Duration	5 seconds
Temperature	0	Resolution	480P
Max. output tokens/model call	4,096	Prompt expansion	Balanced
Model-turn limits (stages 1–4)	4 / 6 / 6 / 6	Frames for agent reasoning	9

We used $R = 3$ evolution rounds. At each stage, the supervisor evaluates the initial candidate and up to $B = 5$ revisions, with early stopping. Execution-code revisions are restricted to the current stage.

For transfer to Grok-4.5, we initialized the harness from the completed GPT-5.6 co-evolution, replaced the VLM with Grok-4.5, and ran one additional co-evolution round before evaluation.

A.4 VWM-Driven Skills and Visual Operations

Table 6 summarizes the three skills and four visual operation categories, using the abbreviations in Figure 3b.

Table 6: Skills and visual operations used in the experiments.

Component	Abbr.	Function
VWM-driven skills		
Cross-view	CV	Guides reasoning about identity, viewpoint, and layout relations across images.
Temporal	Temp	Guides reasoning about transitions between ordered observations and about short temporal continuations.
Causal intervention	Caus	Guides reasoning about the outcome of a specified action on a visible target.
VWM-driven visual operations		
View change	View	Generates a new or connecting viewpoint to expose a spatial relation.
Temporal interpolation	Intp	Generates intermediate states between ordered observations.
Sequence continuation	Cont	Extends an observed state forward or backward in time.
Intervention	Act	Generates the predicted outcome of a specified action on a visible target.

A.5 Stage Assessment and Candidate Selection

Within each evolution round, the supervisor selects the best completed candidate at each stage using stage-specific assessments, sampled execution traces, and available answer-correctness feedback.

During capability boundary assessment, the supervisor combines skill-loading rates with sampled load/skip decisions to assess skill loading, skill choice, and whether the VWM can provide the needed evidence. During Request VWM Assistance, it checks the validity of each request, its source-image selection and order, and whether the tool performs the intended operation. During Video Interpretation, it assesses the interpretation of visible events, generation artifacts, and reference frames. During Evidence Integration and Reasoning, it considers answer correctness and the use of original and generated evidence.

The supervisor compares sampled execution traces across benchmarks, including execution failures and cases with and without generation, to weigh improvements against regressions across benchmarks.

A.6 Development and Test Performance

Figure 6 compares the final evolved agents on the 120 development examples and 8,044 held-out test examples from MMSI, Spin, CausalSpatial, and PerceptionBench. Aggregate accuracy is weighted by the number of examples within each split.

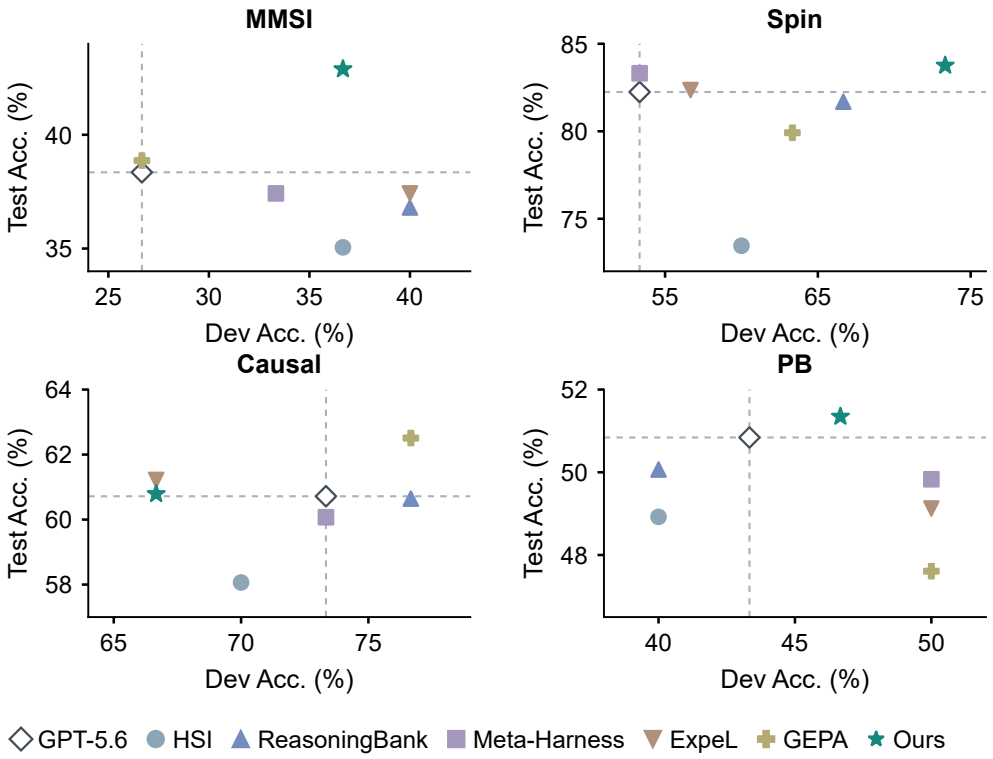


Figure 6: Development and test accuracy. Dashed lines show the GPT-5.6 direct reasoning baseline; stars mark CoWorld Reasoner.

ReasoningBank improves over GPT-5.6 on MMSI, Spin, and CausalSpatial during development, yet falls below it on all three test sets. Across the four benchmarks, CoWorld Reasoner and ReasoningBank both reach an aggregate development accuracy of 55.83%, but their test accuracies are 62.88% and 60.95%, respectively, compared with 61.62% for GPT-5.6. CoWorld Reasoner improves over GPT-5.6 on each of the four test benchmarks.

