

Predictive Safety in Embodied AI: From Control-Theoretic Filters to Foundation Models

Kleyton da Costa, Adriano Koshiyama, Dimitrios Kanoulas, and Philip Treleaven

Abstract—Predictive safety links anticipated consequences to interventions that prevent or limit harm. This review examines that link across control-theoretic filters, constrained learning, learned world models, and vision-language-action policies. A structured review organizes 288 candidate records, with 26 full-text inspections and 262 AI-assisted abstract assessments. We distinguish conditional invariance and probabilistic guarantees from calibrated prediction and empirical cost reduction, and organize the literature by safety objective, assurance mechanism, representation, platform, coordination, and evaluation setting. Three recurring failures structure the synthesis: representations omit hazard-relevant information, predictions become unreliable under deployment shift, or warnings leave insufficient time or authority to intervene. We compare representative methods and benchmarks and propose evaluation records that report task success alongside harm-specific violations, uncertainty, and intervention behavior. Effective predictive safety requires a grounded specification, reliable prediction, and an executable safeguard within a documented operating domain.

Index Terms—Robot safety, predictive safety filters, world models, safe reinforcement learning, vision-language-action models.

I. INTRODUCTION

EMBODIED agents turn computational decisions into physical consequences. A navigation policy can enter a region from which braking is no longer sufficient; a manipulator can select a successful grasp that damages its surroundings; and a language-conditioned robot can correctly follow an instruction whose intended outcome is unsafe. Safety therefore requires more than task competence. It requires a specification of unacceptable outcomes, an account of what the controller can know before acting, and a mechanism capable of preventing or limiting those outcomes.

We use *predictive safety* to describe the use of anticipated consequences when assessing or modifying an embodied agent’s actions. The anticipation may involve a trajectory, a reachable set, a future violation probability, or a learned value function that summarizes longer-term outcomes. This definition does not require photorealistic simulation. Nor does it imply that every safety method explicitly rolls out candidate futures. A control barrier function (CBF), for example, can constrain the local state derivative to preserve a safe set. Its relationship to future safety comes from the invariance argument, not necessarily from a generated trajectory.

The foundations of this approach precede current foundation models. Predictive safety filters construct feasible backup trajectories for learning-based controllers, while CBF formulations already address automotive cruise control and lane keeping [1], [2]. Recent learned models make it possible to anticipate consequences from richer observations and to propose actions conditioned on language. Their promise motivates a comparison with established assurance mechanisms, rather than a replacement of the assumptions on which those mechanisms depend.

A central question is therefore: *when does a prediction justify a safety intervention, and when does that intervention justify a claim about the deployed robot?* An accurate-looking video is insufficient if it omits contact force. A well-calibrated risk score is insufficient if calibration fails under the encountered shift. A warning is insufficient if no feasible action remains. Conversely, a conservative analytical filter may provide a meaningful guarantee for a limited specification even when it lacks broad semantic understanding.

The literature spans control, reinforcement learning (RL), formal verification, robotics, and foundation-model research. These communities often use the word “safe” for different properties: never leaving a specified set, bounding a violation probability, keeping expected cumulative cost below a threshold, or reducing observed unsafe behavior. These statements are not interchangeable. The review makes their assumptions and evidence explicit and uses them to organize the comparison.

The review makes four contributions: a systems account linking specification, prediction, action, and recovery; a taxonomy that separates safety objectives from mechanisms and representations; a comparison of control, learned-model, and vision-language-action (VLA) approaches by evidence strength; and a reproducible literature audit with an evaluation agenda. The contribution is a synthesis of existing evidence; no new controller or robot experiment is reported.

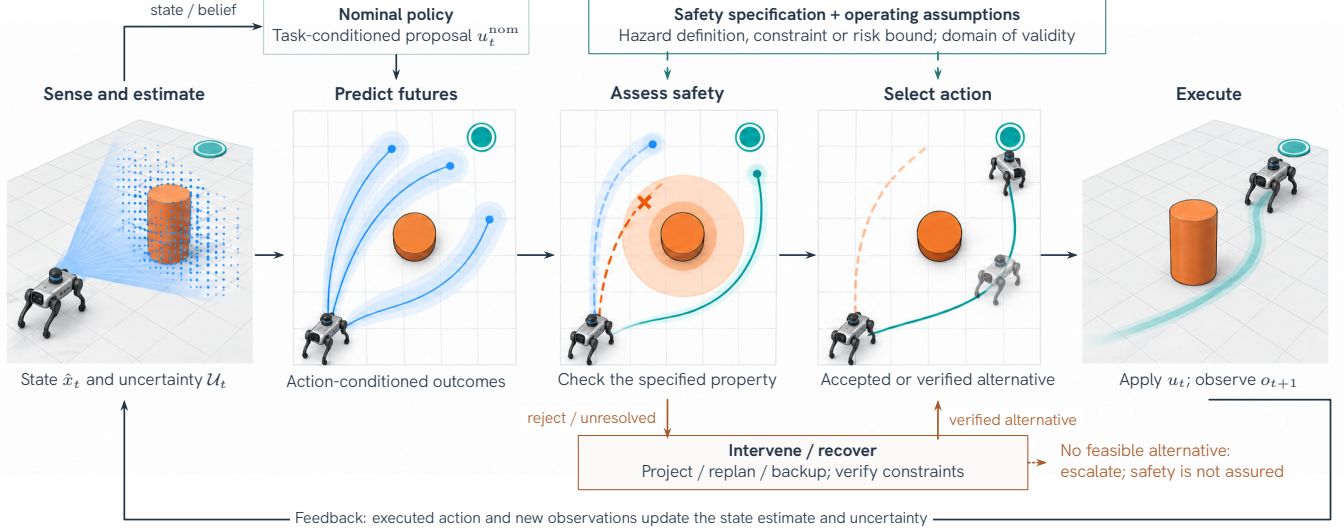
II. RELATIONSHIP TO EXISTING REVIEWS

Safe learning in robotics already has a substantial synthesis connecting control-theoretic models, uncertainty, learning, and certification [3]. Reviews of safe RL and constraint formulations further explain why an expected-cost constraint, a chance constraint, and a statewise constraint lead to different algorithms [4], [5], [6]. Krasowski et al. distinguish action replacement, projection, and masking and compare their behavior on control tasks [7]. These contributions establish the foundations of our intervention vocabulary.

Recent reviews explicitly address the safety of embodied foundation models. The world-model safety review examines

Kleyton da Costa is with University College London and Holistic AI, London, UK (corresponding author: kleyton.costa.25@ucl.ac.uk). Adriano Koshiyama is with Holistic AI. Dimitrios Kanoulas and Philip Treleaven are with University College London.

Predictive safety: from observation to physical action



Three conditions for an enforceable safety claim



Hazard-relevant state
Retain variables that determine harm.
Represent missing information.



Valid prediction
Check model error and calibration.
Reassess assumptions under shift.



Feasible, timely intervention
Maintain an executable alternative.
Meet actuation and timing limits.

Illustrative navigation scenario, not experimental data. Blue: prediction. Orange: hazard / rejected route. Teal: admissible route / goal.

Fig. 1. Predictive safety from observation to action. A nominal policy proposes u_t^{nom} from the estimated state $\hat{x}_t$ and uncertainty $\mathcal{U}_t$. Candidate futures (blue) are assessed against a specified hazard (orange); the teal route illustrates an admissible alternative. Rejected or unresolved candidates trigger intervention, and alternatives require safety and feasibility checks before execution. Applied inputs u_t and observations o_{t+1} close the loop. The lower strip identifies three requirements: hazard-relevant state, valid prediction, and timely, feasible intervention. Solid arrows show runtime flow; dashed arrows supply specifications and assumptions. If no feasible alternative exists, escalation alone does not assure safety. The AI-generated illustrations are conceptual, not experimental trajectories or quantitative uncertainty estimates. Functional roles may share an implementation; barrier- and value-based methods need not generate explicit trajectories.

scene and control generation and includes empirical analysis of prediction pathologies [8]. Embodied-security and navigation-safety surveys organize vulnerabilities, attacks, defenses, and evaluation within their respective scopes [9], [10]. A VLA safety survey maps training-time and inference-time threats and mechanisms [11]; a world-model lifecycle survey follows security failures from data and grounding through imagination and execution [12]. Together, these reviews establish a broad foundation for examining embodied safety.

We focus on the *conditions connecting prediction to an enforceable safety property*. This provides a common basis for comparing geometric filters and generative models while preserving differences in their assurances. Table I summarizes how this focus complements existing reviews; it describes their principal emphases rather than exhaustively cataloguing their contents.

III. REVIEW PROTOCOL AND EVIDENCE

A. Search and Collection

The targeted search was completed on 10 September 2026 and covers predictive safety filters, safe learning with world models, VLA safeguards, and embodied and world-model safety reviews. Web discovery was followed by inspection of selected primary-source PDFs and publication metadata. The repository records queries, retrieval dates, identifiers, and source PDF

hashes. The temporal scope extends from 2015 to the search cutoff, with earlier foundational work included as background.

The dataset contains 292 collected records. Normalized identifiers and titles yield 280 unique records; 8 targeted additions produce 288 candidates (Fig. 2). The collection contains methods, architectures, reviews, benchmarks, and potentially out-of-scope records. The technical synthesis draws on 26 purposively selected full-text sources, including 15 primary method papers.

Bibliographic normalization preserves traceability without deciding scientific eligibility. Collected rows remain linked through identifier and title matches, and unresolved citation keys are treated as metadata issues. The candidate total therefore measures the assembled collection, not the number of eligible primary safety studies.

Coverage is limited by search-engine ranking, source accessibility, and purposive selection. The targeted query log does not establish exhaustive retrieval for the full collection, and no authenticated Scopus or IEEE Xplore search was conducted. We therefore describe the work as a *structured review*. Figure 2 reports normalization and assessment depth, without inferring database-specific retrieval or full-text exclusion counts.

B. Selection and Assessment

The synthesis includes methods that specify, predict, enforce, or evaluate safety-relevant properties of embodied controllers.

TABLE I
POSITIONING RELATIVE TO SELECTED REVIEWS. THE FINAL COLUMN SPECIFIES THE ADDITIONAL FOCUS OF THIS REVIEW.

