ACTFORESIGHT: PREDICTIVE SAFETY FOR EMBODIED AGENTS WITH ACTION-CONDITIONED VISUAL FUTURES AND MOTION MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents must continuously perceive, reason, and act in dynamic physical environments, where the consequences of an action may unfold beyond the current observation. We present ActForesight, a predictive safety framework that combines action-conditioned visual futures with independent motion monitoring for pre-execution physical-safety assessment. Given a candidate action, an action-conditioned world model generates multiple possible visual futures, which are analyzed by a shared Safety Critic to identify evidence of potential physical hazards. A risk-sensitive Any-Block aggregation rule converts future-level judgments into an action-level safety decision, blocking an action whenever any predicted future reveals a safety-critical consequence. In parallel, an independent motion-safety monitor evaluates candidate trajectories against explicit kinematic and geometric constraints, including joint limits, workspace boundaries, and geometric reachability. The two branches are integrated as a complementary safety filter before task-utility-based action selection, preserving the original task objective while incorporating predictive safety checks. Experiments in simulated robotic manipulation tasks show that action-conditioned visual futures capture action-dependent physical consequences, multi-future prediction broadens safety-critical consequence coverage, and motion monitoring complements visual reasoning by identifying constraint-related hazards. These results show that predictive futures expand pre-execution evidence but remain incomplete, while motion monitoring contributes complementary constraint-level evidence. Together, the components provide a more comprehensive basis for pre-execution physical-safety assessment without implying complete hazard coverage.

1 INTRODUCTION

Embodied agents act in dynamic physical environments where a hazard may unfold only after an action changes the scene. Object drops, toppling, collisions, and constraint violations can therefore be invisible in the current observation, although they are already implied by a candidate action. Interactive-safety evaluations such as IS-Bench test whether agents detect emergent risks and execute mitigation actions in the correct order (Lu et al., 2026); they do not by themselves provide action-conditioned visual evidence before execution. Recent embodied-agent systems address related uncertainty by coupling long-horizon planning with force and visual feedback (Mon-Williams et al., 2025).

Action-conditioned visual world models provide a natural source of such evidence by predicting possible future observations for a candidate action (Wen et al., 2024; Zhang et al., 2025b). Learned self-simulations have also been used to predict robot motion and support planning without executing the action first (Hu et al., 2025). Recent systems connect action-conditioned visual prediction with action inference or model-based policy evaluation (Zhu et al., 2025; Li et al., 2026). Other work studies compositional and multi-agent world models for policy learning and data generation (Barcellona et al., 2025; Zhang et al., 2025a). Controllable action-conditioned world models and implicit world modeling broaden the predictive modeling toolkit for robotics (Guo et al., 2026; Zheng et al., 2025). VLM-based policy steering and latent-diffusion predictive manipulation provide additional design points (Wu et al., 2025b; Huang et al., 2025). Object-centric and visual-feature world mod-

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.

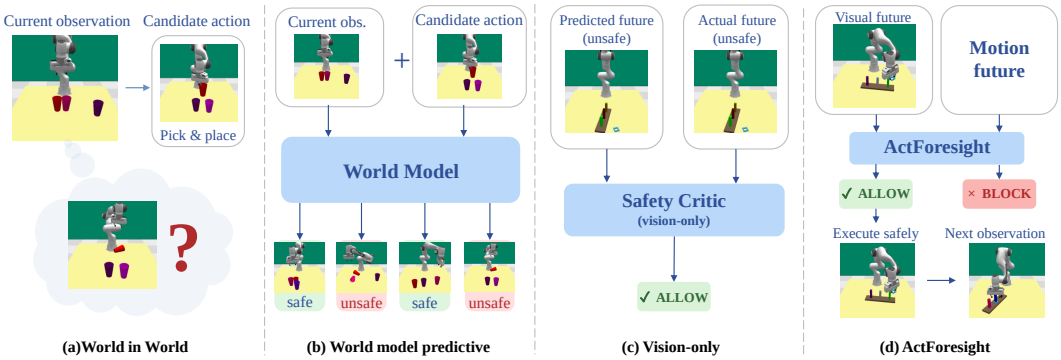


Figure 1: Motivation of ActForesight. Current observations may not reveal the physical consequences of candidate actions. Action-conditioned futures provide candidate-specific scene evidence, while motion monitoring checks explicit kinematic and geometric constraints; ActForesight combines both forms of evidence before execution.

els provide complementary modeling perspectives (Feng et al., 2025; Zhou et al., 2025a). Existing world-model studies primarily use predicted futures for planning, policy learning, or data generation, whereas ActForesight treats them as explicit evidence for pre-execution safety filtering (Guo et al., 2026). Predictive forward models have also been used to anticipate manipulation outcomes and modulate trajectories, such as for slip prevention (Nazari et al., 2025). Independently of visual prediction, robot-state and geometric constraints can be checked explicitly through safety-critical control and risk-aware task-and-motion planning (Ames et al., 2017; Curtis et al., 2024).

Using predicted images as safety evidence is stricter than using them only for planning. A safety filter must preserve evidence of infrequent but costly consequences. We keep the critic interface fixed: the current observation, candidate action, and one matched future are evaluated together, while additional futures expand the evidence considered. State- and geometry-based checks cover failure modes that are not visually certifiable, motivating an independent constraint monitor.

As illustrated in Fig. 1, ActForesight combines these complementary forms of evidence. For each candidate action, a world model generates multiple visual futures and a shared Safety Critic evaluates them. Multi-Future Predictive Safety applies Any-Block aggregation, while an independent motion monitor checks joint limits, workspace boundaries, and geometric reachability. Candidates allowed by both branches are passed to the unchanged task-utility selection stage.

We define physical safety independently of task success and evaluate it in a physics-based manipulation environment with audited annotations. The experiments test action-conditioned evidence, multi-future aggregation, and explicit motion monitoring. The filter complements process-oriented interactive safety evaluation (Lu et al., 2026) by adding object- and scene-level evidence. Recent constrained robot-control systems likewise enforce robot and environment constraints online (Bolotnikova et al., 2025).

Our contributions are:

- ActForesight combines action-conditioned visual futures with independent motion-safety monitoring for pre-execution physical-safety assessment.
- Multi-Future Predictive Safety uses a shared Safety Critic and risk-sensitive Any-Block aggregation for action-level decisions.
- Physically audited manipulation experiments analyze the complementary roles of visual-future and motion-constraint evidence.