A.7 Results by Task Category

Tables 7–12 report category accuracies for Spin, SpatiaLab, ERQA, CausalSpatial, and Perception-Bench. All denotes overall benchmark accuracy, and Gain (pp) is the difference from GPT-5.6 in percentage points. Best results are bolded. Category results for MMSI and SAT appear in Table 2.

A.8 Examples of Evolved Guidance

Binding requests to their visual sources. The request-construction instructions align image references with the source order supplied to the VWM. In Figure 7, the second input image becomes Picture 1 in the generation request. They also ground object-relative movement in visible front and rear features. The interpretation instructions guide the agent to check object identity and express spatial relations in the reference frame specified by the question.

Distinguishing view changes from object motion. The interpretation instructions guide the agent and verifier to compare the subject with background landmarks across early, middle, and late frames. A turning subject against a fixed background can indicate object rotation rather than camera motion. In Figure 10, the agent tracks the SUV relative to the stationary chair, plant, and vase, then relates its motion to the original image and question.

Table 7: Spin accuracy by task category (%, ↑).

Method	Identity	Relation	Translation	Rotation	View	Mental rot.	Perspective	All
Evolving Agents								
GPT-5.6	98.51	88.68	**100.00**	**90.65**	67.17	66.97	69.00	82.24
+ *HSI*	99.01	64.78	94.87	82.72	66.26	64.68	61.83	73.46
+ *ReasoningBank*	**99.26**	84.43	**100.00**	**90.65**	68.09	72.48	68.03	81.69
+ *Meta-Harness*	98.76	88.52	98.72	89.24	65.65	**76.61**	**72.27**	83.31
+ *ExpeL*	98.76	87.74	**100.00**	86.12	66.87	71.56	71.45	82.36
+ *GEPA*	98.51	78.14	98.08	88.10	65.96	72.48	70.31	79.92
+ Ours	99.01	**91.35**	**100.00**	87.25	**71.73**	69.72	71.13	**83.76**
Gain (pp)	+0.50	+2.67	+0.00	-3.40	+4.56	+2.75	+2.12	+1.51

Table 8: SpatiaLab (MCQ) accuracy by task category (%, ↑).

Method	3D geo.	Depth/occl.	Orient.	Rel. pos.	Size/scale	Navigation	All
Evolving Agents							
GPT-5.6	47.90	55.21	60.40	**67.92**	53.17	57.38	56.64
+ *HSI*	47.48	57.53	60.40	64.62	50.40	55.70	55.71
+ *ReasoningBank*	47.48	59.46	60.40	63.68	**54.37**	55.27	56.57
+ *Meta-Harness*	47.06	57.53	59.41	66.04	49.60	57.38	55.86
+ *ExpeL*	49.58	**59.85**	60.89	67.45	52.78	51.05	56.64
+ *GEPA*	49.58	55.98	59.90	61.32	46.83	57.38	54.86
+ Ours	**52.10**	57.14	**64.36**	67.45	48.02	**57.81**	**57.36**
Gain (pp)	+4.20	+1.93	+3.96	-0.47	-5.16	+0.42	+0.71

Table 9: SpatiaLab (Open) accuracy by task category (%, ↑).

Method	3D geo.	Depth/occl.	Orient.	Rel. pos.	Size/scale	Navigation	All
Evolving Agents							
GPT-5.6	34.87	28.19	36.14	**41.51**	40.48	37.13	36.21
+ *HSI*	26.89	23.55	30.69	37.26	30.56	29.96	29.57
+ *ReasoningBank*	27.73	23.55	32.18	33.96	34.13	31.22	30.29
+ *Meta-Harness*	28.57	23.55	33.17	37.74	32.94	29.11	30.57
+ *ExpeL*	28.15	22.39	29.70	34.43	30.95	32.91	29.57
+ *GEPA*	25.63	23.55	32.67	31.60	27.38	29.96	28.21
+ Ours	**37.82**	**30.12**	**38.12**	41.04	**40.87**	**38.82**	**37.64**
Gain (pp)	+2.94	+1.93	+1.98	-0.47	+0.40	+1.69	+1.43

Figures 7–12 show reasoning trajectories for six examples. In Figure 9, the generated sequence shows the remaining plates staying in place after the bottom plate is removed, prompting the agent to revise its initial answer. Boldface highlights key cues, and `[...]` indicates omitted skill text.

Table 10: ERQA accuracy by task category (%, ↑).

Method	Action	Multi-view	Other	Pointing	Spatial	State	Task	Trajectory	All
Evolving Agents									
GPT-5.6	45.83	35.14	21.43	79.41	57.14	47.27	**65.79**	48.48	51.75
+ *HSI*	54.17	40.54	42.86	79.41	59.52	56.36	**65.79**	54.55	57.25
+ *ReasoningBank*	48.61	**48.65**	**57.14**	79.41	60.71	65.45	63.16	56.06	59.00
+ *Meta-Harness*	52.78	43.24	**57.14**	**85.29**	57.14	**70.91**	**65.79**	56.06	**60.00**
+ *ExpeL*	50.00	43.24	42.86	79.41	59.52	63.64	60.53	53.03	57.00
+ *GEPA*	54.17	**48.65**	42.86	73.53	**63.10**	63.64	57.89	**62.12**	59.75
+ Ours	**56.94**	**48.65**	35.71	**85.29**	58.33	52.73	60.53	60.61	58.50
Gain (pp)	+11.11	+13.51	+14.29	+5.88	+1.19	+5.45	-5.26	+12.12	+6.75

Table 11: CausalSpatial accuracy by task category (%, ↑).

