ASKING THE WORLD: GENERALIST PHYSICAL REASONING THROUGH AGENTIC WORLD MODELING AND PROBING

Anonymous authors
Paper under double-blind review

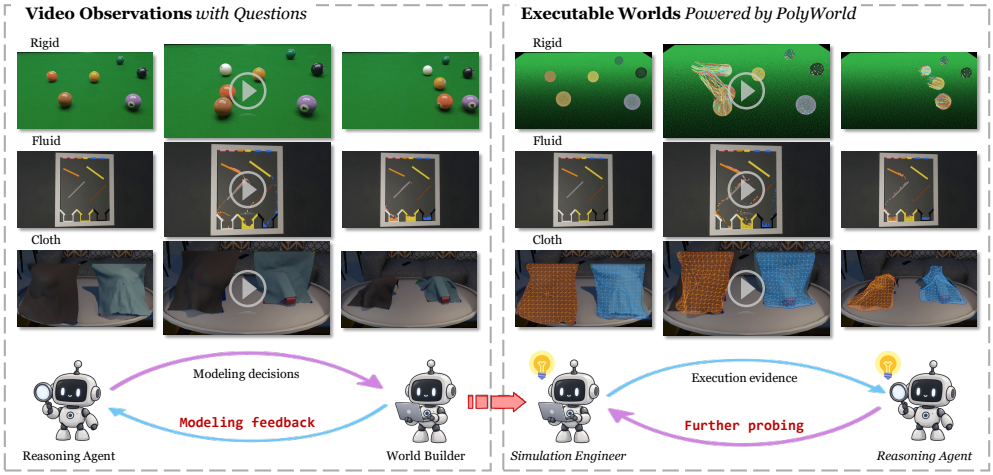

Figure 1: **Physical reasoning through world modeling and probing.** Given videos and questions, ATW builds and revises executable worlds, probes them with PolyWorld Engine, and reasons from execution evidence. Examples span rigid bodies, fluids, and cloth.

ABSTRACT

Physical reasoning from video requires inferring latent physical properties and dynamics beyond direct observation. Direct VLM inference remains unreliable on complex physical tasks without explicit modeling and validation, while predefined tool pipelines rely on task- and domain-specific priors that limit generalization across materials, dynamics, and reasoning tasks. We introduce *Asking the World* (ATW), a generalist agent that constructs and interrogates task-relevant executable worlds through two adaptive stages: *World Modeling* calibrates a world from video, while *World Probing* queries, simulates, and intervenes on it to obtain question-relevant evidence. Rather than prescribing the operations in either stage, ATW determines how to model and probe according to the scene and question. We develop *PolyWorld Engine*, a lightweight and highly programmable Warp-based multiphysics simulator for constructing and probing worlds with rigid bodies, soft bodies, cloth, ropes, fluids, and their coupled interactions. CEM-based system identification recovers task-relevant dynamics during World Modeling. The resulting world becomes an active workspace for question-directed physical experiments rather than a predetermined downstream tool. We evaluate ATW on CLEVRER, ContPhy, and three real-world scenarios. Using Gemini-3-Flash as its base VLM, ATW achieves 80.82% overall per-question accuracy on CLEVRER, improving direct Gemini-3-Flash by 46.50 points, GPT-5.5 by 13.58 points, and PhysMind by 8.27 points. On ContPhy, it reaches 70.56% overall accuracy, surpassing Gemini-3-Flash by 28.10 points and GPT-5.5 by 3.53 points. Across the three real-world scenarios, ATW achieves 71.67% accuracy, 28.33 points above GPT-5.5. These results establish agentic world modeling and probing as an effective, execution-grounded approach to generalist physical reasoning.

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

Physical questions about latent properties, future outcomes, and counterfactual events often require evidence beyond the observed video. A capable system must therefore construct a task-relevant account of scene dynamics and obtain physical evidence from it. Doing so across different materials, dynamics, and question types is the central ambition of generalist physical reasoning, from rigid-body collisions in CLEVRER to continuum phenomena in ContPhy (Yi et al., 2019; Zheng et al., 2024).

Existing approaches largely follow two paradigms. VLM-based methods exploit broad visual and semantic priors, either through direct inference or task-specific adaptation (Chow et al., 2025; Wang et al., 2025; Feng et al., 2025). Yet recent evaluations continue to expose failures in physical consistency and causal evolution across multimodal and generative video models (Li et al., 2025c; Bansal et al., 2024; Meng et al., 2024; Li et al., 2025a). Physics-grounded methods instead recover dynamics, call specialized tools, or execute simulators (Ding et al., 2021; Li et al., 2023; Xie et al., 2024; Zhang et al., 2024; Fan et al., 2025; Cherian et al., 2026; Yang et al., 2026). Although they provide explicit physical evidence, they usually commit in advance to a particular representation, physical model, or inference procedure, limiting adaptation across materials and questions. What remains missing is a general agent that can construct an appropriate executable world for the scene and determine how to interrogate it for the question at hand.

Recent agents show that tools, generated programs, and execution can support iterative reasoning rather than only terminal prediction (Yao et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023; Liang et al., 2023; Yin et al., 2026). Physical question answering, however, has no fully observed target scene against which every update can be compared; the agent must decide for itself what hidden mechanisms and interventions matter. We therefore introduce *Asking the World* (ATW), a generalist physical reasoning agent that organizes its reasoning into two stages. *World Modeling* constructs a task-relevant executable world by identifying relevant entities, selecting physical representations, and recovering latent dynamics. *World Probing* queries, simulates, and intervenes on that world to gather evidence for the associated questions. Rather than prescribing the operations in either stage, ATW determines how to model and probe according to the scene and question. In a nutshell, ATW asks the world by building and probing it (Figure 1). This process is enabled by *PolyWorld Engine*, our lightweight and highly programmable multiphysics simulator built on Warp (Macklin, 2022). It provides a unified environment for rigid bodies, soft bodies, cloth, ropes, fluids, and their coupled interactions, which ATW can instantiate and modify on demand. CEM-based system identification supports World Modeling by recovering task-relevant parameters from visual evidence. Together, these capabilities turn simulation from a predetermined downstream tool into an active workspace for constructing and interrogating physical explanations.

Without task-specific training, ATW applies the same agentic formulation across rigid-body, continuum, and real-world settings. It achieves 80.82% overall per-question accuracy on CLEVRER, improving Gemini-3-Flash by 46.50 points, GPT-5.5 by 13.58 points, and PhysMind by 8.27 points. On ContPhy, it reaches 70.56%, surpassing Gemini-3-Flash by 28.10 points and GPT-5.5 by 3.53 points; across three real-world scenarios, it achieves 71.67% accuracy, 28.33 points above GPT-5.5. These results establish agentic world modeling and probing, supported by a general multiphysics engine, as an effective path toward generalist physical reasoning.

In summary, our contributions are threefold:

- **Agentic physical reasoning.** We introduce ATW, a generalist physical reasoning agent that organizes reasoning into World Modeling, which constructs task-relevant executable worlds, and World Probing, which conducts question-directed queries, simulations, and interventions.
- **Generalist multiphysics execution.** We develop PolyWorld Engine, a lightweight and highly programmable environment that unifies rigid bodies, soft bodies, cloth, ropes, fluids, and their coupled interactions for on-demand world construction, intervention, and validation.
- **Broad empirical validation.** We demonstrate substantial improvements over direct VLM reasoning and strong baselines across CLEVRER, ContPhy, and three real-world scenarios, spanning multiple materials, dynamical systems, and reasoning tasks.

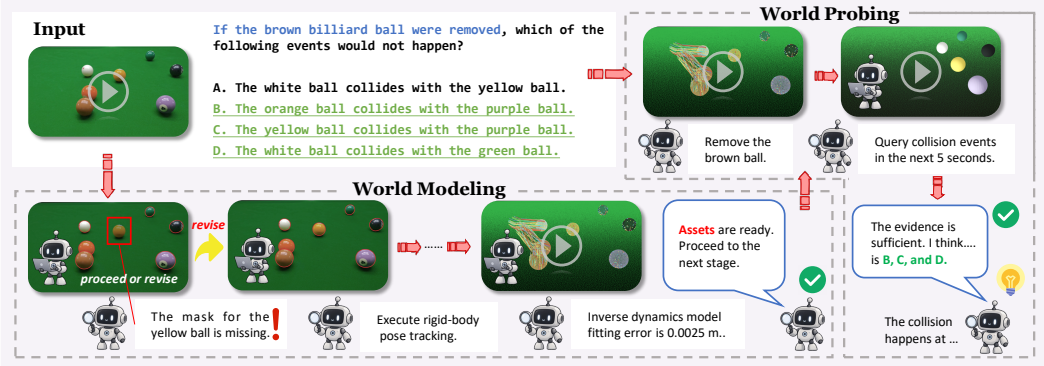

Figure 2: **ATW system overview.** A billiards counterfactual illustrates the one-way transition from World Modeling to World Probing. The modeling agent corrects a missing object mask, tracks object poses, and fits dynamics before committing the executable world. The probing agent then removes the brown ball, queries collision events, and answers from the resulting evidence.