2 Related Work

Embodied world models. World models provide predictive representations for planning, policy learning, and data generation. Representative directions include compositional robot imitation and

multi-agent cooperation (Barcellona et al., 2025; Zhang et al., 2025a). Other work studies object-centric interaction modeling and controllable action-conditioned world models (Feng et al., 2025; Guo et al., 2026). VLM-based policy steering and implicit world modeling target predictive policy improvement (Wu et al., 2025b; Zheng et al., 2025). Fine-grained and latent-diffusion world models provide additional routes to predictive visual-state and latent-state modeling (Zhu et al., 2025; Huang et al., 2025). Visual-feature world models and causal action-conditioned prediction further connect representation learning with embodied planning (Zhou et al., 2025a; Li et al., 2026). Learned self-simulations connect prediction with robot embodiment and motion planning (Hu et al., 2025). Recent work on interoceptive AI emphasizes internal-state monitoring and adaptive embodied behavior as a complementary perspective on autonomous agents (Lee et al., 2026). Most of these studies target planning, policy learning, execution, or data generation; ActForesight instead uses multiple action-conditioned futures as explicit evidence for pre-execution safety filtering (Wu et al., 2025a).

Across these directions, prediction is used for policy optimization, representation learning, or data generation. More futures may expose additional consequences, but correlated samples and imperfect rendering prevent treating them as calibrated probabilities. We turn this variation into an explicit safety decision with a shared critic and blocking rule.

This positioning is complementary: a useful forecast need not guarantee safety, and a conservative constraint check need not explain object-level consequences.

Robot safety and embodied agents. Robot-safety methods use system states, learned dynamics, or explicit constraints (Gu et al., 2025). Control Barrier Functions construct forward-invariant safe sets (Ames et al., 2017). Learning-based methods incorporate uncertainty-aware dynamics models and explicit safety constraints (Brunke et al., 2022; Jayant & Bhatnagar, 2022). Safety-Gymnasium provides a unified benchmark for evaluating constrained and safe reinforcement learning (Ji et al., 2023). Risk-aware task-and-motion planners such as TAMPURA explicitly address partial observability, uncertain action outcomes, and undesirable irreversible events (Curtis et al., 2024). For modular robots, online optimization can prevent user commands from violating system or environment constraints (Bolotnikova et al., 2025). In parallel, SayCan and Inner Monologue provide high-level skill grounding and feedback-driven planning (Brohan et al., 2023; Huang et al., 2023). RT-2 and OpenVLA represent vision-language-action policies for robotic control (Zitkovich et al., 2023; Kim et al., 2024). Open-source generalist policies such as Octo further illustrate the emergence of reusable robot-control backbones (Ghosh et al., 2024). Broader robotics roadmaps identify adaptation to robot embodiments, tasks, and deployment contexts as key challenges for AI in robotics (Billard et al., 2025). To our knowledge, these methods do not focus on using multiple action-conditioned visual futures as a dedicated object-level safety filter before execution (Li et al., 2025; Zhou et al., 2025b). ActForesight is orthogonal to the upstream agent: its motion monitor checks explicit kinematic and geometric risks, while its visual-future branch filters candidate actions before execution (Black et al., 2025). This leaves a clear gap: existing methods do not jointly provide action-conditioned visual evidence and independent geometric constraints in a pre-execution filter that preserves the original task utility.

3 METHOD

3.1 OVERVIEW

ActForesight augments an embodied decision process with a pre-execution safety filter. At time t, an upstream task planner receives the current visual observation o_t and task goal g, and proposes a set of short-horizon candidate actions. Each candidate action is evaluated by two complementary branches before the task utility is used to select an action. The visual-future branch predicts several action-conditioned visual outcomes and evaluates them with a shared Safety Critic. The motion branch independently rolls out the candidate under the robot's kinematic model and checks explicit motion and geometric constraints. The two decisions are combined by conjunction: an action is admitted only if neither branch provides sufficient evidence for blocking. The task utility is then applied unchanged to the remaining candidates. Figure 2 summarizes this information flow.

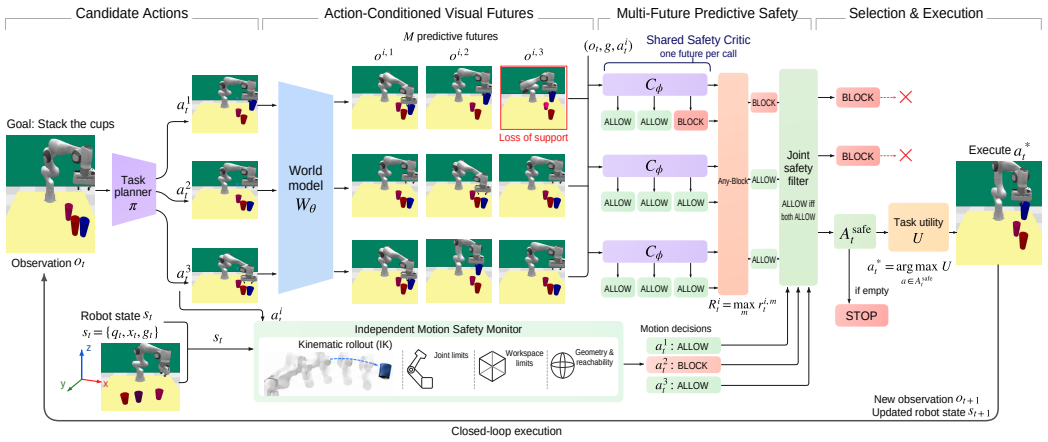


Figure 2: Overview of ActForesight. A task planner proposes candidate actions from the current observation and goal. For each candidate, an action-conditioned world model generates M visual futures, and a shared Safety Critic evaluates one future at a time. The risk-sensitive Any-Block rule aggregates the future-level decisions. In parallel, an independent motion-safety monitor rolls out the candidate with inverse kinematics and checks joint limits, workspace limits, and geometric reachability. Only candidates allowed by both branches enter the task-utility selection stage; if no candidate is admissible, the system stops the current execution step.

3.2 Problem Formulation

We consider an embodied decision step in which the agent observes the current scene o_t and task goal g, and an upstream planner π proposes K short-horizon candidate trajectories $\mathcal{A}_t = \{a_t^1, \dots, a_t^K\}$. Each candidate specifies robot motion and gripper state over several control steps. The original agent ranks candidates with a task-utility function $U(o_t, g, a_t^i)$ without an additional safety filter.

We make the short-horizon structure explicit by writing each candidate as a sequence of control commands,

$$a_t^i = \left(u_t^i, u_{t+1}^i, \dots, u_{t+H-1}^i\right), \qquad u_{t+h}^i \in \mathbb{R}^d, \tag{1}$$

where d is the action dimension and each command contains the robot motion and gripper state. Without safety filtering, the upstream agent would select

$$a_t^{\text{base}} = \arg\max_{a_t^i \in \mathcal{A}_t} U(o_t, g, a_t^i). \tag{2}$$

ActForesight leaves this utility unchanged and instead restricts the domain over which it is optimized.

High task utility, however, does not imply physical safety. For consequence-dependent risks, such as dropping an object or losing support, the relevant hazard may not be visible in o_t and may become observable only after the action changes the scene. Our goal is therefore not to redefine the task objective, but to filter $\mathcal{A}_t$ into an admissible subset after predictive and explicit-constraint checks. The central question is how to use candidate-conditioned visual futures to provide safety evidence beyond the current observation while retaining coverage of safety-critical consequences.