Method	Collision	Compat.	Occlusion	Physics	All
Evolving Agents					
GPT-5.6	54.92	**64.58**	57.74	76.62	60.72
+ *HSI*	52.13	60.42	51.79	76.62	58.06
+ *ReasoningBank*	53.95	**64.58**	58.93	78.25	60.65
+ *Meta-Harness*	54.68	59.38	58.93	75.32	60.07
+ *ExpeL*	53.83	**64.58**	63.10	78.90	61.22
+ *GEPA*	**55.65**	59.38	**64.29**	80.84	**62.51**
+ Ours	54.56	62.50	52.98	**81.17**	60.79
Gain (pp)	-0.36	-2.08	-4.76	+4.55	+0.07

Table 12: PerceptionBench accuracy by error category (%, ↑).

Method	Context	Depth/ 3D	Fine grained	Halluc.	OCR	Attr.	Comp.	Count	Loc.	Rel.	All
Evolving Agents											
GPT-5.6	46.03	44.95	**48.60**	26.49	50.98	53.37	56.36	**49.24**	65.44	**62.08**	50.84
+ *HSI*	41.67	43.12	46.15	27.24	49.02	50.00	51.64	**49.24**	**66.06**	59.63	48.92
+ *ReasoningBank*	47.22	42.20	45.10	29.10	**52.16**	54.29	**57.82**	47.40	63.30	58.72	50.07
+ *Meta-Harness*	44.84	43.73	46.15	26.49	51.76	53.68	57.45	47.40	62.08	60.55	49.83
+ *ExpeL*	44.05	40.98	46.85	29.10	**52.16**	53.99	52.36	45.57	63.00	59.33	49.12
+ *GEPA*	42.46	41.59	42.31	22.39	50.59	50.92	54.55	48.93	60.86	56.88	47.61
+ Ours	**48.81**	**46.79**	46.50	**30.97**	49.41	**57.36**	54.18	48.62	64.53	61.47	**51.35**
Gain (pp)	+2.78	+1.83	-2.10	+4.48	-1.57	+3.99	-2.18	-0.61	-0.92	-0.61	+0.51

A.9 Qualitative Case Studies

Input

Question:
These three photos were taken in the same square meeting room. Assuming that the camera in Photo 3 is facing east, if I am currently taking Photo 2, then what is the direction of the camera relative to me when taking Photo 1?
Options: A: Approximately east, B: Approximately west, **C: Approximately north**, D: Approximately south

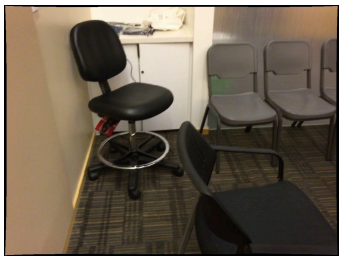


Photo 1


Photo 2

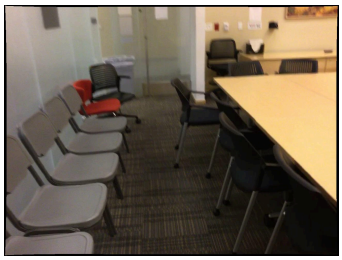


Photo 3

P1 · Capability Boundary Assessment

<tool_call>

{
"name": "load_wm_task_skill",
"arguments": {
"skill_id": "**complete-cross-view-spatial**"
}
}

</tool_call>

<tool_response>

{
"skill_body": "[...]\n- Use `vwm_new_view_tool` for a missing grounded viewpoint or cross-view mapping.\n[...]"
}

</tool_response>

{
"decision_reason": "The three images provide separate viewpoints rather than a single jointly visible map. Using the shared table, chairs, desk, and wall layout, Photo 1 is inferred to be on the south side of Photo 2.",
"missing_state": "A connecting, reference-frame-preserving view that maps the Photo 1 camera position to the Photo 2 position and establishes the room's north/south side.",
"expected_discriminative_value": "A grounded connecting view of the room could distinguish whether the Photo 1 viewpoint is on the north or south side of the Photo 2 viewpoint by preserving the table, chair rows, and east-wall desk as anchors.",
"provisional_answer": "**D: Approximately south**",
"selected_task_skill": "**complete-cross-view-spatial**"
}

Figure 7: Reasoning trajectory for `MMSI-Bench/MMSI-0071`.

P2 · Request VWM Assistance

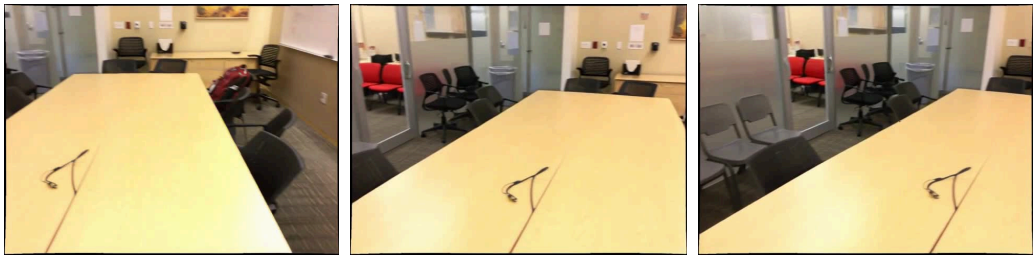
<tool_call>


```
{
  "name": "vwm_new_view_tool",
  "arguments": {
    "generation_instruction": "Local Picture 1 shows a large light meeting table filling the foreground, a rear desk with wall controls beyond it, office chairs around the table, and a red backpack near the right-side chairs. Local Picture 2 shows gray chairs along one wall, red and black chairs near glass doors, and the meeting table along the right. Render one coherent same-room connecting view during the repositioning, retaining the identities and layout of the meeting table, rear desk, carpet, walls, glass doors, chair rows, and red backpack wherever visible. Show enough of the intervening room layout to map the two viewpoints without adding new objects, labels, or an assumed answer direction.",
    "camera_instruction": "Starting from the viewpoint shown in local Picture 1, make a grounded lateral repositioning and modest reorientation through the same square meeting room toward the complementary room layout represented by local Picture 2. Keep the path inside the room and preserve stable landmarks. Do not add compass labels or encode any preferred cardinal direction. Expose the spatial arrangement between the two viewpoints using the meeting table, chair rows, glass doors, and rear desk as anchors.",
    "source_image_indices": [1, 0],
    "source_order_mode": "reference_to_anchor",
    "reference_image_index": 1
  }
}
```