Review	Principal emphasis	Complementary question addressed here
Safe learning in robotics [3]	Learning-based control, safe RL, and certification	How do those assumptions translate to learned representations and foundation-model policies?
Provably safe RL [7]	Intervention categories and benchmarking	What information and timing does each intervention need at deployment?
World-model safety [8]	Scene/control generation and prediction pathologies	When can predicted consequences support an executable safeguard?
Embodied and navigation security [9], [10]	Vulnerabilities, attacks, defenses, and domain evaluation	What physical property survives an uncertain or compromised predictor?
VLA safety [11]	Threats and mechanisms across training and inference	How do empirical alignment and conditional control guarantees differ?
World-model lifecycle security [12]	Threat propagation through the model lifecycle	Which assumptions must be checked at the prediction–control interface?

General control results are included when their assumptions inform robotic deployment. Architectures provide representation context, benchmarks define evaluation settings, and reviews position the contribution. These roles are coded separately from evidence that a safeguard is effective.

On 11 September 2026, the supplied titles and complete abstracts of the remaining 262 candidates were individually screened with AI assistance. All 288 candidates thus have at least an abstract-level assessment. Each abstract record contains a contribution summary, relevance judgment, full-text follow-up priority, verification question, and text hash. This pass used supplied abstracts rather than freshly verified publisher text.

Abstract judgments remain separate from historical provisional roles and keyword-derived facets, which support only the descriptive audit. Full-text evidence records identify the method, claimed property, assumptions, evaluation setting, and source location. The review does not claim independent human double screening or inter-rater agreement. Borderline decisions and technical interpretations require author validation.

IV. TAXONOMY AND SAFETY PROPERTIES

A. Objectives, Mechanisms, and Representations

A useful taxonomy must distinguish *what is to be protected* from *how protection is attempted*. Collision avoidance, contact-force limits, human safety, and compliance with semantic prohibitions are objectives. CBFs, predictive filtering, reachability, shielding, and constrained learning are mechanisms. Robustness qualifies a property under perturbations; it is not automatically physical safety. A robust policy may reliably execute an unsafe instruction, while a safety filter may reject instructions without producing a robust task policy.

Training paradigms and representations also differ. RL optimizes behavior through interaction, whereas a world model represents dynamics or consequences. A VLA can be fine-tuned with RL and can incorporate learned dynamics. A joint world-action architecture can predict both observations and actions. These descriptions overlap legitimately; treating “safe RL” and “RL” as independent axes builds semantic dependence into the taxonomy before any empirical analysis begins.

We use six facets: safety objective; assurance mechanism; model representation; platform; coordination setting; and

evaluation setting. Platforms include manipulation, navigation, driving, aerial robots, and legged robots. Multi-agent operation is a coordination attribute that can apply to any platform. Simulation and hardware are evaluation settings, not robot types. No facet is assumed statistically independent of another. Unreported information is marked explicitly rather than converted into a negative scientific judgment.

B. Safety Properties

Consider dynamics and observations

$$x_{t+1} = f(x_t, u_t, w_t), \quad o_t = h(x_t, v_t), \quad (1)$$

where disturbances w_t and observation errors v_t may be uncertain. A state estimator produces a state or belief from the available history, and a nominal policy proposes an action. The safety claim must identify the property evaluated on the resulting physical trajectory.

A hard state constraint requires $x_t \in \mathcal{S}$ for every relevant time. A chance constraint may instead require

$$\Pr(\exists t \leq H : x_t \notin \mathcal{S}) \leq \delta. \quad (2)$$

A constrained learning objective commonly bounds expected cumulative cost:

$$\max_{\pi} J_r(\pi) \quad \text{subject to} \quad J_c(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^H c(x_t, u_t) \right] \leq d. \quad (3)$$

Equation (3) allows cost to be distributed unevenly among trajectories. Equation (2) permits a specified failure probability. Neither is equivalent to unconditional invariance. Likewise, a calibrated estimate of safety probability is a statement about predictive reliability under sampling assumptions, not by itself an intervention guarantee.

Semantic constraints may concern sequences rather than instantaneous states: an object must remain away from a heat source, a robot must not enter a private space without permission, or a liquid container must remain upright throughout a transfer. Such properties require explicit temporal scope and a measurable grounding in observations or state. A benchmark’s predicate is an operational definition of the hazard, not necessarily a complete specification of real-world harm.

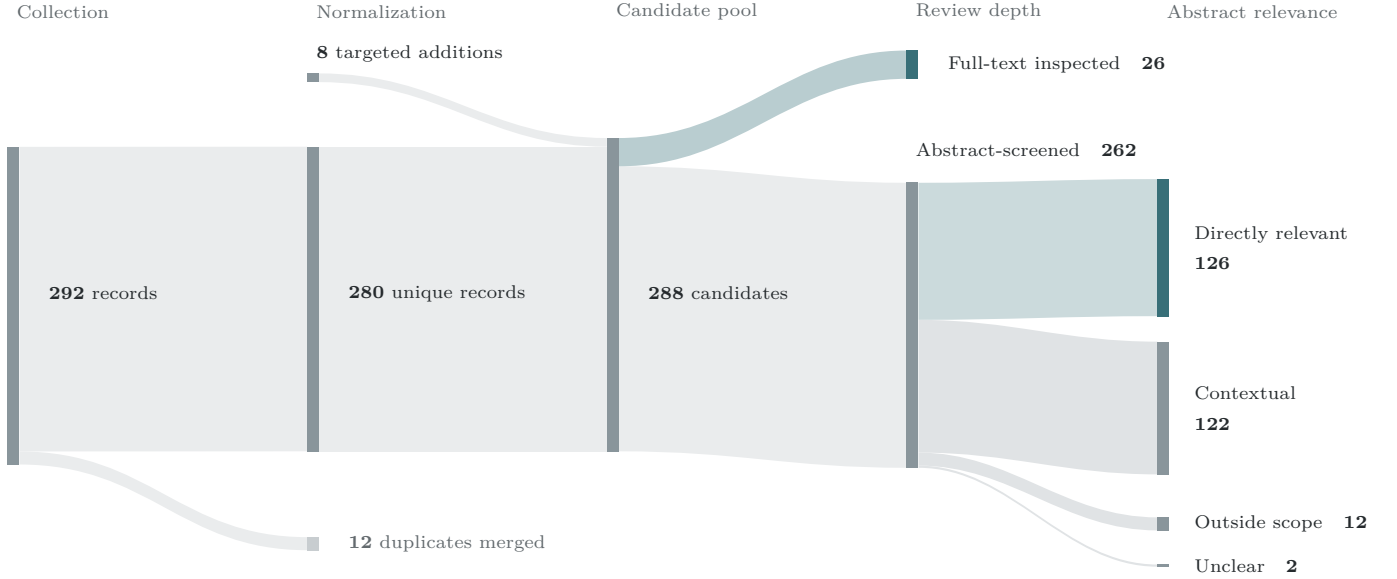


Fig. 2. Collection provenance and assessment depth. Ribbon widths and node heights share a linear record-count scale. Normalization yields 280 unique records from 292 entries; 8 additions bring the collection to 288 candidates. The 26 full-text sources comprise 15 methods, seven reviews, two architectures, and two benchmarks. The remaining 262 records are partitioned by AI-assisted abstract relevance judgments, pending author validation. All candidates are retained. Assessment depth and study role do not establish safety effectiveness; the diagram does not imply exhaustive retrieval or database-specific screening counts.

C. Prediction in the Control Loop

Prediction can support training, runtime monitoring, action selection, or recovery. These placements impose different computational and validation requirements. A model trained to predict future frames may improve representations without generating runtime rollouts. A risk critic may summarize future costs without reconstructing the scene. A model predictive control (MPC) filter may simulate a short horizon and rely on a terminal invariant set to account for what happens afterward.

Figure 3 places selected control and learned-prediction methods on a common timeline using their first public source years. The chronology traces connections across method families without assigning priority for their underlying ideas.

V. PREDICTION AND ENFORCEMENT MECHANISMS

Figure 4 distinguishes learned prediction, finite-horizon optimization, and pointwise admissibility. A world model can supply dynamics to MPC, and a CBF can constrain the optimizer or filter its output. Their composition requires compatible representations, uncertainty assumptions, and feasibility conditions. The following subsections examine these roles and the evidence supporting each mechanism; Table II summarizes the comparison.

A. Barrier Functions and Predictive Control

For continuous-time control-affine dynamics $\dot{x} = f(x) + g(x)u$, a safe set can be represented as $\mathcal{S} = \{x : b(x) \geq 0\}$. Under the regularity and feasibility assumptions of the barrier formulation, constraining

$$\nabla b(x)^\top [f(x) + g(x)u] + \alpha(b(x)) \geq 0 \quad (4)$$

supports forward invariance. A quadratic program can seek the closest admissible input to a nominal command. The foundational automotive examples apply barrier constraints to cruise control and lane keeping [2]. The open deployment issues concern model uncertainty, perception, actuator limits, feasibility, and the completeness of the hazard specification.

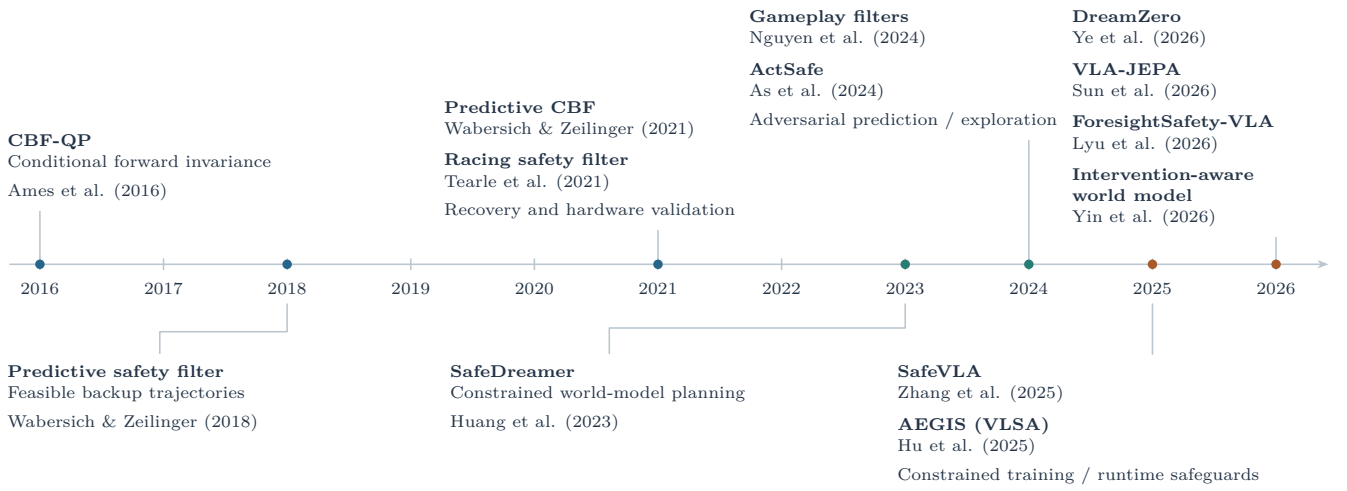
A predictive safety filter searches over a horizon for admissible inputs and a terminal state from which a safe backup controller exists. The nonlinear formulation links probabilistic safety to uncertainty bounds, constraint tightening, and initial feasibility [1]. Its terminal condition supports continued safety beyond the predicted horizon by maintaining an executable backup plan.