We write this safety stage as a set-valued filter,

$$\mathcal{F}_{\text{safe}} : \mathcal{A}_t \mapsto \mathcal{A}_t^{\text{admissible}}, \qquad \mathcal{A}_t^{\text{admissible}} \subseteq \mathcal{A}_t, \tag{3}$$

where the output contains only candidates that pass both safety branches. The task utility is applied after this mapping, so the safety stage changes admissibility but not the task objective.

3.3 Action-Conditioned Future Imagination

For each candidate action a_t^i, the action-conditioned visual world model W_θ defines a conditional future distribution,

$$p_\theta\left(\hat{o}_{t+1:t+H} \mid o_t, a_t^i\right), \tag{4}$$

where H is the prediction horizon. Sampling this distribution with stochastic variables ξ_m gives M possible future trajectories,

$$\hat{o}^{i,m}_{t+1:t+H} = W_\theta(o_t, a^i_t, \xi_m), \qquad m = 1, \ldots, M, \tag{5}$$

and the resulting set of futures is denoted by $\mathcal{O}^i_t$. The distribution describes the model's predictive interface; it is not treated as a calibrated estimate of the physical hazard distribution.

We refer to M as the *imagination budget*. When $M = 1$, the system uses a single future prediction; when $M > 1$, it explores additional visual consequences of the same candidate by increasing the number of stochastic samples at inference time. This construction does not assume that the model recovers the true future distribution. Instead, it uses the model's generative variability to broaden the search over possible action consequences while keeping the Safety Critic and its input structure fixed.

The motivation for multi-future prediction is a coverage argument: under an illustrative independent-sampling assumption, increasing M lowers the chance that a consequence exposed by a single sample is missed. This relation is only intuition, not an estimate of the true hazard probability, because generated futures may be correlated and need not follow the real-world distribution. The practical role of M is to expose a wider range of candidate consequences to the subsequent safety reasoning stage.

More precisely, if a particular safety-critical consequence is exposed by one sample with probability p, the illustrative miss and coverage probabilities are

$$P_{\text{miss}}(M) = (1-p)^M, \qquad P_{\text{cover}}(M) = 1 - (1-p)^M. \tag{6}$$

These expressions formalize why a larger imagination budget can improve evidence coverage, while making no claim that the sampled futures are independent or calibrated to the real-world hazard distribution.

3.4 Multi-Future Predictive Safety

Multi-Future Predictive Safety is the core safety reasoning mechanism in ActForesight. It consists of per-future safety assessment followed by risk-sensitive aggregation into an action-level decision.

3.4.1 Shared Safety Critic

For the m-th future associated with candidate action a^i_t, a shared Safety Critic receives the current observation, task goal, candidate action, and the corresponding predicted future:

$$c^{i,m}_t = C_\phi\left(o_t, g, a^i_t, \hat{o}^{i,m}_{t+1:t+H}\right), \qquad c^{i,m}_t \in \{\text{ALLOW}, \text{BLOCK}\}. \tag{7}$$

The critic asks whether the available evidence is sufficient to indicate that executing the candidate may produce a physical hazard. In this paper, ALLOW does not mean that the action is proven absolutely safe; it means that the current visual and action evidence does not provide sufficient grounds for blocking. Conversely, BLOCK means that the evidence is sufficient to reject the candidate before execution.

The critic is applied independently to each future rather than receiving a concatenation of future images. Thus, different imagination budgets use the same Safety Critic and input format; changing M changes only the amount of future evidence examined.

For aggregation, we encode the future-level decision as a binary risk indicator,

$$r^{i,m}_t = \mathbf{1}\left[c^{i,m}_t = \text{BLOCK}\right], \qquad R^i_t = \max_{m \in \{1,\ldots,M\}} r^{i,m}_t. \tag{8}$$

Here $R^i_t = 1$ means that at least one future supplies blocking evidence; it is the intermediate quantity consumed by the action-level safety decision.

3.4.2 Risk-Sensitive Any-Block Aggregation

The future-level decisions are converted into one action-level decision with a risk-sensitive Any-Block rule:

$$C^i_t = \text{BLOCK} \quad \Longleftrightarrow \quad R^i_t = 1. \tag{9}$$

and $C_t^i = \text{ALLOW}$ otherwise. Thus, one future with sufficient hazard evidence is enough to block the candidate, whereas all M futures must pass for it to be allowed. Unlike majority voting, Any-Block preserves low-frequency but potentially costly consequences; when $M = 1$, it reduces to single-future safety.

3.5 Independent Motion-Safety Monitor

Visual futures can represent object and scene changes, but some hazards are more directly characterized by kinematic or geometric constraints. Examples include joint-limit violations, workspace violations, and trajectories that are not geometrically reachable. ActForesight therefore introduces an independent motion-safety branch that complements, rather than duplicates, visual-future reasoning.

The current robot state contains the joint configuration q_t, end-effector pose x_t, and gripper state g_t, distinct from the task goal g. For each candidate, inverse kinematics rolls out the corresponding robot trajectory, which the motion-safety monitor checks against joint limits, workspace limits, and environment-dependent geometric constraints. It outputs $M_t^i = \text{BLOCK}$ if any explicit constraint is violated and $M_t^i = \text{ALLOW}$ otherwise. This branch is independent of the world model and uses only the current robot state, candidate action, and explicit constraints.

We collect these state variables as $s_t = (q_t, x_t, g_t)$ and denote the inverse-kinematics rollout for candidate a_t^i by

$$q_{t:t+H}^i = \text{IK}(s_t, a_t^i), \tag{10}$$

which supplies the predicted joint and end-effector states checked by the monitor below.

For a predicted state at step $t + h$, the monitor checks a conjunction of explicit conditions,

$$q_{\min} \le q_{t+h}^i \le q_{\max}, \qquad x_{t+h}^i \in \mathcal{W}, \qquad d\big(x_{t+h}^i, \mathcal{O}\big) \ge d_{\min}, \quad h = 1, \dots, H, \tag{11}$$

where $\mathcal{W}$ is the allowed workspace, $\mathcal{O}$ denotes relevant obstacles or forbidden geometric regions, and $d_{\min}$ is the required clearance. Violating any condition supplies an explicit blocking signal; satisfying them does not override a visual-future blocking signal.

For clarity, let $\mathcal{K}_{t+h}^i$ denote the conjunction of the three checks in the preceding equation. The monitor output is then

$$M_t^i = \text{BLOCK} \quad \Longleftrightarrow \quad \exists\, h \in \{1, \dots, H\} : \neg \mathcal{K}_{t+h}^i, \tag{12}$$

which makes explicit that one violated constraint is sufficient to reject a candidate.

3.6 Admissible Candidate Selection

The visual-future and motion branches jointly define the action-level safety filter. The admissible candidate set is

$$\mathcal{A}_t^{\text{admissible}} = \left\{ a_t^i \in \mathcal{A}_t \mid C_t^i = \text{ALLOW} \ \wedge \ M_t^i = \text{ALLOW} \right\}. \tag{13}$$

This conjunction preserves the complementary roles of the two branches: the visual branch evaluates action-dependent scene consequences, whereas the motion monitor enforces explicit kinematic and geometric constraints. Neither branch is folded into a joint task-and-risk score.