</tool_call>

Local Picture 1 = input image 2; Local Picture 2 = input image 1.

<tool_response>



predicted_00_00 predicted_00_02 predicted_00_03

predicted_00_04 predicted_00_06 predicted_00_08

</tool_response>

Figure 7 (continued).

P3 · Video Interpretation

<tool_call>

```
{
  "name": "viewpoint_layout_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Map the camera viewpoint of Photo 1 relative to the photographer's viewpoint in Photo 2, using Photo 3's stipulated east-facing orientation and the persistent meeting-room landmarks."
  }
}
```

</tool_call>

<tool_response>

```
{
  "request_execution": {
    "finding": "The requested connecting view is partially executed as a leftward camera sweep with modest reorientation, while clear lateral camera translation is not distinguishable from rotation.",
    "sources": ["predicted_00_00", "predicted_00_03", "predicted_00_08"],
    "uncertainty": "The supplied frames do not establish the exact camera path or displacement."
  },
  "defects": []
}
```

</tool_response>


```
{
  "observations": [
    "Across generated frames, the meeting table remains consistent while the rear desk and red backpack leave view and the glass doors, red/black chairs, and then gray-chair row become prominent."
  ],
  "defects_and_uncertainty": [
    "The generated sequence preserves the room and furniture identities, but it mainly shows a leftward rotational sweep; exact camera translation and displacement are not clearly established."
  ],
  "reference_frame_mapping": "Keep Photo 3's stipulated heading fixed: image-left when facing east is north, and image-right is south. Use the gray-chair row and cabinet as persistent anchors linking Photo 1 to Photo 3, and the table/rear-desk view as the Photo 2 anchor. This maps the Photo 1 viewpoint toward the north side relative to the Photo 2 viewpoint, subject to the generated sequence's uncertainty about exact translation. No final option is selected in this stage."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "Photo 3 is stipulated to face east. In that view, the distinctive row of gray chairs is on the image-left side, which converts to the north side. Photo 1 is positioned by that same gray-chair row and cabinet, while Photo 2 is the table/rear-desk viewpoint farther from that north-side row. The strongest contrary cue is the earlier provisional south reading, but it incorrectly assigned the chair-wall side without first converting Photo 3's east-facing image axes."
}
```


Answer

C: Approximately north

Figure 7 (continued).

Input

Question:
From a stationary viewpoint, in which direction did the object rotate between these two views? A. clockwise, **B. counterclockwise**
Only answer with the capital letter from (A, B).
The camera is stationary and the object rotate in place.
Clockwise and counterclockwise are defined from a top-down view.



Image 1 Image 2

P1 · Capability Boundary Assessment

<tool_call>

```
{
  "name": "load_wm_task_skill",
  "arguments": {
    "skill_id": "reason-temporal-spatial-continuity"
  }
}
```

</tool_call>

<tool_response>

```
{
  "skill_body": "[...]\nUse `vwm_temporal_interpolation_tool` for the shortest ordered interval containing the missing\nstate. Choose endpoints that constrain the relevant identities and motion. The result is predicted\nvisual continuity, not an exact geometry, distance, speed, or angle measurement.\n[...]"
}
```

</tool_response>


```
{
  "decision_reason": "The two images show ordered endpoint orientations, but not the intermediate rotation. A connecting transition could distinguish clockwise from counterclockwise while preserving the table's stable legs and tabletop geometry.",
  "missing_state": "The ordered intermediate rotation of the table between the front-facing first view and the angled second view, including the signed direction of its turn.",
  "expected_discriminative_value": "A temporally interpolated transition would be useful if it preserves the stationary camera and object identity, revealing which edge swings toward the camera during the rotation.",
  "provisional_answer": "A",
  "selected_task_skill": "reason-temporal-spatial-continuity"
}
```


Figure 8: Reasoning trajectory for `Spin-Bench/SPIN-2511`.

P2 · Request VWM Assistance

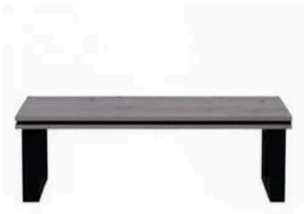
<tool_call>

```
{
  "name": "vwm_temporal_interpolation_tool",
  "arguments": {
    "generation_instruction": "Create a short, visually continuous transition from local Picture 1 to local Picture 2. Begin with the rectangular dark wooden table in the front-facing orientation shown in Picture 1, then show the same table rotating in place while the camera remains completely stationary, and end in the angled orientation shown in Picture 2. Keep the tabletop, four legs, and their visible spacing and silhouettes coherent throughout; show the tabletop edges and leg positions changing smoothly so the actual turn between the two endpoint orientations is visible. Do not translate the table or add any objects.",
    "source_image_indices": [0, 1],
    "source_order_mode": "ordered_transition",
    "reference_image_index": null
  }
}
```

</tool_call>

Local Picture 1 = input image 1; Local Picture 2 = input image 2.