2 Related Work

Visual Physical Reasoning. Object-centric models learn structured representations and dynamics (Chen et al., 2021; Ding et al., 2021; Wu et al., 2023; Li et al., 2025b). VLM methods use targeted training, memories, perception modules, spatiotemporal tools, or simulator search (Balazadeh et al., 2025; Wang et al., 2025; Feng et al., 2025; Ghazanfari et al., 2025; Chow et al., 2025; Fan et al., 2024; 2025; Cherian et al., 2026), while complementary benchmarks assess physical consistency and world-model behavior in generated videos (Bansal et al., 2024; Meng et al., 2024; Li et al., 2025a). Their physics nevertheless remains implicit or procedurally fixed; ATW instead lets execution feedback select what to model and test.

Executable Physical Worlds. Differentiable simulators and video-conditioned models optimize parameters within specified dynamics (Howell et al., 2022; Ding et al., 2021; Li et al., 2023), while black-box approaches search simulator parameters from trajectory error (Cherian et al., 2026). Video-to-simulation methods reconstruct geometry and physical properties (Xie et al., 2024; Zhang et al., 2024; Chen et al., 2025a; Zhao et al., 2025), and executable-world methods use reconstructed dynamics for question answering (Yang et al., 2026). These systems assume a fixed representation, material family, or objective; ATW instead builds task-sufficient worlds with heterogeneous materials and coupled interactions.

Agents Reasoning through Engines. Language and vision agents interleave reasoning with actions, tools, or generated programs (Yao et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023; Liang et al., 2023). Agentic reconstruction combines visual models, code generation, execution, and feedback to construct explicit scenes (Yao et al., 2025; Yin et al., 2026); related agents use 3D programs for spatial reasoning (Luo et al., 2026; Chen et al., 2025c) or interactive environments for manipulation planning (Liu et al., 2025; Xu et al., 2026). These agents target observable scenes or prespecified goals, whereas physical questions may concern hidden, future, or counterfactual dynamics. ATW therefore makes world construction question-conditioned and chooses which physical hypotheses to instantiate and probe.

3 Method

ATW turns physical reasoning into interaction with an executable world. Given a video and its associated questions, the agent first constructs a task-sufficient physical world and then uses that world to conduct question-directed experiments. These are two consecutive stages: *World Modeling* revises perceptual and physical assumptions until it commits a world, after which *World Probing* adapts experiments within that world. Figure 2 illustrates this progression. We first describe the two-stage agentic framework, then present the PolyWorld Engine that makes executable worlds

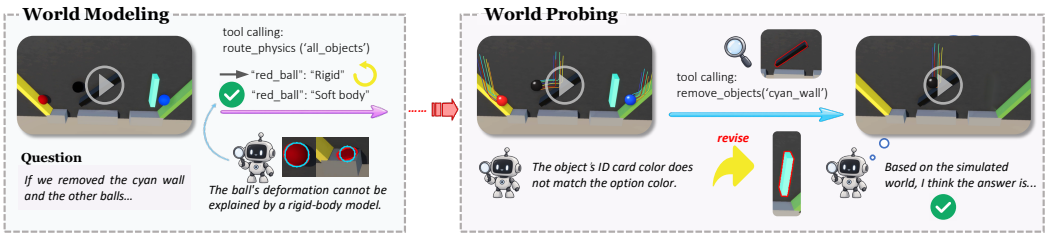

Figure 3: **Stage-local feedback and correction.** Within World Modeling, observed deformation causes the agent to replace a rigid-body hypothesis with a soft-body model. Within World Probing, execution feedback reveals an incorrect intervention target and causes the agent to probe the intended cyan wall. Feedback therefore changes the agent's decisions within each stage.

programmable by the agent, and finally explain how video evidence grounds and calibrates those worlds.

3.1 ATW: Two-Stage Agentic Physical Reasoning

Let $\mathcal{V}$ denote an input video and $\mathcal{Q} = \{q_i\}_{i=1}^{N}$ its physical questions. Rather than reconstructing every visible detail, ATW seeks an executable world $\mathcal{W}^\star$ containing the entities, physical representations, states, and latent parameters needed by these questions. We denote the stage-specific agent policies by $\mathcal{A}_{\mathrm{M}}$ and $\mathcal{A}_{\mathrm{P}}$. The former constructs the world, while the latter obtains evidence $\mathcal{E}_i$ for each question and produces the answer:

$$\mathcal{W}^\star = \mathcal{A}_{\mathrm{M}}(\mathcal{V}, \mathcal{Q}), \qquad \mathcal{E}_i = \mathcal{A}_{\mathrm{P}}(\mathcal{W}^\star, q_i), \qquad \hat{a}_i = \mathrm{Answer}(q_i, \mathcal{E}_i). \tag{1}$$

Both $\mathcal{A}_{\mathrm{M}}$ and $\mathcal{A}_{\mathrm{P}}$ are adaptive agents rather than fixed workflows: within its stage, each agent chooses its next operation from the current state, prior tool results, and execution feedback.

World Modeling. The modeling agent decides what must be represented before deciding how to recover it. It selects question-relevant entities, invokes visual tools to establish their geometry and motion, assigns a physical representation to each entity, and identifies parameters whose values must be inferred. The agent can inspect intermediate masks, tracks, reconstructions, and simulated trajectories; a mismatch can trigger another observation, a corrected object binding, a different physical representation, or renewed parameter fitting. This feedback remains internal to World Modeling. The stage terminates by committing $\mathcal{W}^\star$, including object identities, geometry, initial states, physical models, and fitted parameters.

World Probing. The probing agent treats $\mathcal{W}^\star$ as a fixed base world and decides which physical experiment will resolve each question. It can read an inferred property, continue the unmodified dynamics, instantiate an intervened branch, inspect events or trajectories, and compare outcomes across branches. Thus property questions can query the fitted world directly, predictive questions can extend its trajectory, and counterfactual or goal-driven questions can test edited worlds. Execution feedback may change the next probe or correct its intervention target, but it does not reopen World Modeling or alter the committed base world. Probing ends when the accumulated evidence supports an answer or the interaction budget is exhausted.

3.2 PolyWorld Engine: Agent-Operable Multiphysics Worlds

PolyWorld Engine is our lightweight and highly programmable multiphysics simulator built on Warp (Macklin, 2022). It places rigid bodies, soft bodies, cloth, ropes, and fluids in a common executable scene and supports their coupled interactions. Each physical system retains the state variables and dynamics appropriate to its material, while the shared scene allows the agent to compose heterogeneous systems when required by the observed interaction. Reusable scene data, simulation buffers, and explicit parameterization make repeated fitting and branched rollouts practical within an agent trajectory.

Table 1: **PolyWorld as an agent-facing semantic interface.** High-level operations connect agent decisions to physical execution while preserving explicit world and evidence provenance. Concrete tool signatures and backend-specific details are provided in Appendix A.

Interface role	Representative capabilities	Returned state or evidence
World construction	Bind objects, assign physical models, initialize states, fit parameters	Identities, geometry, model assignments, fitted dynamics
World inspection	Inspect frames or rollouts, read properties, query states	Visual observations, properties, trajectories, events
World experimentation	Continue dynamics, intervene, branch rollouts, compare worlds	Predictive and counterfactual outcomes with temporal scope
Execution control	Describe operations, revisit evidence, recover from failure, terminate	Tool constraints, execution status, retained evidence

PolyWorld exposes physics through semantic operations rather than requiring the agent to manipulate low-level simulator code. Table 1 summarizes the interface used across the two stages. The available operations are conditioned on the instantiated physical systems, and every result records its source world, objects, temporal scope, and execution status. This provenance lets the agent revisit earlier evidence without rerunning an experiment or confusing observations from different branches.

The committed world is reusable but not destructively edited during probing. Predictive rollouts continue the base world, whereas interventions create branches that inherit its fitted parameters and initial conditions. Multiple questions and candidate interventions can therefore share the same calibrated world, while their trajectories and events remain separately addressable. This design turns simulation into a persistent workspace for physical inquiry rather than a terminal tool invocation.