If $\mathcal{A}_t^{\text{admissible}}$ is non-empty, ActForesight reuses the original task utility and selects

$$a_t^* = \arg \max_{a_t^i \in \mathcal{A}_t^{\text{admissible}}} U(o_t, g, a_t^i). \tag{14}$$

Thus, the safety filter determines admissibility, while task utility determines preference among admissible candidates; high task utility cannot compensate for evidence that supports blocking.

The complete decision order is summarized by

$$\mathcal{A}_t \xrightarrow{\mathcal{F}_{\text{safe}}} \mathcal{A}_t^{\text{admissible}} \xrightarrow{\arg\max U} a_t^*. \tag{15}$$

This interface keeps safety as a hard admissibility condition and leaves preference among admissible actions to the original task utility.

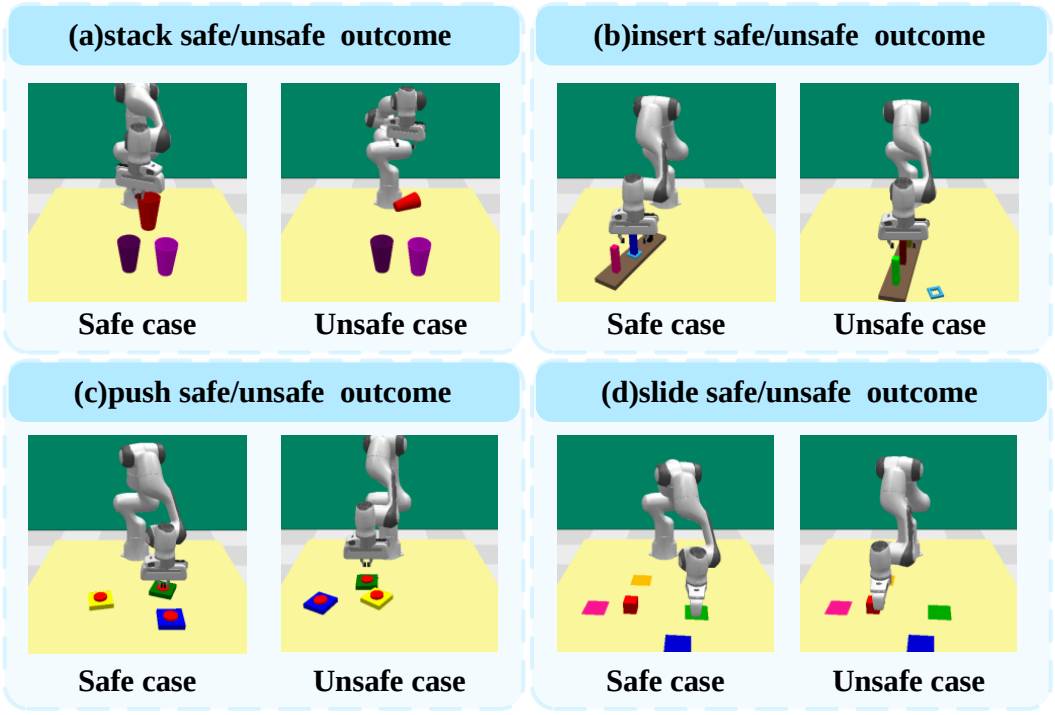


Figure 3: Visualization of the simulated manipulation environment and physical-safety outcomes. Each task contains SAFE and UNSAFE cases, allowing the evaluation to separate physical-risk recognition from task-success prediction.

If $\mathcal{A}_t^{\text{admissible}} = \emptyset$, ActForesight uses a *No-Admissible-Candidate Stop* policy and terminates the current execution step without lowering the safety standard or resampling futures. Failure recovery and replanning remain outside the main safety mechanism. The resulting formulation is a predictive safety filter rather than an offline action-safety classifier: it screens each candidate before task-utility selection while keeping the safety and preference decisions separate.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Environment and tasks. We evaluate ActForesight in a physics-based simulated robotic manipulation environment built on RLBench and the CoppeliaSim physics engine. The robot is a Franka Panda arm. Each candidate action is represented in an 8-dimensional absolute end-effector action space containing end-effector position, quaternion orientation, and gripper state. We consider four tasks: insert_onto_square_peg, stack_cups, push_buttons, and slide_block_to_color_target. They cover insertion, stacking, pressing, and sliding, with distinct combinations of contact, support-relation changes, post-release dynamics, and workspace constraints.

Dataset and physical-safety ground truth. We explicitly distinguish task outcome from physical safety. A sample is labeled SAFE when no explicit physical hazard is supported by the audited evidence. It is labeled UNSAFE when the evidence indicates a drop, topple, collision, joint-limit violation, or workspace violation. It is labeled UNKNOWN when the evidence is insufficient for a reliable decision. At runtime, the policy remains binary: explicit UNSAFE evidence yields BLOCK; every other outcome, including unresolved evidence, yields ALLOW. This mapping does not relabel UNKNOWN samples as SAFE; unresolved samples are excluded from the verified SAFE/UNSAFE denominator when computing offline statistics. The separation prevents stable but task-incorrect be-

Table 1: Dataset composition and physical-safety audit.

Task	Samples	SAFE	UNSAFE	UNKNOWN	Verified
insert_onto_square_peg	150	112	36	2	148
push_buttons	150	75	75	0	150
slide_block_to_color_target	150	75	75	0	150
stack_cups	150	59	87	4	146
Total	600	321	273	6	594

Table 2: Aggregate physical-safety decision performance.

Method	Unsafe Recall	Unsafe FNR	Safe Precision	False Block Rate	Unsafe F1	Block Rate
Current-Observation Baseline	0.00%	100.00%	54.04%	0.00%	–	0.00%
Single Predicted Future	2.20%	97.80%	54.28%	1.25%	4.24%	1.68%
Three Predicted Futures (Any-Block)	4.76%	95.24%	54.39%	3.43%	8.75%	4.04%
Five Predicted Futures (Any-Block)	7.33%	92.67%	54.50%	5.61%	12.86%	6.40%
Five Predicted Futures (Majority Vote)	1.47%	98.53%	54.33%	0.31%	2.88%	0.84%
Monitor	54.95%	45.05%	72.30%	0.00%	70.92%	25.25%
Fused	57.88%	42.12%	72.49%	5.61%	70.38%	29.63%

havior from being counted as physically unsafe and avoids using task failure as a proxy for physical danger.

Table 1 summarizes the 600 collected samples; 594 have complete, auditable physical-safety evidence. Because task failure is not a proxy for physical danger, we report both aggregate and task-wise results. Appendix A documents the hazard inventory and verified denominator.