<tool_response>



predicted_00_00



predicted_00_02



predicted_00_04



predicted_00_08

</tool_response>

Figure 8 (continued).

P3 · Video Interpretation

<tool_call>

```
{
  "name": "temporal_motion_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Inspect the generated transition of the same rectangular table from the first front-facing source view to the second angled source view. Assess whether the table rotates rigidly in place with a stationary camera, and describe the visible direction of its top-down turn using persistent tabletop and leg anchors, without choosing an answer."
  }
}
```

</tool_call>

<tool_response>

```
{
  "request_execution": {
    "finding": "The table visibly turns in place as its tabletop orientation and visible leg spacing change progressively, with no evident translation or added object.",
    "sources": ["predicted_00_00", "predicted_00_04", "predicted_00_08"],
    "uncertainty": "The supplied frames support apparent in-place rotation, but the original source images are not attached for comparison."
  },
  "defects": []
}
```

</tool_response>


```
{
  "observations": [
    "The right-side legs progressively separate in depth, while the left-side structure exposes more side depth, matching the endpoint's angled presentation.",
    "The artifact shows apparent rigid in-place rotation without obvious translation or deformation."
  ],
  "defects_and_uncertainty": [
    "The white background supplies no independent stationary landmarks, though the framing remains stable."
  ],
  "reference_frame_mapping": "The requested convention is clockwise/counterclockwise as viewed from directly above the table, not the viewer-relative screen direction. The generated sequence supplies the visible 3D turn cues needed for that mapping, but the verifier correctly leaves the final top-down label for interpretation against the original ordered views."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "The second view shows the table's left end closer to the stationary camera and its right end farther away, with the long front edge rising toward the right. This is the cue opposite to a clockwise turn from the initial front-facing orientation."
}
```


Answer

B

Figure 8 (continued).

Input

Question:
If you take away the bottom plate from the stack, what is likely to happen?
Choose exactly one option:
A. The other plates will fall
B. The plates will float
C. Nothing will happen
D. All plates will break



Image 1

P1 · Capability Boundary Assessment

<tool_call>


```
{
  "name": "load_wm_task_skill",
  "arguments": {
    "skill_id": "simulate-causal-intervention"
  }
}
```


</tool_call>

<tool_response>


```
{
  "skill_body": "[...]\nUse `vwm_simulate_intervention_tool` for one target and one action supported by one source\nobservation. The structured controls and the generation request must describe the same target and\naction. Do not add a second moving target or an unrelated scene edit.\n[...]"
}
```


</tool_response>


```
{
  "decision_reason": "The original image shows a stack of plates, but not the result of removing the bottom plate. The outcome therefore depends on an unobserved bounded intervention.",
  "missing_state": "The post-removal state and short motion of the remaining plates, including whether they lose support and fall.",
  "expected_discriminative_value": "A causal-intervention simulation could show whether the remaining plates move downward/tip after the bottom plate is removed, distinguishing falling from staying in place.",
  "provisional_answer": "A. The other plates will fall",
  "selected_task_skill": "simulate-causal-intervention"
}
```


Figure 9: Reasoning trajectory for `SpatiaLab/SPATIALAB-590`.

P2 · Request VWM Assistance

<tool_call>

```
{
  "name": "vwm_simulate_intervention_tool",
  "arguments": {
    "generation_instruction": "Show the visible stack of white plates on the lower-left shelf in Picture 1. A hand carefully takes the lowest plate out from underneath the stack, sliding it fully clear while leaving the other plates initially together. Keep the shelf, the remaining plates, and the surrounding toiletries and wooden-handled object visible and otherwise unchanged. Show the immediate short progression after the lowest plate has been removed, without adding any other action or scene edit.",
    "target_object": "The lowest white plate in the stack at the lower-left of the shelf, visibly supporting the plates above it.",
    "source_image_indices": [0]
  }
}
```

</tool_call>

Local Picture 1 = input image 1.

<tool_response>



predicted_00_00



predicted_00_01



predicted_00_03



predicted_00_04



predicted_00_05



predicted_00_08

</tool_response>

Figure 9 (continued).

P3 · Video Interpretation

<tool_call>

```
{
  "name": "intervention_outcome_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Inspect whether removing the lowest white plate from the identified stack causes the remaining plates to lose support and visibly fall, while checking target identity, contact, and continuity against the original shelf scene."
  }
}
```

</tool_call>

<tool_response>

```
{
  "request_execution": {
    "finding": "The requested removal is depicted: a hand slides a white plate out from the stack and the plate is clear of the stack by the later frames, with no other scene edit visible.",
    "sources": ["predicted_00_01", "predicted_00_04", "predicted_00_05", "predicted_00_06"],
    "uncertainty": ""
  },
  "defects": []
}
```

</tool_response>


```
{
  "observations": [
    "Early frames show the hand approaching and beginning to pull the lowest plate outward.",
    "Middle frames show the removed plate moving toward the foreground and becoming clear of the stack.",
    "Late frames show the remaining plates still together in approximately the same shelf position, with no visible fall, float, or breakage."
  ],
  "defects_and_uncertainty": [
    "The simulation depicts removal clearly, but its mechanics are approximate and cannot guarantee the real physical result."
  ],
  "reference_frame_mapping": "The relevant frame is the original shelf scene viewed from the camera-facing perspective. The intervention is evaluated relative to the shelf and the persistent plate stack: the bottom plate is removed horizontally toward the foreground, and the remaining plates are tracked in the same lower-left shelf region. Their lack of visible displacement after removal is the task-relevant generated relation."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "The generated sequence shows the lowest plate being slid out toward the foreground and removed, while the remaining plates stay grouped in the same shelf position. They do not visibly tip, separate, fall, float, or break. The strongest contrary possibility is a physical settling or collapse after support is removed, but no such movement appears in the predicted endpoint frames."
}
```


Answer

C. Nothing will happen

Figure 9 (continued).