Predictive CBFs connect constrained prediction with barrier-based recovery. Their auxiliary soft-constrained construction can restore feasibility under specified conditions [13]. This matters when a disturbance or model error takes the controller outside the original feasible region. Recovery must nevertheless be reported separately from prevention: returning to a safe region does not imply that the intervening trajectory satisfied the original hard constraint. Physical miniature racing experiments demonstrate predictive filtering beyond purely numerical control tasks [14], while also illustrating dependence on a specific vehicle model and track specification.

B. Reachability and Shielding

Reachability methods characterize states from which failure can occur under competing control and disturbance choices. The result can guide action replacement, restrict actions, or define a switching surface. Exact guarantees depend on the modeled dynamics, disturbance set, failure set, and solution accuracy. Approximating a value function with a neural network

Selected milestones in control, learned prediction, and embodied safeguards



Years denote first public versions of the cited sources, not invention of a method family. Axis spacing is proportional to time. Author-year references are clickable; full bibliographic entries appear in the manuscript. Coverage ends 10 September 2026.

Fig. 3. Selected milestones, positioned by first public source year on a linear time axis. The CBF-QP preprint appeared in 2016 (journal publication, 2017), and the nonlinear predictive-filter preprint in 2018. Linked author-year labels identify sources, not dates of invention for entire method families. Sources span control and predictive filters [2], [1], [13], [14], learned prediction [15], [16], [17], and VLA safeguards and predictive architectures [18], [19], [20], [21], [22], [23].

changes the verification obligation: the learned approximation must be related to the underlying safe set before the exact result can be transferred.

Observation-conditioned reachability applies this idea to quadrupedal navigation, using perception and disturbance estimates to adapt a safety value function to new environments. Its simulation and hardware demonstrations show useful transfer across controllers and settings, while its reduced-order model and learned approximation remain material assumptions [24]. The relevant comparison with a language-conditioned safeguard is therefore about state information, intervention authority, and approximation error, not simply network size.

Probabilistic shielding provides a different formal route. An augmented Markov decision process (MDP) can restrict action choices so that a policy satisfies an undiscounted avoidance bound when the safety dynamics are known [25]. The statement protects a specified probability budget. It does not mean every trajectory is failure-free, nor does it establish that a learned perception system has supplied the correct MDP state. The interface between a grounded physical state and the formal abstraction is part of the assurance case.

C. World Models for Constrained Learning

Learned dynamics can improve sample efficiency and make safety evaluation possible from high-dimensional observations. The representation may predict latent states, rewards, costs, or uncertainty. These targets should not be conflated: low reconstruction error may coexist with an inaccurate safety-cost model.

SafeDreamer integrates world-model planning and Lagrangian updates in low-dimensional and visual safe-RL

tasks. It distinguishes online action planning from background planning used to update a policy [15]. The reported benchmark cost reductions support the usefulness of learned prediction, but do not provide a universal per-trajectory safety certificate. A policy deployed without online planning also has a different intervention path from one that repeatedly checks imagined trajectories.

ActSafe combines optimistic exploration with pessimistic constraint evaluation under calibrated models, regularity conditions, and an initial safe policy set. A scalable practical variant addresses more complex visual-control settings [17]. The theoretical algorithm and its neural approximation provide different evidence: transferring the guarantee requires establishing the implementation’s calibration and other assumptions.

Driving-policy work on predictive safety representations provides further empirical evidence that future-violation features can improve a reward–cost trade-off in tested datasets [26]. Such results are distribution- and optimizer-dependent. Dataset-based evaluation is valuable but is not equivalent to deployment in uncontrolled traffic.

D. Adversarial Imagination and Human Intervention

Gameplay filters use imagined interaction between a learned safety policy and an adversary to decide whether to intervene. Their full-order quadruped experiments include physical disturbances and irregular terrain [16]. This differs from sampling likely futures: adversarial imagination attempts to expose unfavorable ones. However, a trained adversary may not span the actual disturbance space, and rollout time must be accounted for before the intervention is applied.

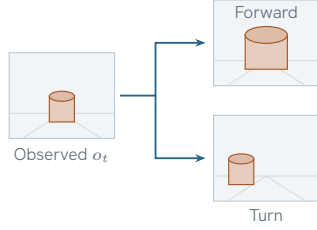
Quadruped navigation
Shared illustrative scenario



Observations and state estimate
Task and safety specification

a World model

Action-conditioned futures



Representation

Latent futures and safety costs;
optional decoded camera views.

Decision

Use predicted costs to plan or learn.
Prediction alone does not enforce safety.

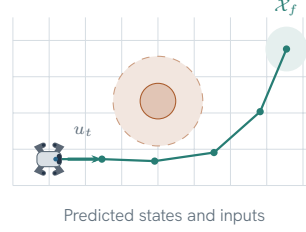
Assurance requires

Hazard-relevant prediction, validated
uncertainty and a safety mechanism.

Huang et al., 2024

b Model predictive control

Constrained state trajectories



Representation

State-input sequence $x_{0:H}, u_{0:H-1}$;
obstacle, input and terminal constraints.

Decision

Optimize the plan; apply its first input.
Re-estimate and replan at each step.

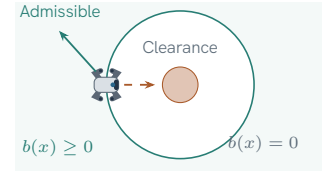
Assurance requires

Valid model/error bounds and recursive
feasibility beyond the planning horizon.

Wabersich & Zeilinger, 2021

c Control barrier function

Safe-set admissibility



Current state and safety margin

Representation

Safe set $\mathcal{S} = \{x : b(x) \geq 0\}$;
input-dependent boundary condition.

Decision

Filter a nominal input to satisfy
 $\dot{b}(x, u) + \alpha(b(x)) \geq 0$.

Assurance requires

Valid barrier and dynamics; feasible timely
control from the certified safe set.

Ames et al., 2017

Closed-loop example

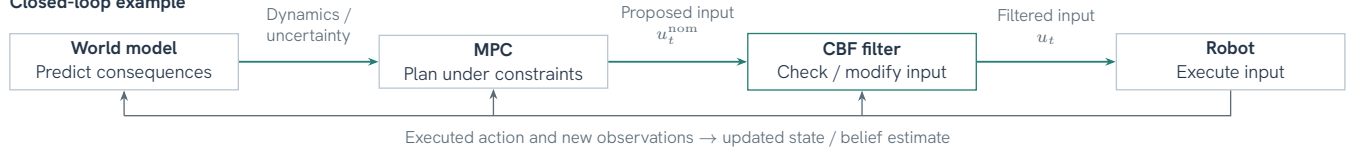


Fig. 4. Three representations of a shared quadruped navigation scenario. (a) A visual world model predicts latent futures and may decode action-conditioned camera views; safety-cost prediction additionally requires an appropriate model head and decision mechanism [15]. (b) MPC optimizes a trajectory under constraints: dots denote predicted robot states, the obstacle envelope a clearance constraint, and $\mathcal{X}_f$ a terminal set [1]. (c) A CBF restricts current inputs through a safe-set derivative condition [2]; predictive CBFs extend this construction [13]. The clearance sketch assumes planar velocity control; full quadruped dynamics require an appropriate barrier formulation. The bottom row shows a possible closed loop. Model validity, barrier feasibility, and the effect of filtering on MPC recursive feasibility must be established jointly. All views are conceptual vector illustrations, not sensor captures, model outputs, or hardware results.

A recent intervention-aware world model uses human takeover signals to predict situations in which an operator would intervene and to shape dexterous control [23]. This connects human feedback to future-action evaluation, but takeover behavior is a proxy for risk, influenced by operator caution and task context. The source reports limited experimental replication and local task variation. A lower intervention rate can indicate improved autonomy, reduced operator vigilance, or both; task success and physical violations must therefore be reported alongside it.

E. VLA Training and Runtime Safeguards

SafeVLA demonstrates constrained Markov decision process (CMDP) learning for VLA policies and introduces Safety-CHORES for long-horizon mobile manipulation [18]. Its evidence concerns specified predicates, elicited hazards, and empirical cumulative costs. The useful research question is how those gains transfer across embodiments, hazard definitions, and policy architectures.

AEGIS supplies a complementary runtime mechanism: a CBF-based constraint layer adjusts VLA actions. Its theorem assumes accurate safety assessment, complete obstacle representation, enclosing geometry, and safe initialization; the stated collision property concerns the modeled end-effector and obstacle [19]. Those conditions must accompany any guarantee claim. Whole-arm collision, force-sensitive damage, missed obstacles, or semantic misuse are not automatically covered by an end-effector separation constraint. Its SafeLIBERO and hardware experiments provide evidence for the evaluated setting rather than for all embodied hazards.

Training-time and runtime mechanisms can be combined. Constrained training may reduce how often a filter needs to intervene, while a filter can provide a narrower explicit constraint at execution. Their composition must still be evaluated: intervention changes the distribution of states visited by the policy, and policy adaptation may change the assumptions used to build the filter.

F. Joint World–Action Prediction

DreamZero jointly predicts future video and actions and learns from heterogeneous robot data. Its zero-shot task and environment generalization builds on training with robot demonstrations [20]. The architecture makes joint prediction available, but a joint generator is not automatically a safety evaluator. It still needs a hazard representation, a decision rule, and an intervention mechanism whose behavior is checked independently of the generated appearance.