3.3 Grounding and Calibrating Executable Worlds

World Modeling converts video evidence into a world that is physically executable and sufficient for the questions at hand. The agent obtains object masks, motion tracks, and scene geometry from visual tools, and retains explicit bindings between referenced entities and objects in the executable scene. These observations play complementary roles: masks delimit objects and their visible deformation, tracks constrain temporal motion, and geometry establishes spatial configuration and initial conditions. The agent requests or refines only the evidence required by the selected representation, avoiding a mandatory reconstruction pipeline shared by all scenes.

For each instantiated physical representation h, PolyWorld maps an initial state s_0 and parameters $\boldsymbol{\theta}$ to a simulated trajectory, from which a model-specific readout is compared with visual evidence. Rigid systems use tracked positions and orientations; deformable systems additionally use shape, surface, centerline, or occupancy observations as appropriate. We fit the exposed parameters with the cross-entropy method (CEM). At iteration k, CEM evaluates bounded parameter samples under the corresponding trajectory discrepancy and uses the lowest-loss elite set to update its diagonal Gaussian proposal:

$$\begin{aligned} \boldsymbol{\mu}_{k+1} &= \alpha \boldsymbol{\mu}_k + (1-\alpha)\widehat{\boldsymbol{\mu}}_k, \\ \boldsymbol{\sigma}_{k+1} &= \max(\boldsymbol{\sigma}_{\min}, \alpha \boldsymbol{\sigma}_k + (1-\alpha)\widehat{\boldsymbol{\sigma}}_k)\,, \end{aligned} \tag{2}$$

where $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\boldsymbol{\sigma}}_k$ are the elite statistics, α controls smoothing, and $\boldsymbol{\sigma}_{\min}$ preserves exploration. Fitting progressively incorporates longer observation windows so that later interactions constrain the recovered dynamics (system-identification details: Appendix A.2).

Calibration also tests whether the current world hypothesis is adequate. Persistent motion or shape discrepancies can indicate that the error lies in the representation or object binding rather than its numerical parameters. The modeling agent can therefore revise those choices and repeat fitting before committing $\mathcal{W}^\star$ as the shared physical basis for subsequent experiments.

4 EXPERIMENTS

We evaluate ATW on CLEVRER (Yi et al., 2019), ContPhy (Zheng et al., 2024), and real-world videos against foundation VLMs and physical-reasoning methods. We also present qualitative results and ablations of adaptive reasoning, physical simulation, and system identification.

4.1 EXPERIMENTAL PROTOCOL

Benchmarks and metrics. CLEVRER (Yi et al., 2019) tests explanatory, predictive, and counterfactual rigid-body reasoning. Its validation subset contains 1,000 videos, 4,280 questions, and 14,228 options; we report per-option and per-question accuracy, with the latter requiring every option to be correct. ContPhy (Zheng et al., 2024) covers physical-property, predictive, counterfactual, and goal-driven reasoning. On its 600 videos and 1,950 questions, we report category accuracy and a question-count-weighted overall score. We additionally report category-wise accuracy on three real-world scenarios totaling 60 videos and questions (evaluation details: Appendix B).

Baselines and controlled variants. Baselines comprise foundation VLMs—Qwen3-VL-235B-A22B (Bai et al., 2025), GLM-4.6V (Z.ai, 2025), Gemini-3-Flash (Google DeepMind, 2025), Gemini-3.1-Pro (Google DeepMind, 2026), GPT-4o (OpenAI, 2024), and GPT-5.5 (OpenAI, 2026); training-based video reasoners—VideoRFT (Wang et al., 2025), Video-R1 (Feng et al., 2025), VideoThinker-R1 (Wu et al., 2026), and Chain-of-Frames (Ghazanfari et al., 2025); and training-free methods—VideoAgent (Fan et al., 2024), STAR (Fan et al., 2025), and PhysMind (Yang et al., 2026). Controlled variants isolate structured perception, physical execution, and adaptive reasoning.

Implementation details. ATW uses Gemini-3-Flash and PolyWorld Engine, without training or fine-tuning on either benchmark. Gemini baselines receive the video; GPT, Qwen, and GLM receive eight uniformly sampled frames (implementation details: Appendix A.1; prompts: Appendix D).

4.2 MAIN RESULTS

CLEVRER. Table 2 shows that ATW achieves 80.82% overall per-question accuracy, outperforming Gemini-3-Flash by 46.50 percentage points, GPT-5.5 by 13.58 points, and PhysMind, the strongest training-free baseline, by 8.27 points. ATW leads all three categories with 83.87% explanatory, 88.60% predictive, and 75.16% counterfactual accuracy. Its largest gain over GPT-5.5 is on counterfactual questions (+23.83 points), where it also exceeds PhysMind by 4.58 points.

ContPhy. Table 3 evaluates continuous-media and diverse-material tasks. ATW achieves 70.56% overall accuracy, 28.10 percentage points above Gemini-3-Flash and 3.53 points above GPT-5.5; it also outperforms Gemini-3-Flash in each of the four categories. Its gains over GPT-5.5 concentrate in physical-property and predictive reasoning, reaching 81.73% and 62.20% (+6.26 and +11.39 points). It nearly matches GPT-5.5 on goal-driven questions (a 1.93-point gap) but trails it by 6.46 points on counterfactual questions.

Real-world evaluation. Across the three real-world tasks, ATW achieves 85.00%, 70.00%, and 60.00% per-question accuracy on billiards counterfactual, bouncing-ball prediction, and toy-car prediction, respectively, outperforming both GPT-5.5 and Gemini-3-Flash in every category. Its overall accuracy is 71.67%, compared with 43.33% for GPT-5.5 and 33.33% for Gemini-3-Flash. Across these 60 videos, the gains show that executable-world reasoning extends to the three evaluated real-world settings despite appearance variation. The lower predictive-question score also reflects the greater difficulty of forecasting future trajectories in real-world scenes (detailed results: Appendix C.1).

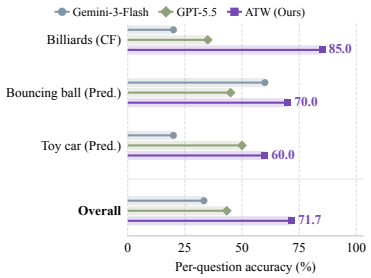

Figure 4: Real-world per-question accuracy (CF: counterfactual; Pred.: prediction).

Table 2: CLEVRER accuracy (%) on a validation subset of 1,000 videos. Explanatory questions identify causes, while predictive and counterfactual questions concern future and altered outcomes. Each category reports per-question and per-option accuracy; a question is correct only when all associated options are correct. Bold and underlined values indicate the best and second-best results in each column, respectively.

Method	Explanatory		Predictive		Counterfactual		Overall	
	per ques.	per opt.	per ques.	per opt.	per ques.	per opt.	per ques.	per opt.
Baselines								
Random	7.19	49.33	25.40	49.71	9.75	50.02	11.26	49.69
Blind Gemini-3-Flash	27.00	64.64	24.68	48.92	10.07	50.13	19.21	56.31
Foundation VLMs								
Qwen3-VL-235B-A22B	53.19	75.84	40.40	62.12	22.17	61.79	37.52	67.92
GLM-4.6V	39.33	73.75	63.64	66.74	24.31	58.70	36.68	66.02
Gemini-3-Flash	50.32	74.66	42.71	53.46	16.63	58.38	34.32	64.97
Gemini-3.1-Pro	60.90	79.17	61.76	71.28	25.37	63.50	45.47	71.06
GPT-4o	36.76	70.48	35.64	54.98	15.57	56.54	27.29	62.44
GPT-5.5	<u>77.21</u>	<u>89.55</u>	<u>85.71</u>	<u>90.69</u>	51.33	76.74	67.24	83.66
Training-Based Methods								
VideoRFT	16.66	61.98	39.25	47.40	15.35	54.98	19.74	57.28
Video-R1	32.79	71.14	41.56	46.25	19.62	59.10	28.43	63.08
VideoThinker-R1	17.53	63.60	41.56	42.50	15.14	53.71	20.37	56.92
Chain-of-Frames	44.71	77.26	84.42	87.88	33.48	70.84	46.21	75.29
Training-Free Methods								
VideoAgent	34.42	61.55	14.29	48.56	7.57	51.61	19.39	55.63
STAR	55.58	78.62	24.24	57.14	7.57	53.47	29.46	64.75
PhysMind	76.97	87.38	66.96	82.32	<u>70.58</u>	<u>88.08</u>	<u>72.55</u>	<u>87.22</u>
ATW (Ours)	**83.87**	**92.31**	**88.60**	**91.41**	**75.16**	**90.29**	**80.82**	**91.28**

Table 3: ContPhy accuracy (%) over 600 videos and 1,950 questions. Property questions test physical attributes, predictive and counterfactual questions test future and intervened outcomes, and goal-driven questions test action selection toward a target state. Bold and underlined values indicate the best and second-best results in each column, respectively.