Input

Question:
A toy armed SUV with a rugged and tactical appearance lies on the floor. The toy car is driving forward. Is there any risk if it moves forward? A. Yes; **B. No**; C. Not Sure.



Image 1

P1 · Capability Boundary Assessment

{
"decision_reason": "The SUV's front points toward the open floor on the right, with no immediate obstacle in that direction. The plant and vase are offset from the apparent path, so the static scene supports no evident risk.",
"missing_state": null,
"expected_discriminative_value": "No additional visual-world operation would add meaningful evidence; the relevant path and nearby obstacles are already visible together.",
"provisional_answer": "**B. No**",
"selected_task_skill": null
}

P2 · Request VWM Assistance

<tool_call>
{
"name": "vwm_simulate_intervention_tool",
"arguments": {
"generation_instruction": "Keep the camera fixed and keep the entire visible floor scene stable. The blue-and-orange rugged toy SUV starts in its current position and drives straight forward in the direction indicated by its visible front: the **bright white headlight at the right-facing end**, opposite the red rear lights. Let it continue a moderate distance along that heading across the wooden floor, preserving the potted plant and gray vase as fixed objects and showing any visible approach, contact, or clearance **without inventing an outcome in advance**.",
"target_object": "The blue-and-orange rugged toy SUV on the wooden floor near the upper-left/central area, identifiable by its bright white headlight on the right-facing front and red lights at the rear",
"source_image_indices": [0]
}
}
</tool_call>

Figure 10: Reasoning trajectory for `CausalSpatial/CS-COLL-6169db902026`.

P2 · Request VWM Assistance (continued)

Local Picture 1 = input image 1.

<tool_response>



predicted_00_00



predicted_00_02



predicted_00_03



predicted_00_04



predicted_00_06



predicted_00_08

</tool_response>

P3 · Video Interpretation

<tool_call>


```
{
  "name": "intervention_outcome_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Assess whether the identified toy SUV's forward motion, using its intrinsic front
    marked by the bright headlights (right-facing end in the source), creates a visible collision risk with
    the potted plant or any other floor obstacle."
  }
}
```


</tool_call>

Figure 10 (continued).

P3 · Video Interpretation (continued)

<tool_response>


```
{
  "request_execution": {
    "finding": "The identified SUV moves continuously along a roughly straight path with its bright
    headlights leading, while the camera viewpoint and floor scene remain stable.,",
    "sources": ["predicted_00_00", "predicted_00_04", "predicted_00_08"],
    "uncertainty": "The path is judged from image-plane motion; exact 3D heading cannot be
    confirmed."
  },
  "defects": []
}
```


</tool_response>


```
{
  "observations": [
    "Generated frames: the same SUV advances while the chair, wall/door, plant, vase, and floor remain
    stable, supporting object motion rather than camera movement.",
    "The SUV comes near the potted plant around the middle frames but visibly continues past it; the
    plant does not visibly move.",
    "The SUV continues along the right side and does not visibly contact the gray vase."
  ],
  "defects_and_uncertainty": [
    "The SUV passes close to the plant, so the apparent lack of contact depends on the generated path and
    may not establish a guaranteed real-world outcome."
  ],
  "reference_frame_mapping": "The task asks about the SUV's intrinsic forward direction, not merely
  viewer-relative left/right. The source's front is mapped to the headlight-bearing, right-facing end,
  opposite the red rear lights. In that reference frame, the generated trajectory is forward and passes beside
  the plant before continuing toward the lower-right; the vase is offset from that trajectory. The original
  scene and generated motion therefore support assessment of visible obstacle risk along this particular
  forward path, while leaving unshown or exact 3D hazards unresolved."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "The SUV's bright headlight marks its front, opposite the red rear lights.
  In the generated forward trajectory, it passes to the right of the potted plant without moving or
  contacting it, and continues past the gray vase with visible clearance. No obstacle obstructs the shown
  endpoint."
}
```


Answer

B. No

Figure 10 (continued).

Input

Question:
The image is a continuous first-person perspective shot. In which direction are you moving?
Options: A: Left, B: Up, **C: Cannot be determined**, D: Right

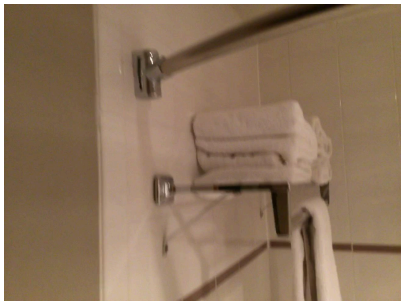


Image 1

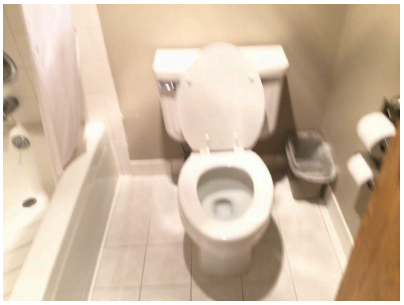


Image 2

P1 · Capability Boundary Assessment

<tool_call>

{
"name": "load_wm_task_skill",
"arguments": {
"skill_id": "**reason-temporal-spatial-continuity**"
}
}