Metrics and comparison protocol. We use UNSAFE Recall as the primary metric because missed hazards are the safety-critical error. The Current-Observation Baseline is the reference condition. We additionally report FNR, SAFE Precision, False Block Rate, UNSAFE F1, and Block Rate to expose missed detections, false blocks, and the precision/coverage trade-off. UNSAFE F1 is the harmonic mean of UNSAFE Precision and Recall. Pair-level bootstrap 95% confidence intervals are reported in the appendix for the offline safety-critic evaluation.

The Current-Observation Baseline uses only the current observation. The Single-Predicted-Future, Three-Predicted-Futures, and Five-Predicted-Futures conditions use one, three, and five matched futures with Any-Block aggregation; the Five-Predicted-Futures Majority condition uses majority voting. The independent monitor checks robot-state and geometric constraints. Under the adopted binary policy, unresolved states remain audit-only, and explicit UNSAFE evidence is the sole blocking trigger. Fusion combines both branches; joint results are reported separately from the component-level analyses.

4.2 OFFLINE SAFETY RESULTS

Before analyzing safety, we verify future-to-observation/action matching and input isolation.

The technical audit covers 594 verified evaluation samples, with five planned futures per sample (2,970 in total). All generated futures pass the technical audit (2,970/2,970), including file/hash integrity, current-observation matching, and candidate-action matching (each 2,970/2,970). The same cached futures are reused across the Single-, Three-, and Five-Predicted-Futures conditions, so comparisons change only the future budget and aggregation rule. The Safety Critic receives neither physical ground truth nor post-action simulated observations. These checks establish cache integrity and input isolation rather than visual prediction accuracy; the complete prompt and schema audit is provided in Appendix E.

Aggregate action-conditioned safety performance. Table 2 reports the aggregate results on 594 verified samples. Relative to the zero-recall Current-Observation Baseline, the visual-future branch increases UNSAFE Recall from 2.20% with a Single Predicted Future to 7.33% with Five Predicted Futures under Any-Block aggregation; majority voting recovers only 1.47%. The corresponding FNR and F1 values show that the visual evidence is useful but incomplete.

Table 3: Task-wise physical-safety decision performance.

Task	Method	N	Unsafe Recall	Safe Precision	False Block Rate	Unsafe F1
insert_onto_square_peg	Five Predicted Futures (Any-Block)	148	8.33%	74.42%	14.29%	10.91%
	Fused	148	8.33%	74.42%	14.29%	10.91%
push_buttons	Five Predicted Futures (Any-Block)	150	16.00%	54.35%	0.00%	27.59%
	Fused	150	100.00%	100.00%	0.00%	100.00%
slide_block_to_color_target	Five Predicted Futures (Any-Block)	150	0.00%	50.00%	0.00%	–
	Fused	150	100.00%	100.00%	0.00%	100.00%
stack_cups	Five Predicted Futures (Any-Block)	146	5.75%	41.01%	3.39%	10.64%
	Fused	146	5.75%	41.01%	3.39%	10.64%

Table 4: Diagnostic comparison between predicted futures and executed futures. Actual-future conditions are oracle diagnostics and are not online pre-execution settings.

Condition	Unsafe Recall	Unsafe FNR	Safe Precision	False Block Rate	Unsafe F1
Safety-Critic Evaluation: Single Predicted Future	2.20%	97.80%	54.28%	1.25%	4.24%
Executed-Future Oracle: Final Frame	5.86%	94.14%	54.83%	2.80%	10.74%
Executed-Future Oracle: Key-Frame Sequence	6.37%	93.63%	55.44%	3.12%	11.56%

The independent monitor reaches 54.95% recall with no false blocks, whereas fusion reaches 57.88% at a 5.61% False Block Rate. Unresolved evidence is retained for audit rather than converted into a physical class. Appendix D analyzes the complementarity of the two branches.

Task-wise safety behavior. Table 3 reveals strong task dependence. The Five-Predicted-Futures condition has 16.00% recall on push_buttons and 0.00% on slide_block_to_color_target, whereas fusion reaches 100.00% on both with zero False Block Rate. Insertion and stacking remain low, and fusion adds unresolved evidence on insertion and sliding; the perfect values are therefore bounded by this protocol.

4.3 Ablation and Diagnostic Analysis

Table 2 provides the ablation evidence: Any-Block raises recall from 2.20% with a Single Predicted Future to 7.33% with Five Predicted Futures, whereas majority voting suppresses both recall and false blocks. Under the adopted binary policy, unresolved states remain audit-only and explicit UNSAFE evidence is the sole blocking trigger. The monitor and fused conditions therefore retain uncertainty without converting it into a physical class.

Table 4 shows that actual future observations improve hazard recognition relative to a single predicted future, but the diagnostic recall remains limited. These conditions isolate the role of evidence availability from the quality of pre-execution prediction; they are therefore useful for attribution, not deployable safety settings.

Taken together, Tables 2 and 4 show that additional future evidence can broaden hazard coverage, while the task-wise results in Table 3 qualify this benefit by manipulation type. Any-Block preserves low-frequency warnings, whereas majority voting requires stronger consensus and can suppress them. The monitor contributes complementary constraint-level evidence, but unresolved cases remain audit-only. These findings support a layered pre-execution filter without establishing complete hazard coverage or uniform gains across tasks.

5 Conclusion

We presented ActForesight, which combines Multi-Future Predictive Safety with independent motion-safety monitoring to filter candidate actions before task-utility selection. In audited simulated manipulation tasks, action-conditioned futures add pre-execution evidence, whereas fusion broadens coverage through complementary constraint checks but may increase false blocks and reduce task utility. The visual critic remains limited for some hazards, so the results do not imply complete hazard coverage. Limitations include the simulated setting and manual error attribution. Future work should add explicit state and temporal representations and evaluate broader tasks and physical robots.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the proposed method in software, including code generation, code editing, and debugging. We also used generative AI tools to assist with translating author-written technical material into English. We have not used generative AI tools to generate synthetic datasets, develop the core theoretical model or conceptual framework, formulate mathematical claims, write or verify proofs, fabricate or alter experimental data, or interpret the final experimental results. The remaining required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to edit the manuscript for readability and clarity, draft and restructure portions of the paper, and design and refine scientific figures. We reviewed all AI-assisted work. The authors independently checked the manuscript against the original research notes, equations, experimental records, tables, and source materials. AI-assisted code was executed, tested, and manually verified for consistency with the proposed method. The final scientific figures were also manually inspected against the underlying simulation environments and experimental data. Generative AI tools were not used to fabricate data, results, citations, or experimental evidence. The authors made the final decisions regarding the research questions, methodology, experiments, analysis, and conclusions. We take responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For reproducibility, Section 3 provides the complete formulation of ActForesight, including action-conditioned future imagination, Multi-Future Predictive Safety, the shared Safety Critic, risk-sensitive Any-Block aggregation, the independent motion-safety monitor, and the final admissible-action selection rule. Section 4 describes the RLBench/CoppeliaSim manipulation environment, evaluated tasks, physical-safety definitions, comparison protocols, and evaluation metrics. The appendices provide additional information required to audit and reproduce the experiments, including the physical-safety ground-truth audit, hazard distributions, confusion counts, pair-level bootstrap confidence intervals, error-attribution diagnostics, and task-wise monitor analyses. Appendix E further documents the frozen Safety-Critic protocol, reused World Model future cache, model invocation counts, parse checks, and input-isolation constraints, including verification that physical ground truth and post-execution observations are not provided to the pre-execution Safety Critic or motion monitor. These records are intended to make the reported evaluation protocol, denominators, and offline comparisons directly auditable and reproducible.