VLA-JEPA uses future latent states to supervise representation pretraining, followed by action-head fine-tuning [21]. This objective does not itself establish runtime candidate screening. Its placement of prediction differs from DreamZero’s joint generation, with corresponding differences in latency, uncertainty, and runtime monitoring.

VI. ASSURANCE REQUIREMENTS AND FAILURE MODES

A. Hazard-Relevant Representation

The physical property of interest must survive representation. If a latent state preserves object position but not fragility, mass, temperature, or contact force, a safety evaluator may be unable to distinguish two scenes with different consequences. Improving prediction on irrelevant pixels does not solve this identifiability problem. Object-focused safety prediction demonstrates why state-specific metrics can be more informative than aggregate image metrics [28].

A specification should identify the protected entity, failure event, time interval, observables, and allowable uncertainty. For a manipulator, “avoid collision” may mean end-effector separation, whole-body separation, limiting contact impulse, or preventing contact with a particular object. Each interpretation leads to a different state representation and controller. A benchmark should publish the operational predicate used to mark a violation, including which interactions are allowed during successful completion.

Semantic interpretation introduces another interface. A language model can propose a prohibited action while the low-level controller correctly enforces geometric limits. Conversely, a harmless instruction can produce dangerous motion. Instruction refusal, geometric feasibility, and execution safety must therefore be tested separately, then jointly. A refusal can be the right response to an unsafe goal, but is not a substitute for collision protection during permissible work.

B. Prediction Error at the Safety Boundary

Consider a safety function b that is L_b -Lipschitz and an error bound $\|x_{t+k} - \hat{x}_{t+k}\| \leq \epsilon_k$. A sufficient margin for the predicted state is

$$b(\hat{x}_{t+k}) \geq L_b \epsilon_k. \quad (5)$$

This elementary bound illustrates the required link between predictive accuracy and a physical constraint. It is not a guarantee for an arbitrary learned model: the norm, bound, Lipschitz constant, and covered state domain must be justified. When the bound holds only with a stated probability, so does the resulting implication. A model that omits the relevant physical

variable cannot satisfy the premise merely by predicting observations well.

Prediction errors can compound over the horizon. Increasing the number of generated futures does not compensate for shared model bias, and uncertainty estimates must distinguish stochastic variability from lack of knowledge. Evaluation should inspect false-safe predictions near the boundary and after interventions, where the visited state distribution may differ most from the training data.

Calibrated safety-chance prediction provides one route to quantifying reliability. The inspected formulation uses independent trajectories in its sampling argument, and its calibration result concerns confidence-score error at the dataset level [27]. It must not be reinterpreted as a guarantee for temporally correlated frames, individual trajectories, or arbitrary deployment shifts. Adaptation results and a calibration theorem answer different questions and should be reported separately.

C. Intervention Timing and Authority

Prediction has operational value only while an effective intervention remains possible. Let T_{sense} , T_{infer} , T_{assess} , T_{solve} , and T_{actuate} denote sensing, inference, assessment, optimization, and actuation delays. A necessary timing condition is

$$T_{\text{sense}} + T_{\text{infer}} + T_{\text{assess}} + T_{\text{solve}} + T_{\text{actuate}} < T_{\text{recoverable}}, \quad (6)$$

where $T_{\text{recoverable}}$ is the remaining time before safe recovery becomes impossible under the specified dynamics. The quantity is scenario-dependent. A nominal inference frequency alone does not establish the condition: queued commands, sensor age, worst-case solver time, and actuator response also matter.

Recovery authority is equally important. A high-level language planner may update slowly while a faster low-level controller executes a buffered action sequence. If the monitor can only reject the next high-level plan, its authority may be insufficient to stop an ongoing hazardous motion. The system description should identify which component can interrupt execution, what happens to buffered commands, and how the fallback remains feasible.

D. Composition and Operating Domain

An assurance claim should be written as an implication from stated assumptions to a stated property. It should identify perception validity, model uncertainty, admissible disturbances, control feasibility, and timing conditions. If a runtime monitor checks an assumption, the response to a failed check must also be specified. An alarm whose fallback depends on the same failed assumption does not close the argument.

Shared models create correlated failure modes. When one network proposes an action, imagines its consequence, and judges the imagined result, a representational omission can pass through all stages. Independent geometric checks, separately calibrated risk estimators, and monitors of model residuals can provide complementary evidence, but independence should not be assumed merely because components have different names. World-model lifecycle security also matters here: compromised

TABLE II
REPRESENTATIVE METHODS AND THEIR ASSURANCE EVIDENCE

Method	Mechanism	Assurance evidence	Hardware
<i>Conditional formal or calibration results</i>			
CBF-QP [2]	CBF constraint	Invariance*	–
Predictive safety filter [1]	Horizon + backup	Probability bound*	–
Predictive CBF [13]	Soft-constrained prediction	Recovery conditions*	–
Probabilistic shielding [25]	Restricted action set	Probability bound*	–
ActSafe [17]	Constrained learning	Theory + approximation*	–
AEGIS / VLISA [19]	Runtime CBF constraint	End-effector separation*	Yes
Safety-chance calibration [27]	Risk prediction	Confidence calibration*	–
<i>Empirical safety or intervention evidence</i>			
Racing safety filter [14]	Horizon + backup	Hardware safety filtering	Yes
SafeDreamer [15]	Constrained planning	Empirical cost reduction	–
SafeVLA [18]	Constrained VLA learning	Empirical cost reduction	–
Gameplay filter [16]	Adversarial rollout	Empirical safety	Yes
Observation-conditioned reachability [24]	Safety value function	Empirical safety	Yes
Foundation safety predictor [28]	Object-focused prediction	Empirical prediction	–
Predictive safety representations [26]	Future-violation features	Empirical cost reduction	–
Intervention-aware WM [23]	Takeover prediction	Empirical behavior	Yes
<i>Predictive architectures</i>			
DreamZero [20]	Joint world–action model	Safety not established	Yes
VLA-JEPA [21]	Predictive pretraining	Safety not established	Yes

*Conditional on the source’s stated assumptions and safety property; see the technical discussion and evidence dataset for details. “Yes” denotes reported hardware evaluation; “–” means hardware evidence is not established in this comparison. Hardware evaluation alone does not establish a safety guarantee.

inputs or learned dynamics can produce convincing but unsafe predictions [12].

An operational design domain (ODD) should therefore be explicit. It includes the environment, embodiment, payload, disturbances, sensing conditions, and human interactions for which evidence is available. Outside that domain, the appropriate behavior may be degraded operation, requesting assistance, or stopping where a safe stop is feasible. No single fallback is universally safe; the recovery policy must be evaluated under the same physical conditions as the nominal behavior.

VII. EVALUATION AND REPORTING

A. Benchmark Scope

Safety-Gymnasium provides safe-RL tasks and algorithm infrastructure, including multiple observation and agent settings [29]. Safety-CHORES evaluates specified hazards in long-horizon mobile manipulation [18]. SafeLIBERO focuses on safety-critical manipulation scenarios for VLA evaluation [19]. SafeAgentBench supplies executable hazardous tasks for embodied language-agent planning [30]. ForesightSafety-VLA separates physical, language, and vision-related variations and reports process-level cost and risk exposure [22]. Their differing hazard definitions and measurement procedures motivate a closer comparison.

The remaining challenge is comparability and coverage. A language-planning benchmark can evaluate unsafe goal completion without measuring torque or contact damage. A collision benchmark can enforce distance while overlooking semantic misuse. A simulator can provide exact labels for

its internal dynamics while omitting the failure modes of a physical sensor or actuator. Table III distinguishes these roles without aggregating incomparable scores.

B. Minimum Evaluation Record

Every method comparison should define the hazard and record the number of episodes, independent seeds, hardware trials, and exposure duration. Report task success jointly with at least one harm-specific violation measure. Distinguish frequency from severity: a brief low-force contact and a severe collision need not contribute equal cost. When combining hazards into a scalar, disclose the weights and retain disaggregated results.

Report the joint outcomes of safe success, unsafe success, safe failure, and unsafe failure. A policy that stops immediately can have a low violation rate without useful competence; a successful policy can still accumulate unacceptable risk. Intervention frequency, duration, recovery success, and residual violations describe what the safeguard actually contributes. Human takeover rate should include operator protocol and opportunities to intervene.

Prediction evaluation should include false-safe and false-unsafe rates, calibration, horizon dependence, and performance near decision boundaries. Split data by independent episodes or deployment contexts rather than adjacent video frames. State whether probabilities refer to a single step, an entire episode, or an aggregate sample. Report uncertainty intervals with the resampling unit and avoid interpreting a no-failure sample as proof of zero failure probability.

TABLE III
REPRESENTATIVE EVALUATION RESOURCES AND THE LIMITS OF CROSS-BENCHMARK COMPARISON. THESE ARE PUBLISHED RESOURCES, NOT NEW EXPERIMENTS IN THIS REVIEW.

Resource	Safety evidence emphasized	Additional deployment evidence needed
Safety-Gymnasium [29]	Task reward and safety cost in controlled RL environments	Whether task costs cover the deployed robot’s physical hazards
Safety-CHORES [18]	Predicate-based costs in long-horizon mobile manipulation	Transfer across embodiment, perception, and hazard definitions
SafeLIBERO [19]	Manipulation under obstacles and safety interventions	Whole-body, contact-force, and incomplete-perception behavior
SafeAgentBench [30]	Executable and semantic evaluation of hazardous plans	Low-level motion and physical consequence validation
ForesightSafety-VLA [22]	Structured scene/language/vision variation; cost and exposure	External validity beyond the specified simulated scenarios

C. Closed-Loop Stress Testing

A useful stress test varies both the task and the assumptions behind the safeguard. Perturb observation quality, dynamics, payload, contact properties, distractors, instruction phrasing, and other-agent behavior. Include stale sensor data, delayed prediction, infeasible optimization, missed obstacles, and interruptions while commands are buffered. Separate failures of perception, prediction, safety assessment, and intervention wherever instrumentation permits.