</tool_call>

<tool_response>

{
"skill_body": "[...]\nUse `vwm_temporal_interpolation_tool` for the shortest ordered interval containing the missing\nstate. Choose endpoints that constrain the relevant identities and motion. The result is predicted\nvisual continuity, not an exact geometry, distance, speed, or angle measurement.\n[...]"
}

</tool_response>

{
"decision_reason": "The ordered endpoint images show different viewpoints, but do not directly reveal the intermediate camera path or separate camera translation from rotation and framing changes.",
"missing_state": "The intermediate first-person camera poses and the resulting continuous motion relative to stable anchors such as the toilet, wall, shower, and bin.",
"expected_discriminative_value": "An interpolated sequence preserving stable bathroom anchors could show whether the viewpoint progresses laterally left or right, if the transition is visually constrained.",
"provisional_answer": "**C: Cannot be determined**.",
"selected_task_skill": "**reason-temporal-spatial-continuity**"
}

Figure 11: Reasoning trajectory for `MMSI-Bench/MMSI-0821`.

P2 · Request VWM Assistance

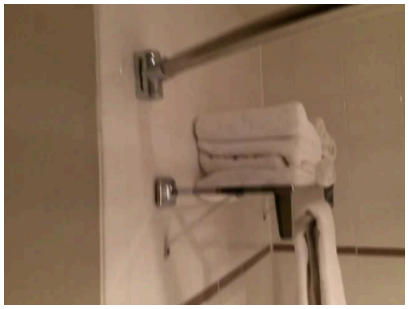
<tool_call>


```
{
  "name": "vwm_temporal_interpolation_tool",
  "arguments": {
    "generation_instruction": "Create a single plausible continuous first-person-perspective transition from Picture 1 to Picture 2. Begin with the close, tilted view of the bathroom wall and shower curtain, including the horizontal grab bar and folded white towel, then progress smoothly through the bathroom while keeping the same room and fixtures consistent. End at the wider view showing the toilet centered, the bathtub and shower curtain on the left, the small waste bin on the right, and the wooden door edge at the far right. Make the intermediate camera motion and changing relative positions of these fixed bathroom landmarks clearly visible, with natural handheld framing and no cuts.",
    "source_image_indices": [0, 1],
    "source_order_mode": "ordered_transition",
    "reference_image_index": null
  }
}
```

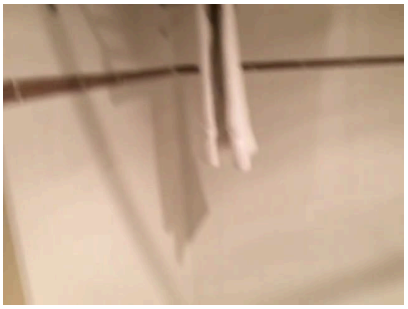

</tool_call>

Local Picture 1 = input image 1; Local Picture 2 = input image 2.

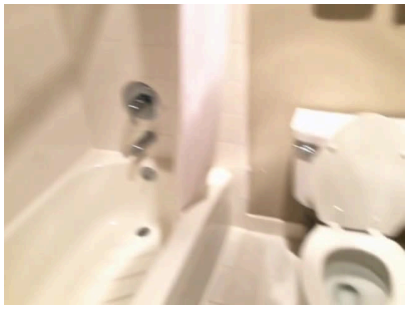
<tool_response>



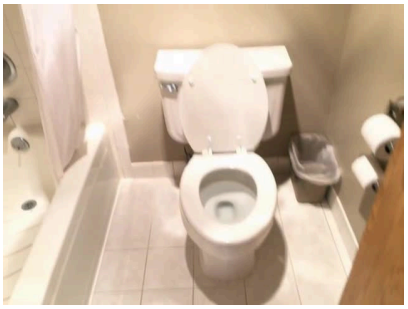
predicted_00_00



predicted_00_02



predicted_00_04



predicted_00_08

</tool_response>

Figure 11 (continued).

P3 · Video Interpretation

<tool_call>

```
{
  "name": "temporal_motion_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Determine the signed direction of the first-person camera movement from the close towel-rack view to the wider bathroom view, distinguishing lateral motion from camera rotation or forward/backward movement."
  }
}
```

</tool_call>

<tool_response>

```
{
  "request_execution": {
    "finding": "The requested continuous bathroom transition occurs without a visible cut, ending with the toilet centered, bathtub and curtain on the left, waste bin on the right, and wooden door edge at the far right.",
    "sources": ["predicted_00_00", "predicted_00_04", "predicted_00_06", "predicted_00_08"],
    "uncertainty": "The supplied frames are generated predictions, and the original endpoint images are not attached for comparison."
  },
  "defects": []
}
```

</tool_response>


```
{
  "observations": [
    "The towel rack and folded towels dominate the early generated frames, then move toward and out of the upper-left.",
    "The same white toilet appears in the middle-to-late frames and remains identifiable by its tank, raised lid, and bowl; the waste bin persists to its right."
  ],
  "defects_and_uncertainty": [
    "The generated transition is visually usable and continuous, but it does not separate lateral translation from camera rotation or forward/backward motion."
  ],
  "reference_frame_mapping": "The relevant reference frame is the first-person camera's actual direction of travel, not merely the direction in which the scene shifts on the screen and not merely the direction of view rotation. In the generated artifact, bathroom landmarks shift as the camera view turns toward the toilet area. Without established translation relative to at least two fixed anchors, that rightward screen reorientation cannot be converted reliably into intrinsic Left or Right motion; vertical motion is likewise not established."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "The only concrete transition cue is a rightward reorientation of the view: the towel rack leaves the upper-left while the bathtub and toilet come into view. However, the fixed anchors do not show whether this resulted from left/right translation, camera rotation, or a combination. The strongest contrary cue is the generated sequence's apparent rightward sweep, but it is predicted footage and does not isolate physical travel direction."
}
```


Answer

C: Cannot be determined

Figure 11 (continued).