Method	Property	Predictive	Counterfactual	Goal-driven	Overall
Baselines					
Random	40.53	31.50	12.92	13.90	28.36
Blind Gemini-3-Flash	47.47	41.26	23.16	29.34	37.90
Foundation VLMs					
Qwen3-VL-235B-A22B	58.80	46.54	30.51	27.41	45.03
Gemini-3-Flash	60.00	43.90	23.16	22.39	42.46
GPT-5.5	<u>75.47</u>	<u>50.81</u>	**69.49**	**69.11**	<u>67.03</u>
Training-Based Methods					
Video-R1-7B	43.47	34.15	11.14	15.44	29.95
Chain-of-Frames-8B	50.27	48.17	10.24	20.85	36.62
Training-Free Methods					
ATW (Ours)	**81.73**	**62.20**	<u>63.03</u>	<u>67.18</u>	**70.56**

Figure 5 visualizes the selected modeling steps across three real-world tasks. Segmentation identifies task-relevant objects, pose tracking preserves identities and motion across key frames, and dynamic reconstruction forms executable states. Together, they provide the geometry, trajectories, and interactions needed for targeted prediction and counterfactual probes.

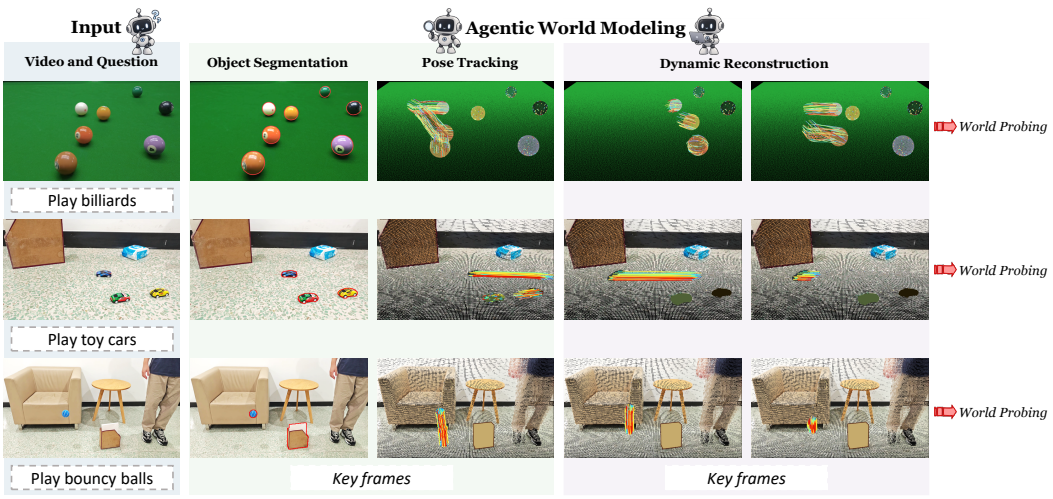

Figure 5: Agentic world modeling from real-world videos of billiards, toy cars, and bouncing balls. Selected key frames show object segmentation, pose tracking, and dynamic reconstruction, producing executable worlds for subsequent probing.

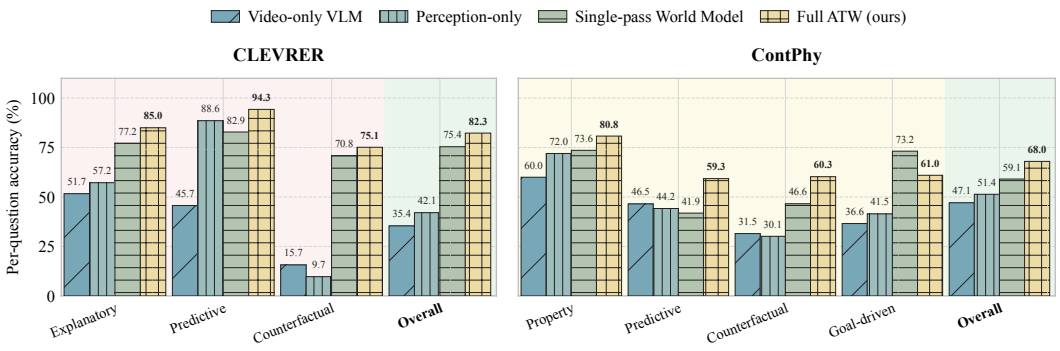

Figure 6: Ablation of agentic world reasoning on the CLEVRER and ContPhy subsets. All four variants use Gemini-3-Flash. Bars and labels show per-question accuracy; overall scores aggregate all questions in each subset.

4.3 Ablation Studies

We conduct ablation studies on CLEVRER and ContPhy subsets, each containing 100 scenes. CLEVRER includes 435 questions and 1,439 answer options, while ContPhy contains 325 questions. These studies examine the roles of executable worlds and adaptive reasoning, the effectiveness of physical simulation and system identification, and the applicability of ATW across foundation models. Within each comparison, the data and evaluation protocol remain fixed so that performance changes isolate the component under study.

Agentic world reasoning. Figure 6 compares four variants: Video-only VLM receives only the video and question; Perception-only (w/o Executable World) additionally receives ATW's structured detections, masks, tracks, and geometry, but neither builds nor executes a physical world; Single-pass World Model builds and executes a world once without feedback-driven revision or additional probing; and Full ATW (ours) adaptively updates its world representation, physical parameters, and subsequent probes based on the question and accumulated feedback. Overall accuracy increases across these variants from 35.40% to 42.07%, 75.40%, and 82.30% on CLEVRER, and from 47.08% to 51.38%, 59.08%, and 68.00% on ContPhy. Structured perception improves video-only reasoning, while physical execution brings a larger gain. Relative to Perception-only, Single-pass World Model raises CLEVRER counterfactual accuracy by 61.08 percentage points,

from 9.73% to 70.81%. Feedback-guided revision in Full ATW then adds 6.90 and 8.92 points over Single-pass on CLEVRER and ContPhy, respectively. Despite some variation across individual reasoning categories, the overall trend across both datasets shows that executable worlds provide substantial gains, while feedback-guided revision brings further improvements (see Appendix C.2).

System identification and physical backends. Figure 7 evaluates both factors on the CLEVRER ablation subset. With Warp fixed and equal simulation budgets, CEM achieves a substantially lower median trajectory RMSE than Random Search, while raising per-question accuracy from 43.45% to 82.76% and per-option accuracy from 69.49% to 93.19%. With CEM fixed, Warp achieves the lowest trajectory error and the highest per-question accuracy among the PhysMind physical model (Yang et al., 2026), MuJoCo (Todorov et al., 2012), and Warp. This paired design separates optimization quality from backend fidelity: stronger search improves the same simulator, while a more suitable backend improves reasoning under the same optimizer. Across both ablations, lower fitting error is consistently associated with stronger downstream reasoning, demonstrating that reliable physical grounding depends on both system identification and simulation (see Appendix C.3).

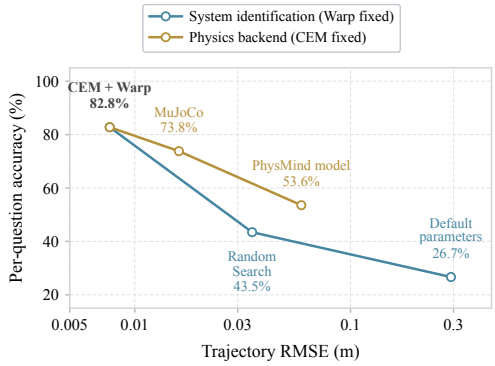

Figure 7: System-identification and physics-backend ablations on CLEVRER. Overall per-question accuracy versus median trajectory RMSE (log scale). System identification fixes Warp; backend comparisons fix CEM. Both ablations share CEM + Warp. CEM and Random Search use equal simulation budgets.

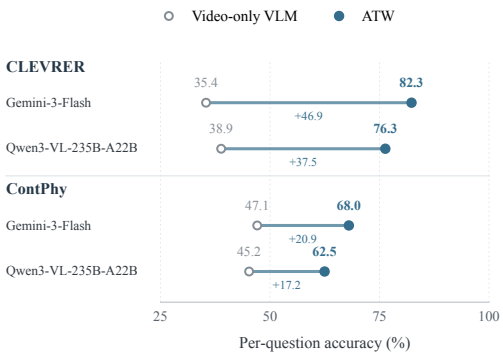

Figure 8: Across foundation models. Overall per-question accuracy; open: Video-only VLM, filled: ATW. Values below the lines show gains in percentage points.