Compare a nominal policy, a safeguard alone, safety-oriented training alone, and their combination on matched conditions. Keep episode budgets, observation access, action authority, and hardware constraints comparable. A method with privileged simulator state should not be presented as directly equivalent to a vision-only method. Ablations should identify whether improvements come from better task competence, a better hazard representation, or a more effective intervention.

Physical validation adds evidence about model mismatch and execution timing, but a small hardware demonstration remains limited evidence. Report the operational envelope and the changes between simulation and hardware. For safety-critical failures, describe containment and recovery rather than only successful demonstrations. When experiments cannot safely reproduce a hazard, distinguish simulated evidence, analytical justification, and physically measured components of the argument.

VIII. DESCRIPTIVE COVERAGE AND LIMITATIONS

Figure 5 summarizes all 288 candidates by study role and bibliographic year. Roles for full-text sources come from technical evidence records; the remaining roles retain historical provisional classifications. The subsequent abstract assessments are recorded separately and do not recode these totals. All records have a year, and each panel sums to the full collection. These counts describe the assembled literature rather than field-wide prevalence or mechanism effectiveness.

Supplementary association outputs use binary paper-level labels. If A_i and B_j identify two labels, observed overlap is $C_{ij} = \sum_n A_{ni} B_{nj}$ and independence gives $E_{ij} = n_i n_j / N$, with N the number of candidate records. The descriptive deficit $(E_{ij} - C_{ij}) / (E_{ij} + 1)$ is a regularized association residual. It

has no normative interpretation: independence is not an ideal allocation of research effort.

Whole-paper bootstrap resampling preserves each record’s label combinations. Permutation shifts an entire label block between papers; an additional analysis permutes within publication-era strata. This preserves within-block label relationships in the permutation reference. The supplementary outputs report intervals and rank frequencies, and explicitly label the inputs as provisional. Resampling cannot repair selection bias, missing studies, or inaccurate coding. Association tests are therefore not used to establish the review’s central conclusions.

Risk weighting is an illustrative sensitivity exercise comparing equal weights with specified platform weights and their perturbations. The weights were not obtained from a documented expert elicitation and do not measure real deployment risk. Likewise, exploratory topic modeling is separated from technical classification. Its selection rule uses mean UMass coherence across five seeds, with smaller topic counts preferred in a tie; the metric and its in-sample character are documented. Topic prevalence is not evidence of a safety guarantee or an important research gap.

The review emphasizes interfaces between control and foundation models. It does not exhaustively cover surgical robotics, wearable assistance, certification practice, or every application domain. Many recent results are preprints, making source version and archival status relevant to interpretation. Purposive full-text selection favors accessible, illustrative methods. The evidence records expose that selection but cannot replace broader domain expertise or independent author validation.

IX. RESEARCH PRIORITIES

A. Safety-Relevant Dynamics

Learned representations should preserve variables that determine harm and expose uncertainty when those variables cannot be inferred. Evaluation should measure hazard-prediction and safety-function error alongside image quality and task success. Counterfactual action tests can assess whether a model responds to interventions rather than merely generating plausible continuations. These are evaluation priorities, not assumed capabilities of current world models.

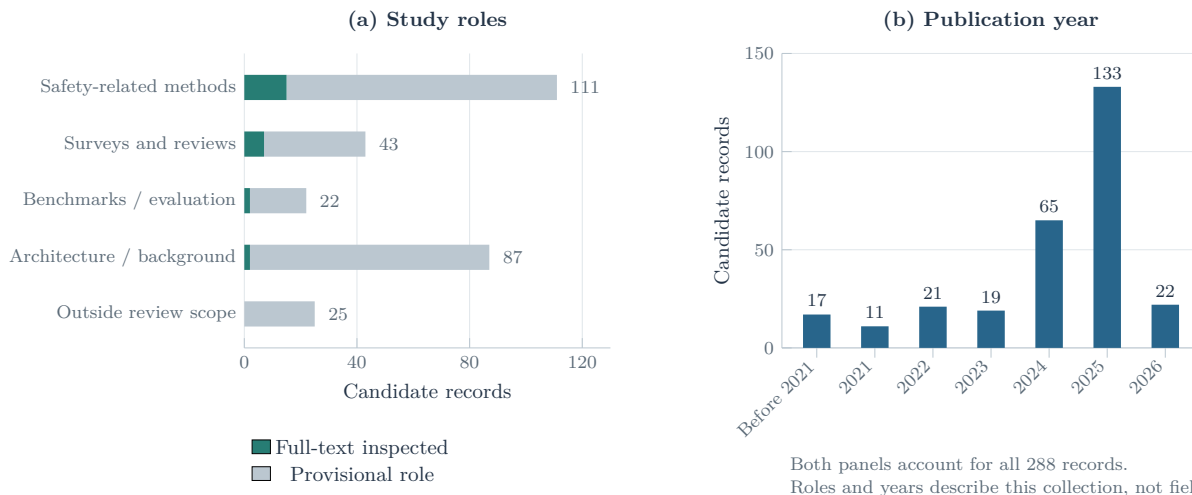


Fig. 5. Coverage of the complete 288-record collection. (a) Mutually exclusive study roles, with full-text-inspected sources distinguished from historical provisional classifications. (b) Recorded publication years, grouping foundational work before 2021. Each panel accounts for every candidate. Bibliographic years can refer to preprints or archival publications; they do not uniformly indicate first public appearance.

B. Recovery Under Distribution Shift

Dependable intervention requires tracking when model error, new obstacles, actuator saturation, or delay invalidate a recovery strategy. Combining a rich nominal policy with a narrower monitored safeguard requires explicit tests of representation and timing compatibility. Evaluation should measure sustained safety and useful task progress within a documented domain, alongside intervention frequency.

C. Predictive Confidence

Predictive confidence should be evaluated at safety-critical decisions. An average calibrated score can conceal errors on rare, consequential states. Studies should state sampling assumptions, test horizon-dependent calibration, and assess post-intervention and shifted distributions separately. Human feedback can help identify hazards, but operator-specific takeover behavior remains a proxy for harm.

D. Comparable Assurance Records

A common assurance record should identify the property, assumptions, intervention authority, runtime budget, independent trials, and operating domain. Benchmarks should publish task outcomes and the hazard predicates used to label failures. These records would support comparison across control, world-model learning, and VLAs while preserving differences in evidence strength. Table IV links the proposed work to the limitations identified in this review.

X. CONCLUSION

Predictive safety depends on the complete perception–prediction–intervention loop. Control-theoretic filters connect explicit assumptions to conditional guarantees; learned models expand the observations, tasks, and consequences that can be represented. Combining them requires a grounded safety

property, reliable uncertainty estimates, timely execution, and feasible recovery.

The central research questions are which hazard-relevant variables a system preserves, where its predictions remain reliable, and whether its safeguard can act effectively. Progress requires evaluating those interfaces in deployment conditions and reporting empirical performance separately from conditional guarantees.

DATA AND REPRODUCIBILITY

The repository provides the raw-to-canonical audit, query log, source manifest, full-text evidence, abstract assessments, taxonomy, analysis code, and generated figures. Quantitative descriptions are regenerated from the canonical candidate collection. Source PDFs are referenced by URL and hash rather than redistributed. These artifacts support replication of the analysis and inspection of its coverage and coding decisions.

The full bibliography includes the candidate collection for traceability. Bibliographic inclusion does not imply eligibility as primary safety evidence or full-text inspection.

REFERENCES

- [1] K. P. Wabersich and M. N. Zeilinger, “A predictive safety filter for learning-based control of constrained nonlinear dynamical systems,” arXiv:1812.05506, 2018. [Online]. Available: <https://arxiv.org/abs/1812.05506>
- [2] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, “Control Barrier Function Based Quadratic Programs for Safety Critical Systems,” arXiv:1609.06408, 2016. [Online]. Available: <https://arxiv.org/abs/1609.06408>
- [3] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, “Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning,” arXiv:2108.06266, 2021. [Online]. Available: <https://arxiv.org/abs/2108.06266>
- [4] J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” *Journal of Machine Learning Research*, vol. 16, no. 42, pp. 1437–1480, 2015.
- [5] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, “A review of safe reinforcement learning: Methods, theories and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 11 216–11 235, 2024.

TABLE IV
RESEARCH DIRECTIONS DERIVED FROM THE ASSURANCE-CHAIN ANALYSIS

Assurance obligation	Failure mode	Evaluation requirement
Safety-relevant representation [28], [8]	Hazard variables are absent from the representation.	Measure hazard error alongside image quality and task success.
Executable intervention [1], [19], [17]	Model error or delay invalidates the safe-guard.	Test missed hazards, infeasible plans, solver failures, and delayed actuation.
Reliable prediction under shift [27], [23]	Calibration fails outside the operating domain.	Test shifted data and report operating limits and fallback behavior.
Comparable safety evidence [22]	Success and violations measure different outcomes.	Report harm, severity, exposure, intervention, and success across independent trials.

These directions synthesize limitations in the inspected sources. They propose evaluation work and do not rank societal risk or infer priorities from sparse keyword counts.