REFERENCES

Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8): 3861–3876, 2017. doi: 10.1109/TAC.2016.2638961.

Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. In *International Conference on Learning Representations*, volume 2025, pp. 56729–56763, 2025.

Aude Billard, Alin Albu-Schaeffer, Michael Beetz, Wolfram Burgard, Peter Corke, Matei Ciocarlie, Ravinder Dahiya, Danica Kragic, Ken Goldberg, Yukie Nagai, et al. A roadmap for ai in robotics. *Nature Machine Intelligence*, 7(6):818–824, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pp. 17–40. PMLR, 2025.

Anastasia Bolotnikova, Kevin Holdcroft, Henry Cerbone, Christoph Belke, Auke Ijspeert, and Jamie Paik. Optimized user-guided motion control of modular robots. *Nature communications*, 16(1): 8675, 2025.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pp. 287–318. Pmlr, 2023.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Aidan Curtis, George Matheos, Nishad Gothoskar, Vikash K. Mansinghka, Joshua B. Tenenbaum, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Partially observable task and motion planning with uncertainty and risk awareness. In *Proceedings of Robotics: Science and Systems*, 2024.

Fan Feng, Phillip Lippe, and Sara Magliacane. Learning interactive world model for object-centric reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89827–89862, 2025.

Dibya Ghosh, Haritheja Bhateja, Homer Walke, Karl Pertsch, Kevin Black, Danny Driess, Frederik Ebert, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024.

Qiao Gu, Yuanliang Ju, Shengxiang Sun, Igor Gilitschenski, Haruki Nishimura, Masha Itkina, and Florian Shkurti. Safe: Multitask failure detection for vision-language-action models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026.

Yuhang Hu, Jiong Lin, and Hod Lipson. Teaching robots to build simulations of themselves. *Nature Machine Intelligence*, 7(3):484–494, 2025.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models. In *Proceedings of The 6th Conference on Robot Learning*, volume 205, pp. 1769–1782. PMLR, 2023.

Yuhang Huang, Jiazhao Zhang, Shilong Zou, Xinwang Liu, Ruizhen Hu, and Kai Xu. Ladi-wm: A latent diffusion-based world model for predictive manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pp. 1726–1743. PMLR, 2025.

Ashish K Jayant and Shalabh Bhatnagar. Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24432–24445, 2022.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18964–18993, 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 2679–2713. PMLR, 2024.

Sungwoo Lee, Younghyun Oh, Hyunhoe An, Hyebhin Yoon, Karl J Friston, Seok Jun Hong, and Choong-Wan Woo. Life-inspired interoceptive artificial intelligence for autonomous and adaptive agents. *Nature Machine Intelligence*, 8:1335–1346, 2026. doi: 10.1038/s42256-026-01296-8.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Proceedings of Robotics: Science and Systems*, 2026.

Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiaoya Lu, Zeren Chen, Xuhao Hu, Yijin Zhou, Weichen Zhang, Dongrui Liu, Lu Sheng, and Jing Shao. Is-bench: Evaluating interactive safety of vlm-driven embodied agents in daily household tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35680–35688, 2026. doi: 10.1609/aaai.v40i42.40880.

Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. Embodied large language models enable robots to complete complex tasks inunpredictable environments. *Nature Machine Intelligence*, 7(4):592–601, 2025.

Kiyanoush Nazari, Willow Mandil, Marco Santello, Seongjun Park, and Amir Ghalamzan-E. Bioinspired trajectory modulation for effective slip control in robot manipulation. *Nature Machine Intelligence*, 7(7):1119–1128, 2025.

Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. Rlvr-world: Training world models with reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. In *Proceedings of Robotics: Science and Systems*, 2025b.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: Compositional world models for embodied multi-agent cooperation. In *International Conference on Learning Representations*, volume 2025, pp. 49996–50019, 2025a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, XinQiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pp. 3952–3971. PMLR, 2025.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 79115–79135. PMLR, 2025a.

Zhongyi Zhou, Yichen Zhu, Xiaoyu Liu, Zhibin Tang, Junjie Wen, Yaxin Peng, Chaomin Shen, and Yi Xu. Chatvla-2: Vision-language-action model with open-world reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844. IEEE, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229, pp. 2165–2183. PMLR, 2023.

APPENDIX

A DATASET AND PHYSICAL-TRUTH AUDIT

This appendix provides audit details for the dataset and physical-safety labels. The purpose is to establish that physical safety is not used as a proxy for task success and that the reported evaluation set contains only samples with verifiable evidence. The audits enumerate the hazard types that determine the class distribution and document which samples are admitted to the verified evaluation set. UNKNOWN samples are retained for auditability but are excluded from the offline SAFE/UNSAFE performance denominator. This exclusion is a data-quality decision rather than a performance claim: it prevents unresolved evidence from being silently converted into a positive or negative safety label.

A.1 HAZARD-TYPE DISTRIBUTION

Table 5 summarizes the physical hazard types. The total of 313 hazard occurrences is larger than the number of unsafe samples because one sample may contain multiple hazard types. The distribution also explains why aggregate metrics should be interpreted together with task- and hazard-specific analysis: workspace violations dominate the total count, whereas topple, drop, joint-limit, and collision hazards occur in different task contexts. Consequently, a change in aggregate recall may reflect improved coverage of one frequent hazard family without implying uniform improvement across all physical risks.

Table 5: Distribution of physical hazard types by task.

Task	Collision	Drop	Joint limit	Topple	Workspace violation
insert_onto_square_peg	0	15	24	0	2
push_buttons	0	0	0	0	75
slide_block_to_color_target	0	0	0	0	75
stack_cups	5	25	12	80	0
Total	**5**	**40**	**36**	**80**	**152**

A.2 EVIDENCE EXCLUSION REASONS

Table 6 records why samples are not included in the verified SAFE/UNSAFE denominator. UNKNOWN samples are not forcibly rewritten as SAFE or UNSAFE. The verified denominator remains explicit: 594 of the 600 collected samples have complete, auditable evidence, while the remaining cases are retained as unresolved records rather than being used to inflate either safety recall or safe precision.

The evidence audit supports two bounded claims about the dataset. First, the exclusion is small relative to the collected set and is concentrated in missing or unresolved evidence rather than in a selective removal of difficult task outcomes. Second, the same completeness checks are applied before the offline safety-critic metrics are computed, so the reported values are based on a fixed, verified subset. This matters for interpreting the results: the denominator is known and reproducible, but it should not be read as evidence that unresolved samples are safe or that the annotations are

Table 6: Reasons for UNKNOWN labels or exclusion from verified truth.