Generalization across foundation models. Figure 8 extends the Gemini-3-Flash evaluation to Qwen3-VL-235B-A22B, replacing the backbone in both world modeling and world probing. With Qwen, ATW improves overall per-question accuracy over Video-only VLM from 38.85% to 76.32% on CLEVRER and from 45.23% to 62.46% on ContPhy. Across the two evaluated backbones, every paired comparison improves in the same direction, demonstrating backbone portability in both the tested rigid-body and continuous-physics settings.

5 Conclusion

We presented Asking the World (ATW), a training-free framework for generalist physical reasoning through agentic world modeling and probing. Rather than treating simulation as a fixed downstream tool, ATW constructs task-relevant executable worlds and uses observations and execution feedback to revise modeling decisions and guide question-directed probes. PolyWorld Engine supports diverse materials and coupled interactions, while system identification grounds world dynamics in visual evidence. Across CLEVRER, ContPhy, and three real-world scenarios, ATW improves reasoning accuracy, while controlled ablations show complementary contributions from physical execution and feedback-guided adaptation. These findings suggest that actively building and interrogating executable worlds provides a promising path toward physical reasoning beyond direct visual inference.

AI USE STATEMENT

Generative AI tools, including large language models, were used to assist with implementing portions of the research code, drafting portions of the manuscript, and improving clarity and readability. They were not used to generate synthetic datasets or to formulate or prove mathematical claims. The authors reviewed and tested all AI-assisted code, verified the reported results, and take responsibility for the final content.

ETHICS STATEMENT

This work uses public benchmarks and author-recorded tabletop scenes and involves no human participants, personal data, or decisions about individuals. Because errors in perception or simulation can produce incorrect conclusions, the method should not be used in safety-critical settings without independent validation.

REPRODUCIBILITY STATEMENT

The paper describes the method, experimental protocol, metrics, baselines, and ablations. Appendices A–D provide additional implementation context, benchmark and real-world protocols, supplementary experimental results, and system-prompt summaries to support interpretation and assessment of the reported findings.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7318–7328, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation, 2024. URL https://arxiv.org/abs/2406.03520.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Chuhao Chen, Zhiyang Dou, Chen Wang, Yiming Huang, Anjun Chen, Qiao Feng, Jiatao Gu, and Lingjie Liu. Vid2Sim: Generalizable, video-based reconstruction of appearance, geometry and physics for mesh-free simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26545–26555, 2025a.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. SAM 3D: 3dfy anything in images, 2025b. URL https://arxiv.org/abs/2511.16624.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning, 2025c. URL https://arxiv.org/abs/2511.22659.

Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bhCDO_cEGCz.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. LLMPhy: Parameter-identifiable physical reasoning combining large language models and physics engines. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q6a9W6kzv5.

Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Joshua B. Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. In *Advances in Neural Information Processing Systems*, volume 34, pp. 887–899, 2021.

Sunqi Fan, Jiashuo Cui, Meng-Hao Guo, and Shuojin Yang. Tool-augmented spatiotemporal reasoning for streamlining video question answering task. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129243–129273, 2025.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *Computer Vision – ECCV 2024*, pp. 75–92. Springer, 2024. doi: 10.1007/978-3-031-72670-5_5.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99114–99137, 2025.

Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal LLMs via frame-aware reasoning, 2025. URL https://arxiv.org/abs/2506.00318.

Google DeepMind. Gemini 3 Flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, 2025. Accessed: 2026-09-24.

Google DeepMind. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026. Accessed: 2026-09-24.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023. doi: 10.1109/CVPR52729.2023.01436.

Taylor A. Howell, Simon Le Cleac'h, Jan Brüdigam, Qianzhong Chen, Jiankai Sun, J. Zico Kolter, Mac Schwager, and Zachary Manchester. Dojo: A differentiable physics engine for robotics, 2022. URL https://arxiv.org/abs/2203.00806.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61064–61088, 2025a. doi: 10.52202/085713-1834.

Jian Li, Wan Han, Ning Lin, Yu-Liang Zhan, Ruizhi Chengze, Haining Wang, Yi Zhang, Hongsheng Liu, Zidong Wang, Fan Yu, et al. SlotPi: Physics-informed object-centric reasoning models, 2025b. URL https://arxiv.org/abs/2506.10778.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Fei-Fei Li, and Ehsan Adeli. QuantiPhy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models, 2025c. URL https://arxiv.org/abs/2512.19526.

Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification, 2023. URL https://arxiv.org/abs/2303.05512.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation*, pp. 9493–9500, 2023. doi: 10.1109/ICRA48891.2023.10160591.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Haowen Liu, Shaoxiong Yao, Haonan Chen, Jiawei Gao, Jiayuan Mao, Jia-Bin Huang, and Yilun Du. SIMPACT: Simulation-enabled action planning using vision-language models, 2025. URL https://arxiv.org/abs/2512.05955.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4world: Feedforward world-centric dense 3d tracking of all pixels. *arXiv preprint arXiv:2603.02573*, 2026.

Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia P. Sycara, and Yaqi Xie. pySpatial: Generating 3d visual programs for zero-shot spatial reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yv15C8ql24.

Miles Macklin. Warp: A high-performance python framework for GPU simulation and graphics. NVIDIA GPU Technology Conference (GTC), March 2022. URL https://github.com/NVIDIA/warp.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation, 2024. URL https://arxiv.org/abs/2410.05363.

OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2026-09-24.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, 2026. Accessed: 2026-09-24.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11854–11864, 2023. doi: 10.1109/ICCV51070.2023.01092.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Learning single-image calibration with geometric optimization. In *Computer Vision – ECCV 2024*, pp. 1–20. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73661-2_1.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. VideoRFT: Incentivizing video reasoning capability in MLLMs via reinforced fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4350–4376, 2025.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17868–17879, 2024. doi: 10.1109/CVPR52733.2024.01692. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wen_FoundationPose_Unified_6D_Pose_Estimation_and_Tracking_of_Novel_Objects_CVPR_2024_paper.html.

Jingze Wu, Quan Zhang, Hongfei Suo, Zeqiang Cai, and Hongbo Chen. Beyond perceptual shortcuts: Causal-inspired debiasing optimization for generalizable video reasoning in lightweight MLLMs, 2026. URL https://arxiv.org/abs/2605.01324.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.05861.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024. doi: 10.1109/CVPR52733.2024.00420.

Wenjiang Xu, Mingkang Zhang, Cindy Wang, Rui Fang, Lusong Li, Jing Xu, Jiayuan Gu, Zecui Zeng, and Rui Chen. Embodied tree of thoughts: Deliberate manipulation planning with embodied world model. *IEEE Robotics and Automation Letters*, pp. 1–8, 2026. doi: 10.1109/LRA.2026.3716133.

Chen Yang, Shenxiang Zeng, Haoyang Zhao, Zhouyuan Xu, Youquan He, Haoyu Li, Mingyi Deng, Jiansheng Fan, and Chen Wang. PhysMind: From video to executable worlds for training-free physical reasoning, 2026. URL https://arxiv.org/abs/2608.04575.

Kaixin Yao, Longwen Zhang, Xinhao Yan, Yan Zeng, Qixuan Zhang, Lan Xu, Wei Yang, Jiayuan Gu, and Jingyi Yu. CAST: Component-aligned 3d scene reconstruction from an RGB image. *ACM Transactions on Graphics*, 44(4):1–19, 2025. doi: 10.1145/3730841.

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.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning, 2019. URL https://arxiv.org/abs/1910.01442.

Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Chenyang Wang, Xiuyu Li, Michael J. Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-inverse-graphics agent via interleaved multimodal reasoning, 2026. URL https://arxiv.org/abs/2601.11109.

Z.ai. GLM-V: GLM-4.6V, GLM-4.5V, and GLM-4.1V-Thinking. https://github.com/zai-org/GLM-V, 2025. Accessed: 2026-09-24.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *Computer Vision – ECCV 2024*, pp. 388–406, 2024. doi: 10.1007/978-3-031-72627-9_22.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. PhysSplat: Efficient physics simulation for 3d scenes via MLLM-guided gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5242–5252, 2025.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. ContPhy: Continuum physical concept learning and reasoning from videos. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61526–61558. PMLR, 2024.

A More Implementation Details

A.1 Models and Visual Foundation Modules

Reasoning models. Gemini-3-Flash (Google DeepMind, 2025) serves as the default backbone for both world modeling and world probing. The cross-backbone experiment replaces the model in both stages with Qwen3-VL-235B-A22B (Bai et al., 2025). These models select operations, interpret returned evidence, and produce answers. We do not train or fine-tune the reasoning models or visual modules on CLEVRER or ContPhy.