- [6] A. Wachi, X. Shen, and Y. Sui, "A survey of constraint formulations in safe reinforcement learning," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 8262–8271.
- [7] H. Krasowski, J. Thumm, M. Müller, L. Schäfer, X. Wang, and M. Althoff, "Provably Safe Reinforcement Learning: Conceptual Analysis, Survey, and Benchmarking," arXiv:2205.06750, 2022. [Online]. Available: <https://arxiv.org/abs/2205.06750>
- [8] L. Baraldi, Z. Zeng, C. Zhang, A. Nayak, H. Zhu, F. Liu *et al.*, "The Safety Challenge of World Models for Embodied AI Agents: A Review," arXiv:2510.05865, 2025. [Online]. Available: <https://arxiv.org/abs/2510.05865>
- [9] W. Xing, M. Li, M. Li, and M. Han, "Towards Robust and Secure Embodied AI: A Survey on Vulnerabilities and Attacks," arXiv:2502.13175, 2025. [Online]. Available: <https://arxiv.org/abs/2502.13175>
- [10] Z. Wang, J. Hu, and R. Mu, "Safety of Embodied Navigation: A Survey," arXiv:2508.05855, 2025. [Online]. Available: <https://arxiv.org/abs/2508.05855>
- [11] Q. Li, B. Yin, W. Huang, R. Liu, B. Zou, R. Yu, J. Ye, W. Yu, and X. Wang, "Vision-Language-Action Safety: Threats, Challenges, Evaluations, and Mechanisms," arXiv:2604.23775, 2026. [Online]. Available: <https://arxiv.org/abs/2604.23775>
- [12] F. Liu, Z. Chen, H. Tan, Y. Meng, G. Chen, and H. Zhu, "Security of World-Model-Based Embodied AI: A Lifecycle of Threats, Defenses, and Evaluation," arXiv:2607.28226, 2026. [Online]. Available: <https://arxiv.org/abs/2607.28226>
- [13] K. P. Wabersich and M. N. Zeilinger, "Predictive control barrier functions: Enhanced safety mechanisms for learning-based control," arXiv:2105.10241, 2021. [Online]. Available: <https://arxiv.org/abs/2105.10241>
- [14] B. Tearle, K. P. Wabersich, A. Carron, and M. N. Zeilinger, "A predictive safety filter for learning-based racing control," arXiv:2102.11907, 2021. [Online]. Available: <https://arxiv.org/abs/2102.11907>
- [15] W. Huang, J. Ji, C. Xia, B. Zhang, and Y. Yang, "SafeDreamer: Safe Reinforcement Learning with World Models," arXiv:2307.07176, 2023. [Online]. Available: <https://arxiv.org/abs/2307.07176>
- [16] D. P. Nguyen, K.-C. Hsu, W. Yu, J. Tan, and J. F. Fisac, "Gameplay Filters: Robust Zero-Shot Safety through Adversarial Imagination," arXiv:2405.00846, 2024. [Online]. Available: <https://arxiv.org/abs/2405.00846>
- [17] Y. As, B. Sukhija, L. Treven, C. Sferrazza, S. Coros, and A. Krause, "ActSafe: Active Exploration with Safety Constraints for Reinforcement Learning," arXiv:2410.09486, 2024. [Online]. Available: <https://arxiv.org/abs/2410.09486>
- [18] B. Zhang, Y. Zhang, J. Ji, Y. Lei, Y. Cai, J. Dai, Y. Chen, and Y. Yang, "SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Constrained Learning," arXiv:2503.03480, 2025. [Online]. Available: <https://arxiv.org/abs/2503.03480>
- [19] S. Hu, Z. Liu, S. Liu, J. Cen, Z. Meng, S. Wang, X. Li, and X. He, "VLSA: Vision-Language-Action Models with Plug-and-Play Safety Constraint Layer," arXiv:2512.11891, 2025. [Online]. Available: <https://arxiv.org/abs/2512.11891>
- [20] S. Ye, Y. Ge, K. Zheng, S. Gao, S. Yu, G. Kurian *et al.*, "World Action Models are Zero-shot Policies," arXiv:2602.15922, 2026. [Online]. Available: <https://arxiv.org/abs/2602.15922>
- [21] J. Sun, W. Zhang, Z. Qi, S. Ren, Z. Liu, H. Zhu, G. Sun, X. Jin, and Z. Chen, "VLA-JEPA: Enhancing Vision-Language-Action Model with Latent World Model," arXiv:2602.10098, 2026. [Online]. Available: <https://arxiv.org/abs/2602.10098>
- [22] M. Lyu, Y. Sun, Y. Jia, S. Shen, M. Sha, H. Li, F. Zhao, and Y. Zeng, "ForesightSafety-VLA: A Unified Diagnostic Safety Benchmark for Vision-Language-Action Models," arXiv:2606.27079, 2026. [Online]. Available: <https://arxiv.org/abs/2606.27079>
- [23] J. Yin, Z. Zhang, L. Xu, H. Zhang, J. Shao, Y. Feng, A. Ajoudani, and R. Xu, "How to Learn from What a Human Would Avoid? Intervention-Aware World Models with Real-World RL for Dexterous Manipulation," arXiv:2609.06009, 2026. [Online]. Available: <https://arxiv.org/abs/2609.06009>
- [24] A. Lin, S. Peng, and S. Bansal, "One Filter to Deploy Them All: Robust Safety for Quadrupedal Navigation in Unknown Environments," arXiv:2412.09989, 2024. [Online]. Available: <https://arxiv.org/abs/2412.09989>
- [25] E. H.-D. I. Court, F. Belardinelli, and A. W. Goodall, "Probabilistic Shielding for Safe Reinforcement Learning," arXiv:2503.07671, 2025. [Online]. Available: <https://arxiv.org/abs/2503.07671>
- [26] M. Keswani and R. Bhattacharyya, "Learning Safe Autonomous Driving Policies Using Predictive Safety Representations," arXiv:2512.17586, 2025. [Online]. Available: <https://arxiv.org/abs/2512.17586>
- [27] Z. Mao, M. E. Umasudhan, and I. Ruchkin, "How Safe Will I Be Given What I Saw? Calibrated Prediction of Safety Chances for Image-Controlled Autonomy," arXiv:2508.09346, 2025. [Online]. Available: <https://arxiv.org/abs/2508.09346>
- [28] Z. Mao, S. Dai, Y. Geng, and I. Ruchkin, "Zero-shot Safety Prediction for Autonomous Robots with Foundation World Models," arXiv:2404.00462, 2024. [Online]. Available: <https://arxiv.org/abs/2404.00462>
- [29] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, and Y. Yang, "Safety Gymnasium: A unified safe reinforcement learning benchmark," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 964–18 993, 2023.
- [30] S. Yin, X. Pang, Y. Ding, M. Chen, Y. Bi, Y. Xiong, W. Huang, Z. Xiang, J. Shao, and S. Chen, "SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents," arXiv:2412.13178, 2024. [Online]. Available: <https://arxiv.org/abs/2412.13178>
- [31] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3D-VLA: A 3D Vision-Language-Action Generative World Model," arXiv:2403.09631, 2024. [Online]. Available: <https://arxiv.org/abs/2403.09631>
- [32] G. Paolo, J. Gonzalez-Billardon, and B. Kégl, "A call for embodied AI," arXiv:2402.03824, 2024. [Online]. Available: <https://arxiv.org/abs/2402.03824>
- [33] F. Sun, R. Chen, T. Ji, Y. Luo, H. Zhou, and H. Liu, "A Comprehensive Survey on Embodied Intelligence: Advancements, Challenges, and Future Perspectives," *CAAI Artificial Intelligence Research*, 2024. [Online]. Available: <https://www.sciopen.com/article/10.26599/AIR.2024.9150042>
- [34] Y. Huang, L. Ding, Z. Tang, T. Wang, X. Lin, W. Zhang, M. Ma, and