Input

Question:
On the floor, there is a toy truck with a classic design. If the toy car is driving forward, is it true that removing the white vase with a bold, oval-like form will prevent all collisions? If removing the white vase with a bold, oval-like form does not prevent all collisions, what additional steps should be taken? A. Remove the pile of books; B. Remove the vase with a twisted loop; C. Remove the moon vase with gray colors; D. There is no collision; **E.Removing the white vase with a bold, oval-like form is sufficient**; F. Not Sure.



Image 1

P1 · Capability Boundary Assessment

<tool_call>
{
"name": "load_wm_task_skill",
"arguments": {
"skill_id": "**simulate-causal-intervention**"
}
}
</tool_call>

<tool_response>
{
"skill_body": "[...]\nUse `vwm_simulate_intervention_tool` for one target and one action supported by one source\nobservation. The structured controls and the generation request must describe the same target and\naction. Do not add a second moving target or an unrelated scene edit.\n[...]"
}
</tool_response>

{
"decision_reason": "The original image shows the truck, its apparent forward heading, and multiple floor objects, but it does not show the counterfactual result after removing the vase and driving forward. A bounded intervention simulation could distinguish whether the truck then reaches another obstacle.",
"missing_state": "The outcome of the toy truck driving straight forward after the white oval-shaped vase is removed, including whether its footprint contacts the gray vase or any other obstacle.",
"expected_discriminative_value": "A useful simulation would preserve the truck's heading and scene layout while removing only the white oval vase; observing whether the truck contacts the gray vase or another object would distinguish sufficiency from the need for an additional removal.",
"provisional_answer": "**F. Not Sure.**",
"selected_task_skill": "**simulate-causal-intervention**"
}

Figure 12: Reasoning trajectory for `CausalSpatial/CS-COLL-6a01e625175f`.

P2 · Request VWM Assistance

<tool_call>


```
{
  "name": "vwm_simulate_intervention_tool",
  "arguments": {
    "generation_instruction": "Starting from Picture 1, keep the fixed overhead camera and the entire room layout unchanged. Identify the classic toy truck near the upper-left floor area and the white oval-shaped vase with a large central opening positioned ahead of it. Remove only that white oval-shaped vase from the floor, leaving the gray mottled vase, furniture, television, and all other objects unchanged. Then show the toy truck traveling straight forward along its existing visible heading across the floor, with the truck's path and any resulting contact with remaining objects clearly visible. Do not remove or move any other object.",
    "target_object": "the white vase with a bold oval-like form and a large central opening on the floor ahead of the toy truck",
    "source_image_indices": [0]
  }
}
```


</tool_call>

Local Picture 1 = input image 1.

<tool_response>



predicted_00_00



predicted_00_01



predicted_00_03



predicted_00_04



predicted_00_05



predicted_00_08

</tool_response>

Figure 12 (continued).

P3 · Video Interpretation

<tool_call>

```
{
  "name": "intervention_outcome_verifier",
  "arguments": {
    "generation_index": 0,
    "question_focus": "Whether the classic toy truck, traveling forward along its original visible heading after only the white oval-shaped vase is removed, collides with the remaining gray mottled vase or any other object."
  }
}
```

</tool_call>

<tool_response>

```
{
  "request_execution": {
    "finding": "The white oval-shaped vase is present initially, absent from the next supplied frame onward, and the truck then travels forward along its visible grille-facing heading with the gray mottled vase left in place.",
    "sources": ["predicted_00_00", "predicted_00_01", "predicted_00_05"],
    "uncertainty": "The supplied frames support the requested removal and forward travel, but do not show the truck contacting any remaining object."
  },
  "defects": []
}
```

</tool_response>


```
{
  "observations": [
    "The white oval vase disappears between the first and second generated frames, while the gray mottled vase remains in place.",
    "By the middle frames, the truck is visibly to the left of the gray vase and there is no visible contact."
  ],
  "defects_and_uncertainty": [
    "The truck is partly or fully offscreen in the final frames, so the complete trajectory and any later collision are unresolved."
  ],
  "reference_frame_mapping": "The task uses the original overhead room view and asks about the toy truck driving forward, meaning forward relative to the truck's intrinsic front, identified by its grille and bumper—not merely downward on the screen. In that fixed viewer-relative frame, the simulated truck moves generally downward and slightly right while the gray vase stays fixed near center-right; the visible path remains left of the gray vase. This mapping supports a noncontact segment but does not establish the full forward route beyond the bottom edge."
}
```


P4 · Evidence Integration and Reasoning


```
{
  "discriminating_evidence": "After the white oval vase is removed, the gray mottled vase remains in place, but the truck's front grille and bumper lead it downward and to the left of that vase. The generated frames show no contact with the gray vase or another remaining obstacle before the truck leaves the visible floor area. Thus no additional removal, including the gray vase, is indicated."
}
```


Answer

E. Removing the white vase with a bold, oval-like form is sufficient.

Figure 12 (continued).