SAM 3. SAM 3 (Carion et al., 2025) supports concept-prompted segmentation of objects in images and videos. In ATW, the agent supplies object descriptions to obtain masks and associated video identities. These masks delimit the visual evidence used for object reconstruction and motion tracking, and connect image regions to retained scene entities. The agent can inspect the segmentation results and request another segmentation when the current masks do not adequately capture the objects relevant to modeling.

Track4World. Track4World (Lu et al., 2026) provides dense 3D tracking in a shared world coordinate system. We use its Depth Anything 3 variant to obtain geometric observations and inter-frame correspondences from the input video. For dynamic rigid objects, correspondences initialize pose propagation, followed by point-cloud registration against the observed geometry. The resulting motion estimates are retained with object identities and confidence information, providing temporal evidence for fitting physical parameters and comparing simulated motion with observations.

Depth Anything 3. Depth Anything 3 (Lin et al., 2025) recovers scene geometry from visual inputs and serves as the geometric backbone of our Track4World configuration. Through this integration, the perception pipeline obtains camera intrinsics, camera poses, and dense 3D points in camera and world coordinates. These outputs provide the spatial reference for aligning reconstructed object geometry with the video. Geometry confidence and validity information support the selection of usable observations for downstream reconstruction and tracking.

SAM 3D Objects. SAM 3D Objects (Chen et al., 2025b) reconstructs object geometry from an image and an object mask. ATW applies it to selected reference observations to obtain canonical object geometry, which is aligned with the recovered scene geometry. For rigid objects, this reference shape is retained while subsequent frames update the object's pose. Keeping geometry and motion separate provides a consistent object representation for tracking, physical modeling, and visualization of the reconstructed world.

FoundationPose. FoundationPose (Wen et al., 2024) estimates and refines the 6D pose of an object using its geometry and visual observations. In our implementation, it refines the reference-frame alignment between the reconstructed object and the observed scene. This pose anchors the canonical object geometry in the recovered coordinate system. Subsequent dynamic poses use Track4World correspondences and point-cloud registration, while static objects retain their aligned reference pose. FoundationPose therefore supplies the initial geometric alignment used by temporal tracking.

A.2 Warp-Based Simulation and CEM Fitting

Shared simulation interface. PolyWorld Engine uses Warp (Macklin, 2022) to execute physical models for rigid bodies, soft bodies, cloth, ropes, and fluids. Each model receives scene geometry, initial conditions, and a candidate parameter vector, and produces a simulated trajectory. Warp kernels accelerate the forward dynamics on the GPU, while reusable scene data and simulation buffers support repeated candidate evaluation. The fitting interface expresses this process as

$$\widehat{Y}_{1:T}(\theta) = \mathcal{O}_h(\mathcal{F}_h(s_0, \theta; T)), \tag{3}$$

where h denotes the physical representation, s_0 its initial state, $\mathcal{F}_h$ the forward simulator, and $\mathcal{O}_h$ the readout that converts simulated states into quantities comparable to visual observations.

Observation-driven fitting. System identification fits the parameters exposed by the instantiated physical model to the retained visual evidence. These parameters describe motion, material response, contact, or boundary conditions, depending on the system. Across the implementations, the fitting objective can be expressed as

$$\theta^\star = \arg\min_{\theta \in \Theta_h} \mathcal{L}_h(\theta), \qquad \mathcal{L}_h(\theta) = D_h\Big(\widehat{Y}_{1:T}(\theta), Y_{1:T}\Big) + \lambda_h R_h(\theta), \tag{4}$$

where $Y_{1:T}$ is the observed evidence, D_h measures the corresponding motion or shape discrepancy, and Θ_h specifies admissible parameters. R_h represents a parameter prior when used, with $\lambda_h = 0$ otherwise. Observation validity and confidence determine which measurements contribute to the discrepancy. This formulation allows each physical model to use the evidence that directly describes its dynamics.

Evidence across physical systems. For *rigid bodies*, fitting compares simulated and tracked positions and rotations to recover motion and contact-response parameters. For *soft bodies*, it combines center motion with directional deformation, linking material response to observed changes in shape. For *cloth*, it compares simulated surface points with tracked trajectories to fit motion, stiffness, damping, and boundary conditions. For *ropes*, it uses centerline geometry and the trajectories of connected objects to fit constrained dynamics. For *fluids*, it compares simulated spatial distributions with observed fluid regions to fit flow behavior and source conditions, including relative density and emission parameters. These observation choices adapt the shared fitting procedure to the representation of each system.

CEM search. CEM maintains a diagonal Gaussian proposal over the fitted parameters. At iteration k, it samples a population, projects candidates onto their bounds, and selects the lowest-loss candidates as an elite set:

$$\theta_k^{(j)} = \Pi_{\Theta_h}(\mu_k + \sigma_k \odot \epsilon_j), \quad \epsilon_j \sim \mathcal{N}(0, I), \qquad \mathcal{E}_k = \mathrm{TopK}_{\mathrm{lowest}\ \mathcal{L}_h}\{\theta_k^{(j)}\}_{j=1}^N. \tag{5}$$

Here N is the population size and K is the elite count. The elite mean and standard deviation update the proposal through the smoothed rule in Equation 2, with a variance floor maintaining exploration. The search reuses promising candidates and progressively extends the fitting horizon so that later interactions contribute to parameter fitting. Model-specific parameterizations and readouts connect this common search procedure to the different forward simulators.

Reusable fitted worlds. The fitted parameters are retained with the modeled geometry and initial conditions, yielding an executable world that can be replayed and queried. World probing uses these assets to continue the dynamics or execute an intervention, such as removing an object or changing an exposed physical condition. The same forward model therefore connects parameter fitting with subsequent question-conditioned experiments.

A.3 Query-Conditioned Physical Rollout Tools

The probing agent receives the retained world assets, object identities, available capabilities, and the current physical question. Tool availability and temporal scope depend on the question and the instantiated world. The following signatures are pseudocode abstractions of general operations; each backend exposes the subset supported by its physical representation.

Asset and property queries. `InspectWorld(objects, fields)` retrieves identities, geometry, visibility, and available capabilities. `ReadProperty(objects, property)` reads supported physical quantities from the fitted world, allowing property questions to use existing parameters without an additional rollout. These queries preserve the distinction between observed attributes and fitted physical properties.

Rollouts and interventions. `Rollout(world, intervention, horizon)` executes an unedited or modified world and returns a rollout identifier. Predictive questions use the unedited world and read the interval after the observation cutoff. Counterfactual questions specify an intervention before execution, while goal-directed questions can test candidate interventions in separate worlds. Object removal is one supported edit; other edits are exposed according to backend capabilities. In the rigid implementation, removal applies from the modeled beginning of the replay. Fitted parameters of the retained objects are reused for the edited rollout.

Evidence readout. `ReadEvents(rollout, objects, interval)` and `ReadTrajectory(rollout, object, interval)` retrieve time-localized outcomes. `ReadTimeline(rollout, interval)` organizes collision, entry, and exit events chronologically, retaining event times and participating object identities to support reasoning about their order. `QueryWorld(rollout, predicate, interval)` requests semantic conditions, such as contact, motion, or region occupancy, with the appropriate temporal aggregation. Readouts refer to an explicit world and time interval. Several queries can accompany a rollout request, and retained rollouts can be queried repeatedly.

Cross-world comparison. `CompareWorlds(base, edited, objects)` compares selected objects across an original world and an intervened world. It aligns their trajectories at shared frames and reports the maximum center displacement for each object, together with the two event sequences and simulation horizons. These outputs show how an intervention changes object motion and event occurrence, providing evidence for comparing the consequences of alternative physical conditions.

Observation-based checks. `InspectObservedMotion(object, interval)` retrieves motion-related measurements from the object's masks in the input video. The returned observations include image-space centroid positions, visible mask areas, and image-boundary contact over the requested interval. Together with the original frames, these measurements let the agent inspect where and when an object moves and compare the observed changes with its tracked or simulated trajectory.

Inspection and control. `DescribeTool(tool)` provides an operation's inputs, parameter constraints, and conditions of use, helping the agent prepare its next call. `InspectFrames(times, object)` retrieves visual observations, while `InspectRollout(rollout, interval)` presents the simulated interval for visual inspection of motion and deformation alongside numerical readouts. `ReadEvidence(id)` revisits an earlier tool result, and `UpdateBindings(correction, reason)` revises question-to-object associations using accumulated evidence. Finally, `Finish(evidence)` ends probing and passes the evidence to answer generation. Tool returns retain their observation or simulation source, object and rollout identities, temporal coverage, and execution status.