- Y. Zhang, "A Framework for Benchmarking and Aligning Task-Planning Safety in LLM-Based Embodied Agents," arXiv:2504.14650, 2025. [Online]. Available: <https://arxiv.org/abs/2504.14650>
- [35] B. Wen, "A Framework for Inherently Safer AGI through Language-Mediated Active Inference," arXiv:2508.05766, 2025. [Online]. Available: <https://arxiv.org/abs/2508.05766>
- [36] C. Dandjinou, "A hierarchical architecture for embodied AI: planner-skills separation, distributed safety contracts, and real-time enforcement," Authorea Preprints, 2025. [Online]. Available: https://d197for5662m48.cloudfront.net/documents/publicationstatus/291247/preprint_pdf/db0e4a2fdbefb68c72e69a6b1d50bfce.pdf
- [37] S. Gu, A. Kshirsagar, Y. Du, G. Chen, J. Peters, and A. Knoll, "A Human-Centered Safe Robot Reinforcement Learning Framework with Interactive Behaviors," arXiv:2302.13137, 2023. [Online]. Available: <https://arxiv.org/abs/2302.13137>
- [38] C. Ellis, M. Wigness, and L. Fiondella, "A Mapping of Assurance Techniques for Learning Enabled Autonomous Systems to the Systems Engineering Lifecycle," 2022 IEEE International Conference on Assured Autonomy (ICAA), 2022. [Online]. Available: <https://doi.org/10.1109/ICAA52185.2022.00013>
- [39] N. Bührer, Z. Zhang, A. Liniger, F. Yu, and L. V. Gool, "A Multiplicative Value Function for Safe and Efficient Reinforcement Learning," arXiv:2303.04118, 2023. [Online]. Available: <https://arxiv.org/abs/2303.04118>
- [40] Y. Zhang, J. Tian, and Q. Xiong, "A review of embodied intelligence systems: a three-layer framework integrating multimodal perception, world modeling, and structured strategies," Frontiers in Robotics and AI, 2025. [Online]. Available: <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2025.1668910/full>
- [41] Y. Zheng, L. Yao, Y. Su, Y. Zhang, Y. Wang, S. Zhao, Y. Zhang, and L.-P. Chau, "A Survey of Embodied Learning for Object-centric Robotic Manipulation," Machine Intelligence Research, 2025. [Online]. Available: <http://dx.doi.org/10.1007/s11633-025-1542-8>
- [42] A. Kushwaha, K. Ravish, P. Lamba, and P. Kumar, "A Survey of Safe Reinforcement Learning and Constrained MDPs: A Technical Survey on Single-Agent and Multi-Agent Safety," arXiv:2505.17342, 2025. [Online]. Available: <https://arxiv.org/abs/2505.17342>
- [43] Y. Xie, "A Survey of Safe Reinforcement Learning Methods in Robotics," ITM Web of Conferences, 2025. [Online]. Available: https://www.itm-conferences.org/articles/itmconf/pdf/2025/09/itmconf_cseit2025_01014.pdf
- [44] T. Feng, W. Wang, and Y. Yang, "A Survey of World Models for Autonomous Driving," arXiv:2501.11260, 2025. [Online]. Available: <https://arxiv.org/abs/2501.11260>
- [45] Z. Xu, K. Wu, J. Wen, J. Li, N. Liu, Z. Che, and J. Tang, "A Survey on Robotics with Foundation Models: toward Embodied AI," arXiv:2402.02385, 2024. [Online]. Available: <https://arxiv.org/abs/2402.02385>
- [46] S. Jiang, Z. Huang, K. Qian, Z. Luo, T. Zhu, Y. Zhong *et al.*, "A Survey on Vision-Language-Action Models for Autonomous Driving," arXiv preprint arXiv:2506.24044, 2025. [Online]. Available: <https://arxiv.org/abs/2506.24044>
- [47] Y. Zhong, F. Bai, S. Cai, X. Huang, Z. Chen, X. Zhang *et al.*, "A Survey on Vision-Language-Action Models: An Action Tokenization Perspective," arXiv:2507.01925, 2025. [Online]. Available: <https://arxiv.org/abs/2507.01925>
- [48] X. Long, Q. Zhao, K. Zhang, Z. Zhang, D. Wang, Y. Liu *et al.*, "A Survey: Learning Embodied Intelligence from Physical Simulators and World Models," arXiv:2507.00917, 2025. [Online]. Available: <https://arxiv.org/abs/2507.00917>
- [49] W. Salhab, D. Ameyed, F. Jaafar, and H. Mccheick, "A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions," IEEE Access, 2024. [Online]. Available: <https://doi.org/10.1109/ACCESS.2024.3440647>
- [50] S. Huang, H. Chang, Y. Liu, Y. Zhu, H. Dong, P. Gao, A. Boularias, and H. Li, "A3VLM: Actionable Articulation-Aware Vision Language Model," arXiv:2406.07549, 2024. [Online]. Available: <https://arxiv.org/abs/2406.07549>
- [51] Z. Ying, L. Wang, Y. Xiao, J. Wang, Y. Ma, J. Guo, Z. Yin, M. Zhang, A. Liu, and X. Liu, "AGENTS SAFE: Benchmarking the Safety of Embodied Agents on Hazardous Instructions," arXiv:2506.14697, 2025. [Online]. Available: <https://arxiv.org/abs/2506.14697>
- [52] X. Qi, Y. Huang, Y. Zeng, E. Debenedetti, J. Geiping, L. He *et al.*, "AI Risk Management Should Incorporate Both Safety and Security," arXiv:2405.19524, 2024. [Online]. Available: <https://arxiv.org/abs/2405.19524>
- [53] B. Gyevar and A. Kasirzadeh, "AI safety for everyone," Nature Machine Intelligence, 2025. [Online]. Available: <https://arxiv.org/abs/2502.09288>
- [54] Z. Chen, G. Li, and X. Wan, "Align, Reason and Learn: Enhancing Medical Vision-and-Language Pre-training with Knowledge," arXiv:2209.07118, 2022. [Online]. Available: <https://arxiv.org/abs/2209.07118>
- [55] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, "Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI," arXiv:2407.06886, 2025. [Online]. Available: <https://arxiv.org/abs/2407.06886>
- [56] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang, "An Embodied Generalist Agent in 3D World," arXiv:2311.12871, 2024. [Online]. Available: <https://arxiv.org/abs/2311.12871>
- [57] Z. Zhang, K. Qian, B. W. Schuller, and D. Wollherr, "An Online Robot Collision Detection and Identification Scheme by Supervised Learning and Bayesian Decision Theory," IEEE Transactions on Automation Science and Engineering, 2021. [Online]. Available: <https://doi.org/10.1109/TASE.2020.2997094>
- [58] W. D'Alessandro and C. D. Kirk-Giannini, "Artificial Intelligence: Approaches to Safety," Philosophy Compass, 2025. [Online]. Available: <https://compass.onlinelibrary.wiley.com/doi/10.1111/phc3.70039>
- [59] L. Hinkeldey, E. Fosong, E. Miller, R. Rubavicius, T. McInroe, P. Wollstadt, C. B. Wiebel-Herboth, S. Ramamoorthy, and S. V. Albrecht, "Assistax: A Hardware-Accelerated Reinforcement Learning Benchmark for Assistive Robotics," arXiv:2507.21638, 2025. [Online]. Available: <https://arxiv.org/abs/2507.21638>
- [60] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li *et al.*, "BadRobot: Jailbreaking Embodied LLMs in the Physical World," arXiv:2407.20242, 2025. [Online]. Available: <https://arxiv.org/abs/2407.20242>
- [61] F. Berkenkamp, A. Krause, and A. P. Schoellig, "Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics," Machine learning, 2023. [Online]. Available: <https://arxiv.org/abs/1602.04450>
- [62] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges," arXiv:2501.02189, 2025. [Online]. Available: <https://arxiv.org/abs/2501.02189>
- [63] A. R. Mahmood, D. Korenkevych, G. Vasan, W. Ma, and J. Bergstra, "Benchmarking Reinforcement Learning Algorithms on Real-World Robots," arXiv:1809.07731, 2018. [Online]. Available: <https://arxiv.org/abs/1809.07731>
- [64] P. Guruprasad, H. Sikka, J. Song, Y. Wang, and P. P. Liang, "Benchmarking Vision, Language, & Action Models on Robotic Learning Tasks," arXiv:2411.05821, 2024. [Online]. Available: <https://arxiv.org/abs/2411.05821>
- [65] K. Mitrokhov, "Between world models and model worlds: on generality, agency, and worlding in machine learning," AI & SOCIETY, 2025. [Online]. Available: <https://link.springer.com/article/10.1007/s00146-024-02086-9>
- [66] E. Idowu, "Bridging the Embodiment Gap: Embodied AI for Enhanced Human-Machine Collaboration and Learning in Dynamic Environments," 2024. [Online]. Available: <https://www.preprints.org/manuscript/202407.0869>
- [67] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025. [Online]. Available: <https://arxiv.org/abs/2502.14420>
- [68] J. Li, Y. Zhu, Z. Tang, J. Wen, M. Zhu, X. Liu *et al.*, "CoA-VLA: Improving Vision-Language-Action Models via Visual-Text Chain-of-Affordance," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025. [Online]. Available: <https://arxiv.org/abs/2412.20451>
- [69] Q. Li *et al.*, "CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation," arXiv preprint arXiv:2411.19650, 2024. [Online]. Available: <https://arxiv.org/abs/2411.19650>
- [70] B. A. Hill, M. K. E. Wei, and T. Jishnuanandh, "Communicating Plans, Not Percepts: Scalable Multi-Agent Coordination with Embodied World Models," arXiv:2508.02912, 2025. [Online]. Available: <https://arxiv.org/abs/2508.02912>
- [71] Z. Liu, Z. Cen, V. Isenbaev, W. Liu, Z. S. Wu, B. Li, and D. Zhao, "Constrained Variational Policy Optimization for Safe Reinforcement Learning," arXiv:2201.11927, 2022. [Online]. Available: <https://arxiv.org/abs/2201.11927>