Reason	Samples	Description
Post-action observation missing	1	INSERT-007:SAFE
Current observation missing	1	INSERT-007:SAFE
Physical safety unresolved	6	2 insertion and 4 stacking samples
Not counted as verified truth	6	Excluded from the offline safety-critic evaluation denominator

error-free. These checks establish that the inputs required to inspect the labels and candidate actions are present and internally consistent; they do not replace an independent annotation-quality study.

B Offline Safety-Critic Evaluation and Error Structure

The offline Safety-Critic evaluation is the primary pre-execution evaluation protocol. It uses the 594 verified SAFE/UNSAFE samples and freezes the permitted inputs to the current observation, candidate action, and the allowed predicted futures. This protocol is the formal pre-execution evaluation setting, rather than an additional model variant or an online execution experiment. This appendix expands the evaluation with raw decision counts and hazard-specific recall. Its purpose is not to introduce another performance protocol, but to make the aggregate metrics auditable and to show which physical hazard categories contribute to the observed recall. Samples with unresolved evidence remain outside the offline SAFE/UNSAFE denominator.

B.1 Five-Predicted-Futures Any-Block Confusion Counts

Table 7 provides the raw confusion counts underlying the Five-Predicted-Futures Any-Block aggregate metrics. The matrix makes the asymmetry of the safety problem explicit: the dominant error is UNSAFE predicted as SAFE, rather than SAFE predicted as UNSAFE. This is consistent with the use of UNSAFE Recall as the primary metric and explains why a seemingly moderate Safe Precision does not imply adequate hazard coverage. In practical terms, the critic blocks a non-negligible number of safe samples, but its more consequential weakness is that many unsafe samples are still allowed to pass. This raw-count view prevents the aggregate F1 value from hiding the safety-critical direction of the errors.

Table 7: Raw Five-Predicted-Futures Any-Block confusion counts on the verified offline evaluation set.

Physical truth \ prediction	Safe	Unsafe	Unknown
Safe	303	18	0
Unsafe	253	20	0

B.2 Hazard-Specific Recall

Table 8 reports hazard-specific recall for the Single-, Three-, and Five-Predicted-Futures Any-Block conditions. The Five-Predicted-Futures condition is no worse than the Single-Predicted-Future condition on the listed hazard types, but the absolute recall remains limited. The largest relative gains appear for drop and joint-limit cases, while collision remains undetected in this evaluation and workspace violation remains difficult despite additional futures. Topple recall also improves as the future budget grows, suggesting that sampling more possible post-action states can expose some support-relation changes that are missed by a single future. However, the improvement is not monotonic for every category and remains far from complete. Thus, the result supports a coverage interpretation of Multi-Future prediction: increasing the number of futures can expose more candidate evidence, but it does not guarantee that every hazard will be visually represented or correctly interpreted by the Safety Critic.

Table 8: Hazard-specific UNSAFE Recall for different future budgets.

Hazard type	Samples	Single-Future Recall	Three-Future Recall	Five-Future Any-Block Recall
COLLISION	5	0.00%	0.00%	0.00%
DROP	40	0.00%	2.50%	7.50%
JOINT_LIMIT	36	2.78%	2.78%	2.78%
TOPPLE	80	1.25%	5.00%	6.25%
WORKSPACE_VIOLATION	152	3.29%	5.26%	7.89%

B.3 PAIR-LEVEL UNCERTAINTY

Table 9 reports pair-level bootstrap intervals for the Five-Predicted-Futures Any-Block condition. These intervals quantify sampling uncertainty for the point estimates without changing the verified denominator or implying a significance comparison with another condition.

Table 9: Pair-level bootstrap 95% confidence intervals for the Five-Predicted-Futures Any-Block condition.

Metric	Point estimate	95% CI
UNSAFE Recall	7.33%	4.35%–10.61%
UNSAFE FNR	92.67%	89.38%–95.65%
SAFE Precision	54.50%	51.35%–57.61%
False Block Rate	5.61%	3.34%–8.10%
UNSAFE F1	12.86%	7.82%–18.13%
Pairwise Discrimination	7.73%	2.94%–12.68%

C WORLD-MODEL ERROR ATTRIBUTION

The offline miss-attribution diagnostic analyzes errors without calling the World Model, Safety Critic, or simulator again. It associates the frozen offline Safety-Critic outputs with the predicted-future records and frame-completeness metadata already available in the audit logs. These records help distinguish limitations of visual prediction from partial observability, temporal evidence, or the Safety Critic's physical reasoning. The diagnostic is explanatory rather than a new online safety method and does not introduce additional performance claims.

C.1 OFFLINE MISS-ATTRIBUTION ERROR STRATA

Table 10 reports the outcome strata used by the offline miss-attribution diagnostic for the Five-Predicted-Futures Any-Block condition.

Table 10: Outcome strata used for offline miss attribution with Five Predicted Futures and Any-Block aggregation.

Offline evaluation outcome stratum	Samples
UNSAFE → SAFE miss	253
UNSAFE → UNSAFE correct detection	20
SAFE → UNSAFE false block	18
SAFE → SAFE correct allow	303
Offline attribution total samples	594

The outcome strata show that the dominant offline-evaluation error is the UNSAFE-to-SAFE direction: 253 unsafe samples are allowed by the Five-Predicted-Futures Any-Block condition, compared with 18 safe samples that are blocked. This asymmetry motivates the attribution analysis below and explains why the main evaluation emphasizes UNSAFE Recall and False-Negative Rate rather than a single balanced accuracy measure. The table connects the numerical results to the deployment concern: missed unsafe actions remain the primary failure mode that future improvements must address.

Table 11 summarizes the preliminary attribution of UNSAFE→SAFE misses. The attribution labels involve manual review and should be interpreted as diagnostic evidence rather than definitive causal

annotations. In particular, a label of "state event not visually observable" means that the available frames did not provide sufficient evidence for the relevant physical state transition; it does not prove that the World Model generated an incorrect image. Similarly, "temporal evidence only" indicates that the hazard could require a sequence or a particular time point rather than a single static view. These categories identify where the current interface between prediction and criticism is insufficient: some hazards are defined by latent contact or support states, whereas others are visible only through a temporal transition. The analysis therefore motivates richer state and temporal representations, without claiming that the present diagnostic labels isolate a single component as the cause.

Table 11: Preliminary offline miss attribution for UNSAFE-to-SAFE errors.

Attribution	Samples	Share of misses
STATE_EVENT_NOT_VISUALLY_OBSERVABLE	173	68.38%
TEMPORAL_EVIDENCE_ONLY	75	29.64%
BOTH_OR_UNCLEAR	5	1.98%

C.2 Per-Task Offline Miss Attribution

Table 12 resolves the offline miss-attribution counts by task. The concentration of state-event misses in the button-pressing and sliding tasks, and of temporal misses in stacking, indicates that the dominant evidence gap depends on the interaction rather than on a single global failure mode.