A.4 SAM 3 Agent for General-Purpose Video Segmentation

Object discovery and seed selection. The SAM 3 agent combines VLM-based visual reasoning with SAM 3 (Carion et al., 2025) segmentation to produce object-level mask tracks. Given a video and indexed frame samples, the VLM identifies distinct entities in the interaction workspace and generates discriminative descriptions and segmentation prompts. It selects a reference frame for each entity, prioritizing complete outlines, visibility, and identity clarity. SAM 3 then generates candidate masks, which the VLM inspects alongside the original image to select a mask covering the target while excluding neighboring objects and background.

Feedback-guided refinement and propagation. When no candidate is accepted, the agent checks the target's visibility and can select another reference frame or propose alternative descriptions of the same entity. Candidate masks and rejection feedback guide these adjustments while preserving the target identity. Accepted seeds initialize bidirectional video-mask propagation in a shared multi-object state. The module returns indexed mask sequences, object-to-track associations, and temporal quality summaries for downstream reconstruction and motion tracking. This interface supports reusable object segmentation across scenes without requiring an object mesh or a rigid-motion assumption.

B Evaluation Benchmark Details

B.1 CLEVRER

CLEVRER (Yi et al., 2019) is a video reasoning benchmark designed to study causal understanding of physical events. It pairs synthetic videos of interacting rigid objects with questions linking object

identities, motion, and collisions. Objects vary in shape, color, and material appearance, allowing questions to refer to specific entities and their interactions. We evaluate ATW on a validation subset containing 1,000 videos, 4,280 questions, and 14,228 answer options.

Explanatory questions. These questions ask which earlier events are responsible for an observed outcome. Answering them requires identifying the relevant objects, recovering the temporal order of their interactions, and relating a target event to preceding motion or collisions. The candidate answers refer to possible explanatory events, so the model must distinguish the interactions that account for the outcome from other events in the video.

Predictive questions. These questions ask what will happen after the observed video ends. The model must infer how the objects will continue moving and whether their subsequent trajectories will lead to the events described by the answer options. This category connects the visible history of motion and collisions with future physical interactions, testing whether the observed evidence supports an accurate continuation of the scene.

Counterfactual questions. These questions ask how events would change under a specified intervention, such as removing an object from the scene. The model must identify the intervention target and reason about how the remaining objects would interact under the modified conditions. Because removing an object can alter subsequent collisions and motion, answering requires considering the consequences of the edit throughout the relevant event sequence.

Metrics. We report per-option and per-question accuracy for each category and overall. Per-option accuracy is the fraction of answer options whose correctness is judged correctly. Per-question accuracy counts a question as correct only when every associated option is judged correctly. We use overall per-question accuracy as the primary metric. Overall scores are computed across all evaluated questions or options, respectively.

B.2 CONTPHY

ContPhy (Zheng et al., 2024) is a benchmark for learning and reasoning about physical concepts from videos of continuum systems. It connects observed dynamics with questions about physical properties and the consequences of changing a scene. Its diverse materials and interactions support evaluation of reasoning across different physical systems. We evaluate ATW on 600 videos containing 1,950 questions.

Physical-property questions. These questions ask about physical attributes that govern the behavior of the objects or materials in a scene. The model must connect observable motion, deformation, or interaction responses with the underlying property being queried. Different scene categories provide different forms of evidence, making this category a test of whether visual dynamics can support judgments about the physical characteristics of the participating entities.

Predictive questions. These questions ask how a physical system will evolve after the observed interval. The model must interpret the current configuration and motion, then anticipate the subsequent behavior of the materials and objects involved. Depending on the scene, relevant evidence may include deformation, contact, constrained movement, or flow. The requested answer concerns the future outcome of the system under its existing conditions.

Counterfactual questions. These questions ask about the outcome of changing a condition in the observed scene. The model must interpret the specified change, identify the affected entities, and reason about how their interactions would unfold in the modified setting. Across the material families, this requires connecting an intervention with changes in motion, deformation, or flow and evaluating the resulting outcome described by the question.

Goal-driven questions. These questions specify a desired physical outcome and ask which intervention would produce it. The model must relate each candidate change to its consequences for the system and select the option that satisfies the stated goal. Reasoning therefore proceeds from a target

outcome to a suitable change in the scene, using the observed configuration and dynamics to assess the available choices.

Cloth scenes. Cloth scenes feature flexible surfaces that bend and deform as they move and interact with other objects. Their evolution depends on the initial configuration, material response, and contact geometry. Understanding these scenes requires following changes in surface shape and spatial relationships over time. They provide visual evidence for questions about material behavior, subsequent deformation, and the effects of changing the physical setup.

Rope scenes. Rope scenes feature connected flexible structures whose motion is constrained by their geometry and attachments, including interactions with pulleys and attached objects. Movement in one part of the system can affect other connected parts. Reasoning about these scenes requires tracking these relationships and understanding how forces and constraints shape the resulting motion, supporting questions about physical properties and the consequences of interventions.

Soft-ball scenes. Soft-ball scenes feature deformable bodies whose shapes change during motion and contact. Their responses to interactions provide evidence about material behavior and influence subsequent trajectories. Understanding these scenes requires relating object-level movement to visible deformation, including how the body responds around contact. The questions use this evidence to probe physical attributes, future outcomes, and changes induced by modifying the scene conditions.

Fluid scenes. Fluid scenes feature flowing materials that interact with containers, boundaries, and obstacles. Their behavior is expressed through changes in spatial distribution and the movement of material between regions. Reasoning requires following the evolving flow and relating it to scene geometry and physical conditions. These scenes support questions about material properties, future distributions, and how changes to the environment affect the resulting flow.

Metrics. We report question-answering accuracy for each question category and overall. Category accuracy is the number of correctly answered questions divided by the number of questions in that category. Overall accuracy is the total number of correctly answered questions divided by 1,950, so each evaluated question contributes equally to the final score.

B.3 REAL-WORLD

We collect a real-world dataset using consumer-grade cameras to evaluate physical reasoning from recorded interactions. The dataset contains 60 videos and 60 questions across three scene categories, with 20 videos and 20 questions per category. It focuses on two challenging reasoning tasks: counterfactual reasoning and prediction. Billiards scenes provide counterfactual questions, while bouncing-ball and toy-car scenes provide predictive questions.

Billiards scenes. These scenes capture balls moving and colliding on a billiards table. Counterfactual questions ask how the interactions would change under a specified intervention, such as removing a ball. Answering requires identifying the relevant balls, interpreting their observed motion, and reasoning about how the intervention changes subsequent collisions. This category evaluates the connection between object-level interventions and the resulting sequence of physical events.

Bouncing-ball scenes. These scenes capture the motion of bouncing balls and their interactions with surrounding surfaces. Predictive questions ask about physical outcomes after the observed motion. The model must connect the ball's trajectory and contact response with its subsequent movement, accounting for changes in direction caused by interactions. This category evaluates whether observed dynamics support predictions of future motion and contact events.

Toy-car scenes. These scenes capture toy cars moving through a physical environment and interacting with other scene elements. Predictive questions concern the subsequent motion or interactions of the cars. Answering requires identifying the relevant vehicles, interpreting their trajectories, and relating their movement to the surrounding geometry. This category evaluates prediction from real object motion under the spatial constraints of the recorded scene.

Ground-truth acquisition. Each scene is recorded in full from multiple synchronized viewpoints. For predictive questions, the reference answers are determined from the future portion of the complete recording, with complementary views used to recover the subsequent trajectory and resolve occlusions. Counterfactual outcomes are not directly observed; we therefore retain only questions for which human observers can determine the answer with high confidence from the multi-view recordings. In particular, cases involving unresolved spatial ambiguity or viewpoint occlusion are excluded.

Metrics. We report per-question accuracy for each scene category and overall. Each category score is the number of correctly answered questions divided by 20, and the overall score is the total number of correct answers divided by 60. Since the categories contain equal numbers of questions, overall accuracy also equals the mean of the three category accuracies.

C More Experimental Results

C.1 Detailed Real-World Results

Table 4 reports detailed accuracy on the 60 real-world videos and questions. In addition to per-question accuracy, we report per-option accuracy for billiards and toy-car tasks. Each bouncing-ball question requires one yes/no judgment. Overall per-judgment accuracy aggregates the 80 billiards options, 20 bouncing-ball judgments, and 60 toy-car options, for 160 judgments in total.

Table 4: Detailed real-world accuracy (%). Each scenario contains 20 videos and 20 questions. Q: per-question; O: per-option; J: per-judgment. Counts are shown in parentheses. Per-question accuracy requires all associated option judgments to be correct. CF: counterfactual; Pred.: predictive. Bold denotes the best result in each column.

Model	Billiards (CF)		Bouncing ball	Toy car (Pred.)		Overall	
	Q (20)	O (80)	Pred. / Q (20)	Q (20)	O (60)	Q (60)	J (160)
Foundation VLMs							
GPT-5.5	35.00	61.25	45.00	50.00	65.00	43.33	60.63
Gemini-3-Flash	20.00	60.00	60.00	20.00	26.67	33.33	47.50
Training-Free Methods							
ATW (Ours)	**85.00**	**91.25**	**70.00**	**60.00**	**71.67**	**71.67**	**81.25**

C.2 Additional Ablation Results

Figure 9 reports additional ablations on the same 100-scene subsets used in Section 4.3, with full ATW included for comparison. The identity-binding ablation removes the appearance cards and identity-binding table.

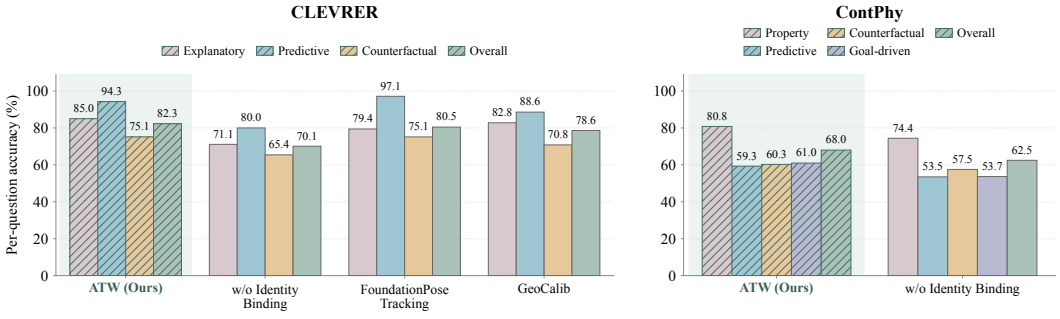

Figure 9: Additional ablations on CLEVRER (left) and ContPhy (right). Bars show per-question accuracy for each reasoning category and overall, evaluated on 100 scenes per benchmark (435 CLEVRER questions and 325 ContPhy questions). Hatched bars denote full ATW.

Effects of identity binding and perception modules. Figure 9 shows that removing appearance cards and identity bindings reduces overall accuracy from 82.30% to 70.11% on CLEVRER and from 68.00% to 62.46% on ContPhy, with declines across every question category. These results highlight the importance of maintaining object identities when connecting visual observations, question references, and executable worlds. On CLEVRER, FoundationPose (Wen et al., 2024) tracking achieves 80.46% overall accuracy and the highest predictive accuracy of 97.14%, while matching ATW on counterfactual questions. Its tracking benefits from an input object mesh and a rigid-body prior, making it effective for these rigid-object scenes. However, its rigid-pose formulation does not capture the evolving shapes of cloth, fluids, or soft bodies. We therefore use Track4World (Lu et al., 2026) for trajectory tracking across the diverse physical systems considered in ATW. The GeoCalib (Veicht et al., 2024) variant, which uses a learning-based single-image camera calibration method to estimate camera intrinsics and gravity direction, achieves 78.62% overall accuracy. Full ATW obtains the highest overall score, while different module choices produce distinct accuracy profiles across reasoning categories.

C.3 Detailed Physical-Grounding Ablations

Tables 5 and 6 provide the category-level results underlying the system-identification and physics-backend ablations. Overall accuracies are aggregated over all questions or options rather than averaged across categories. These experiments are run independently of the agentic world-reasoning ablation.

Table 5: Detailed system-identification ablation on the 100-scene ContPhy subset. Category columns report question-answering accuracy (%). Random Search and CEM use equal simulation budgets.

Method	Property	Predictive	Counterfactual	Goal-driven	Overall
System Identification (Warp Fixed)					
Default parameters	70.40	39.53	56.16	80.49	60.31
Random Search	70.40	51.16	63.01	46.34	60.62
CEM (Ours)	80.00	54.65	63.01	65.85	67.69

ContPhy trends. CEM achieves the strongest aggregate performance among the three configurations. Relative to Random Search, it improves physical-property, predictive, goal-driven, and overall accuracy, while matching counterfactual accuracy at 63.01%. These results indicate that CEM-based system identification improves aggregate reasoning performance, with category-specific differences.

Table 6: Detailed system-identification and physics-backend ablations on the 100-scene CLEVRER subset. Accuracy is reported in percent (Q: per-question; O: per-option), and the final column reports median trajectory RMSE in meters. The CEM and Warp + CEM rows are the same shared configuration.

Method	Explanatory		Predictive		Counterfactual		Overall		Median trajectory RMSE (m)
	Q	O	Q	O	Q	O	Q	O	
System Identification (Warp Fixed)									
Default parameters	25.56	50.62	97.14	98.57	1.08	50.23	26.67	55.11	0.292053
Random Search	41.67	64.77	91.43	92.86	27.03	69.18	43.45	69.49	0.034992
CEM (Ours)	86.11	94.31	94.29	97.14	75.14	91.22	82.76	93.19	0.007655
Physics Backends (CEM Fixed)									
PhysMind model + CEM	58.33	76.15	94.29	95.71	33.51	72.88	53.56	76.58	0.059138
MuJoCo + CEM	68.89	83.38	91.43	94.29	71.89	87.52	73.79	86.31	0.016017
Warp + CEM (Ours)	86.11	94.31	94.29	97.14	75.14	91.22	82.76	93.19	0.007655

CLEVRER trends. Across the system-identification variants, trajectory RMSE decreases from default parameters to Random Search and then to CEM, while overall per-question and per-option

accuracy increase in the same order. The largest reasoning gains appear in explanatory and counterfactual questions, while predictive accuracy remains high across configurations. With CEM fixed, the comparison from the PhysMind model through MuJoCo to Warp shows the same broad relationship between lower trajectory error and higher overall accuracy. Together, these trends show that both system identification and the physical backend determine the quality of the executable world used for downstream reasoning.

D SYSTEM PROMPTS

This section presents the principal system instructions used in ATW. We organize them according to the main stages of the framework: world-representation routing, object discovery and segmentation, world probing, and evidence-based answer generation.

World-representation routing.

Inspect the complete video together with all associated questions, but do not answer them at this stage. First identify whether the main interactions involve free collision and support, constrained relative displacement, or another form of material motion. Examine depth ordering and outline truncation during contact. Use a three-dimensional representation when occlusion makes a visible mask boundary different from the physical contact surface, or when folding, covering, and other out-of-plane states cannot be described by planar outlines. Use a planar representation when the relevant connections, endpoints, displacements, and boundaries provide sufficient interaction geometry. Base the decision on visible evidence rather than object names, dataset conventions, or rendering style, and report the observations and remaining uncertainty that support the route.

Object discovery and SAM 3 segmentation.

Inspect the entire video and indexed frame samples to enumerate every distinct physical entity in the interaction workspace. Include stationary obstacles, supports, and dividers that constrain the experiment, while excluding continuous environment surfaces, shadows, reflections, annotations, and rendering artifacts. Keep same-category instances separate and assign each entity a stable opaque identity. Classify its physical type as rigid, soft body, cloth, or rope from temporal evidence, distinguishing real deformation from occlusion, viewpoint change, and rigid rotation. For each entity, generate short image-grounded descriptions and select a seed frame that prioritizes a complete in-frame outline, visibility, and identity clarity. Accept a SAM 3 candidate only when it covers the intended foreground entity and excludes neighboring objects and background. If no candidate is adequate, refine the description or reselect the seed frame while preserving the target identity.

World probing (shared instructions).

Investigate the literal question before producing an answer and treat its options independently. Match the referenced entities, world, intervention, and time interval. Missing evidence is unknown rather than false, so inspect returned results before deciding that the evidence is sufficient. Use only the tools exposed by the selected route and issue one to six independent tool calls in a round. A rollout identifier must come from an earlier successful operation, and subsequent queries must read from the corresponding world. Follow returned guidance when evidence is missing and do not repeat an unchanged failed request. Termination or video-fallback requests are issued alone. Because the answer stage receives compact textual evidence, preserve every necessary visual observation together with the evidence reference from which it was obtained.

Evidence-based answer generation.

Answer the literal question using only its retained evidence. Match the referenced entities, intervention, and time window, and handle negated statements explicitly. Evaluate each relevant claim independently: unknown evidence does not imply a negative decision, and a claim that an event did not occur requires adequate coverage of the requested interval. Cite the evidence supporting the answer and provide a short factual reason without repeating the full tool history. Return one object that follows the specified output contract. If the retained evidence cannot determine the answer, return an explicit unresolved result that identifies the missing evidence rather than inventing a conclusion.