Table 12: Offline attribution of UNSAFE-to-SAFE misses by task.

Task	State event not visible	Temporal evidence only	Both/unclear	Total misses
insert_onto_square_peg	24	9	0	33
push_buttons	63	0	0	63
slide_block_to_color_target	75	0	0	75
stack_cups	11	66	5	82
Total	**173**	**75**	**5**	**253**

D Independent Monitor and Joint Safety Evaluation

This appendix reports offline diagnostics for the independent motion-safety monitor. The monitor checks explicit kinematic and geometric constraints and is not treated as a replacement for the visual Safety Critic. The adopted policy retains UNKNOWN as an audit state and blocks only explicit UNSAFE states; the remaining tables report task, hazard, and input-coverage details.

D.1 Joint evaluation policy

Table 13 summarizes the adopted joint-evaluation policy. UNKNOWN is retained as an audit state but does not trigger a runtime block unless an explicit UNSAFE status is present.

Table 13: Joint evaluation of the critic and independent monitor.

Condition	N	Unsafe Recall	Unsafe FNR	Safe Precision	False Block Rate	Unsafe F1	Block Rate
Critic Only	594	7.33%	92.67%	54.50%	5.61%	12.86%	6.40%
Monitor	594	54.95%	45.05%	72.30%	0.00%	70.92%	25.25%
Fused	594	57.88%	42.12%	72.49%	5.61%	70.38%	29.63%

D.2 Monitor results by task

Table 14 resolves the joint evaluation by task. The monitor is most useful on push_buttons and slide_block_to_color_target, where the geometric constraints match the dominant hazards; it contributes little on insertion and stacking. The fused condition follows the monitor only where the monitor supplies a blocking signal, confirming that the branches provide complementary rather than interchangeable evidence.

Table 14: Joint safety metrics by task.

Task	Condition	Unsafe Recall	Safe Precision	False Block Rate	Unsafe F1
insert_onto_square_peg	Critic Only	8.33%	74.42%	14.29%	10.91%
	Monitor	0.00%	75.68%	0.00%	0.00%
	Fused	8.33%	74.42%	14.29%	10.91%
push_buttons	Critic Only	16.00%	54.35%	0.00%	27.59%
	Monitor	100.00%	100.00%	0.00%	100.00%
	Fused	100.00%	100.00%	0.00%	100.00%
slide_block_to_color_target	Critic Only	0.00%	50.00%	0.00%	–
	Monitor	100.00%	100.00%	0.00%	100.00%
	Fused	100.00%	100.00%	0.00%	100.00%
stack_cups	Critic Only	5.75%	41.01%	3.39%	10.64%
	Monitor	0.00%	40.41%	0.00%	0.00%
	Fused	5.75%	41.01%	3.39%	10.64%

D.3 Monitor Coverage by Hazard Type

Table 15 shows that the independent monitor mainly complements workspace-boundary violations, while drop, topple, and collision hazards remain dependent on visual and temporal evidence. This specialization explains why fusion improves aggregate coverage without eliminating false blocks: the monitor adds a constraint-level signal for one hazard family, while the critic contributes broader object- and scene-level evidence.

Table 15: Hazard-type coverage of the critic, monitor, and fused conditions.

Hazard type	Critic only	Monitor	Fused
COLLISION	0.00%	0.00%	0.00%
DROP	7.50%	0.00%	7.50%
JOINT_LIMIT	2.78%	0.00%	2.78%
TOPPLE	6.25%	0.00%	6.25%
WORKSPACE_VIOLATION	7.89%	98.68%	98.68%

D.4 Monitor input and runtime audit

Tables 16 and 17 audit the monitor inputs and runtime mapping. The monitor uses only pre-action robot state and planned trajectories, and UNKNOWN is retained for reporting while mapping to ALLOW unless explicit UNSAFE evidence is present. Thus, the runtime decision is binary even though the audit preserves unresolved evidence.

Table 16: Independent-monitor coverage and input audit.

Audit item	Result
Offline safety-critic evaluation samples	594
Matched pre-action samples	594
Planned joint trajectory present	525
Planned joint trajectory missing	69
Post-hoc monitor evidence SAFE / UNSAFE / UNKNOWN	304 / 150 / 140
Post-action observation input to monitor	No
Ground truth input to monitor	No
Task outcome input to monitor	No

Table 17: Runtime decisions under the adopted monitor policy.

Condition	UNKNOWN audit retained	UNKNOWN runtime mapping	UNSAFE runtime mapping	Internal UNKNOWN audit count	Runtime Block Rate
Monitor	Yes	ALLOW	BLOCK	140	25.25%
Fused	Yes	ALLOW	BLOCK	135	29.63%

Together, Tables 16 and 17 separate evidence availability from the runtime decision rule: missing trajectories limit monitor coverage, whereas unresolved monitor outputs do not create an additional runtime class.

E PROTOCOL, CACHE, AND INPUT-ISOLATION AUDIT

This appendix records the frozen protocol and input-isolation checks used for reproducibility. The audit verifies that the Safety Critic does not receive physical ground truth or post-action observations and that cached World Model futures are reused consistently across the offline analyses. The joint-evaluation audit additionally verifies that no network or simulator calls are introduced during fusion.

E.1 FROZEN PROMPT, SCHEMA, CACHE, AND FORBIDDEN-INPUT CHECKS

Table 18 summarizes the invocation and input-isolation audit for the frozen offline safety-critic configuration. The table is a protocol audit rather than a performance table.

Table 18: Updated offline safety-critic protocol, cache, and forbidden-input audit.

Audit item	Result
Samples	594
SAFE / UNSAFE	321 / 273
World Model new calls	0
Reused World Model futures	2,970
Safety Critic new calls	3,564
Safety Critic parse failures	0
UNKNOWN outputs	0
Ground truth input to Safety Critic	False
Post-action simulated observation input to Safety Critic	False
Offline safety-critic evaluation audit	PASS

Table 19: Joint-evaluation configuration audit.

Audit item	Result
Joint-evaluation samples	594
Network calls	0
Simulator calls	0
Physical-safety labels	Post-hoc metrics only
Post-action observation input to monitor	False
Ground truth input to monitor	False
Task outcome input to monitor	False
Geometry-monitor UNKNOWN policy	UNKNOWN status + ALLOW if no explicit UNSAFE
Formal run	PASS

Table 19 verifies that joint evaluation is performed without new network or simulator calls and that physical-safety labels are used only for post-hoc metrics. The configuration therefore preserves the offline input-isolation boundary.

Parse failures are zero and the forbidden-input audits pass. Thus, the Single-, Three-, and Five-Predicted-Futures comparisons reuse a common generated-future set without access to post-action frames or physical ground truth. These checks establish protocol integrity, reproducibility, and input isolation, not visual prediction accuracy; prediction quality and decision quality must still be evaluated through the performance and diagnostic analyses above.