- [72] M. Neunert, A. Abdolmaleki, M. Wulfmeier, T. Lampe, J. T. Springenberg, R. Hafner, F. Romano, J. Buchli, N. Heess, and M. Riedmiller, "Continuous-Discrete Reinforcement Learning for Hybrid Control in Robotics," arXiv:2001.00449, 2020. [Online]. Available: <https://arxiv.org/abs/2001.00449>
- [73] M. Argus, J. Bratulic, H. Masnavi, M. Velikanov, N. Heppert, A. Valada, and T. Brox, "cVLA: Towards Efficient Camera-Space VLAs," arXiv:2507.02190, 2025. [Online]. Available: <https://arxiv.org/abs/2507.02190>
- [74] H. Hu, Z. Zhang, K. Nakamura, A. Bajcsy, and J. F. Fisac, "Deception Game: Closing the Safety-Learning Loop in Interactive Robot Autonomy," arXiv:2309.01267, 2023. [Online]. Available: <https://arxiv.org/abs/2309.01267>
- [75] B. Wu and C. S. Suh, "Deep Reinforcement Learning for Decentralized Multi-Robot Control: A DQN Approach to Robustness and Information Integration," arXiv:2408.11339, 2024. [Online]. Available: <https://arxiv.org/abs/2408.11339>
- [76] H. U. Sheikh and L. Bölöni, "Designing a Multi-Objective Reward Function for Creating Teams of Robotic Bodyguards Using Deep Reinforcement Learning," arXiv:1901.09837, 2019. [Online]. Available: <https://arxiv.org/abs/1901.09837>
- [77] L. Marzari, F. Trotti, E. Marchesini, and A. Farinelli, "Designing Control Barrier Function via Probabilistic Enumeration for Safe Reinforcement Learning Navigation," IEEE Robotics and Automation Letters, 2025. [Online]. Available: <https://doi.org/10.1109/LRA.2025.3596431>
- [78] J. Li and S. X. Yang, "Digital twins to embodied artificial intelligence: review and perspective," Intelligence & Robotics, 2025. [Online]. Available: <https://www.oapublish.com/articles/ir.2025.11>
- [79] W.-H. Hsieh, E. Hsieh, D. Niu, T. Darrell, R. Herzig, and D. M. Chan, "Do What? Teaching Vision-Language-Action Models to Reject the Impossible," arXiv:2508.16292, 2025. [Online]. Available: <https://arxiv.org/abs/2508.16292>
- [80] E. Nikonova, C. Xue, and J. Renz, "Don't do it: Safer Reinforcement Learning With Rule-based Guidance," arXiv:2212.13819, 2022. [Online]. Available: <https://arxiv.org/abs/2212.13819>
- [81] M. Ni, L. Zhang, Z. Chen, K. Bai, Z. Chen, J. Zhang, L. Zhang, and W. Zuo, "Don't Let Your Robot be Harmful: Responsible Robotic Manipulation via Safety-as-Policy," arXiv:2411.18289, 2025. [Online]. Available: <https://arxiv.org/abs/2411.18289>
- [82] Z. Zhu, B. Wu, Z. Zhang, L. Han, Q. Liu, and B. Wu, "EARBench: Towards Evaluating Physical Risk Awareness for Task Planning of Foundation Model-based Embodied AI Agents," arXiv:2408.04449, 2024. [Online]. Available: <https://arxiv.org/abs/2408.04449>
- [83] C. Zhao, G. Liu, R. Zhang, Y. Liu, J. Wang, J. Kang *et al.*, "Edge General Intelligence Through World Models and Agentic AI: Fundamentals, Solutions, and Challenges," arXiv:2508.09561, 2025. [Online]. Available: <https://arxiv.org/abs/2508.09561>
- [84] D. Hu, D. Lan, Y. Liu, J. Ning, J. Wang, Y. Yang, and Z. Pang, "Embodied AI Through Cloud-Fog Computing: A Framework for Everywhere Intelligence," 2024 IEEE 33rd International Symposium on Industrial Electronics (ISIE), 2024. [Online]. Available: <https://doi.org/10.1109/ISIE54533.2024.10595837>
- [85] J. Varley, S. Singh, D. Jain, K. Choromanski, A. Zeng, S. B. R. Chowdhury, A. Dubey, and V. Sindhwani, "Embodied AI with Two Arms: Zero-shot Learning, Safety and Modularity," arXiv:2404.03570, 2024. [Online]. Available: <https://arxiv.org/abs/2404.03570>
- [86] J. Perlo, A. Robey, F. Barez, L. Floridi, and J. MÅskander, "Embodied ai: Emerging risks and opportunities for policy action," arXiv preprint arXiv:2509.00117, 2025. [Online]. Available: <https://arxiv.org/abs/2509.00117>
- [87] T. Feng, X. Wang, Y.-G. Jiang, and W. Zhu, "Embodied AI: From LLMs to World Models," arXiv:2509.20021, 2025. [Online]. Available: <https://arxiv.org/abs/2509.20021>
- [88] F. Ni, M. Zhang, P. Li, Y. Yuan, L. Zhang, Y. Liu *et al.*, "Embodied Arena: A Comprehensive, Unified, and Evolving Evaluation Platform for Embodied AI," arXiv:2509.15273, 2025. [Online]. Available: <https://arxiv.org/abs/2509.15273>
- [89] W. XU, X. JI, C. YAN, Y. CHENG *et al.*, "Embodied artificial intelligence security and governance," Bulletin of Chinese Academy of Sciences (Chinese Version), 2024. [Online]. Available: <https://bulletinofcas.researchcommons.org/journal/vol40/iss3/4/>
- [90] N. Bredeche, E. Haasdijk, and A. Prieto, "Embodied Evolution in Collective Robotics: A Review," arXiv:1709.08992, 2017. [Online]. Available: <https://arxiv.org/abs/1709.08992>
- [91] G. Huang, "Embodied Intelligence," 2025. [Online]. Available: https://doi.org/10.1007/978-981-97-8440-0_8-1
- [92] D. Hafner, "Embodied Intelligence Through World Models," 2024. [Online]. Available: <https://utoronto.scholaris.ca/items/a99c7648-cc8f-4641-835d-5abb15389dde>
- [93] R. V. W. Putra, A. Marchisio, F. Zayer, J. Dias, and M. Shafique, "Embodied Neuromorphic Artificial Intelligence for Robotics: Perspectives, Challenges, and Research Development Stack," arXiv:2404.03325, 2024. [Online]. Available: <https://arxiv.org/abs/2404.03325>
- [94] Y. Yuan, H. Cui, Y. Huang, Y. Chen, F. Ni, Z. Dong, P. Li, Y. Zheng, H. Tang, and J. Hao, "Embodied-RL: Reinforced Embodied Reasoning for General Robotic Manipulation," arXiv:2508.13998, 2025. [Online]. Available: <https://arxiv.org/abs/2508.13998>
- [95] Q. Xie, S. Y. Min, P. Ji, Y. Yang, T. Zhang, K. Xu, A. Bajaj, R. Salakhutdinov, M. Johnson-Roberson, and Y. Bisk, "Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation," arXiv:2409.18313, 2024. [Online]. Available: <https://arxiv.org/abs/2409.18313>
- [96] C. Gao, B. Zhao, W. Zhang, J. Mao, J. Zhang, Z. Zheng *et al.*, "EmbodiedCity: A Benchmark Platform for Embodied Agent in Real-world City Environment," arXiv:2410.09604, 2024. [Online]. Available: <https://arxiv.org/abs/2410.09604>
- [97] W. Khan and M. Ishrat, "Embracing the Future: Navigating the Challenges and Solutions in Embodied Artificial Intelligence," Building Embodied AI Systems: The Agents, the Architecture Principles, Challenges, and Application Domains, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-68256-8_13
- [98] M. D. Buhl, B. Bucknall, and T. Masterson, "Emerging Practices in Frontier AI Safety Frameworks," arXiv preprint arXiv:2503.04746, 2025. [Online]. Available: <https://arxiv.org/abs/2503.04746>
- [99] J. Perlo, A. Robey, F. Barez, and J. MÅskander, "Emerging Risks from Embodied AI Require Urgent Policy Action," Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Position Paper Track, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2025/hash/e9f5279bb8c2e194503630623f81c9d4-Abstract-Position_Paper_Track.html
- [100] Q. Sun, P. Hong, T. D. Pala, V. Toh, U.-X. Tan, D. Ghosal, and S. Poria, "Emma-X: An Embodied Multimodal Action Model with Grounded Chain of Thought and Look-ahead Spatial Reasoning," arXiv:2412.11974, 2024. [Online]. Available: <https://arxiv.org/abs/2412.11974>
- [101] D. Li, T. Cai, T. Tang, W. Chai, K. R. Driggs-Campbell, and G. Wang, "EMMOE: A Comprehensive Benchmark for Embodied Mobile Manipulation in Open Environments," arXiv preprint arXiv:2503.08604, 2025. [Online]. Available: <https://arxiv.org/abs/2503.08604>
- [102] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: a survey," Machine Learning, 2017. [Online]. Available: <http://dx.doi.org/10.1007/s10994-017-5666-0>
- [103] X. Chi, C.-K. Fan, H. Zhang, X. Qi, R. Zhang, A. Chen *et al.*, "EVA: An Embodied World Model for Future Video Anticipation," arXiv:2410.15461, 2025. [Online]. Available: <https://arxiv.org/abs/2410.15461>
- [104] P. Valle, C. Lu, S. Ali, and A. Arrieta, "Evaluating uncertainty and quality of visual language action-enabled robots," arXiv preprint arXiv:2507.17049, 2025. [Online]. Available: <https://arxiv.org/abs/2507.17049>
- [105] T. Lin, G. Li, Y. Zhong, Y. Zou, Y. Du, J. Liu, E. Gu, and B. Zhao, "Evo-0: Vision-Language-Action Model with Implicit Spatial Understanding," arXiv:2507.00416, 2025. [Online]. Available: <https://arxiv.org/abs/2507.00416>
- [106] T. Wang, C. Han, J. C. Liang, W. Yang, D. Liu, L. X. Zhang, Q. Wang, J. Luo, and R. Tang, "Exploring the Adversarial Vulnerabilities of Vision-Language-Action Models in Robotics," arXiv:2411.13587, 2025. [Online]. Available: <https://arxiv.org/abs/2411.13587>
- [107] H. Ma, Y. Guan, S. E. Li, X. Zhang, S. Zheng, and J. Chen, "Feasible Actor-Critic: Constrained Reinforcement Learning for Ensuring Statewise Safety," arXiv:2105.10682, 2021. [Online]. Available: <https://arxiv.org/abs/2105.10682>
- [108] W. Wang, X. Zhou, B. Xu, S. Chen, M. Lu, J. Li, and Y. Gu, "Few-shot reasoning-based safe reinforcement learning framework for autonomous robot navigation," International Journal of Automation and Control, 2024. [Online]. Available: <https://www.inderscienceonline.com/doi/abs/10.1504/IJAAC.2024.135093>
- [109] C. Cao, Y. Xin, S. Wu, L. He, Z. Yan, J. Tan, and X. Wang, "FOSP: Fine-tuning Offline Safe Policy through World Models," arXiv:2407.04942, 2025. [Online]. Available: <https://arxiv.org/abs/2407.04942>
- [110] M. T. Khan and A. Waheed, "Foundation Model Driven Robotics: A Comprehensive Review," arXiv:2507.10087, 2025. [Online]. Available: <https://arxiv.org/abs/2507.10087>

- [111] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu *et al.*, “Foundation Models in Robotics: Applications, Challenges, and the Future,” arXiv:2312.07843, 2023. [Online]. Available: <https://arxiv.org/abs/2312.07843>
- [112] Y. Liu, X. Cao, T. Chen, Y. Jiang, J. You, M. Wu, X. Wang, M. Feng, Y. Jin, and J. Chen, “From screens to scenes: A survey of embodied AI in healthcare,” *Information Fusion*, 2025. [Online]. Available: <http://dx.doi.org/10.1016/j.inffus.2025.103033>
- [113] J. Jabbour and V. Janapa Reddi, “Generative AI Agents in Autonomous Machines: A Safety Perspective,” *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024. [Online]. Available: <https://doi.org/10.1145/3676536.3698390>
- [114] B. Hill, “Generative World Models of Tasks: LLM-Driven Hierarchical Scaffolding for Embodied Agents,” arXiv:2509.04731, 2025. [Online]. Available: <https://arxiv.org/abs/2509.04731>
- [115] Z. Zhou, X. Xie, J. Song, Z. Shu, and L. Ma, “GenSafe: A Generalizable Safety Enhancer for Safe Reinforcement Learning Algorithms Based on Reduced Order Markov Decision Process Model,” *IEEE Transactions on Neural Networks and Learning Systems*, 2025. [Online]. Available: <http://dx.doi.org/10.1109/TNNLS.2024.3496492>
- [116] L. Sun, B. Xie, Y. Liu, H. Shi, T. Wang, and J. Cao, “GeoVLA: Empowering 3D Representations in Vision-Language-Action Models,” arXiv:2508.09071, 2025. [Online]. Available: <https://arxiv.org/abs/2508.09071>
- [117] A. Ye, B. Wang, C. Ni, G. Huang, G. Zhao, H. Li *et al.*, “GigaWorld-0: World Models as Data Engine to Empower Embodied AI,” arXiv:2511.19861, 2025. [Online]. Available: <https://arxiv.org/abs/2511.19861>
- [118] T. Zhang, H. Duan, H. Hao, Y. Qiao, J. Dai, and Z. Hou, “Grounding Actions in Camera Space: Observation-Centric Vision-Language-Action Policy,” arXiv:2508.13103, 2025. [Online]. Available: <https://arxiv.org/abs/2508.13103>
- [119] J. Günster, P. Liu, J. Peters, and D. Tateo, “Handling Long-Term Safety and Uncertainty in Safe Reinforcement Learning,” arXiv:2409.12045, 2024. [Online]. Available: <https://arxiv.org/abs/2409.12045>
- [120] M. Guo, T. Liao, J. Wang, and Z. Li, “Hierarchical Motion Planning Under Probabilistic Temporal Tasks and Safe-Return Constraints,” *IEEE Transactions on Automatic Control*, 2023. [Online]. Available: <https://doi.org/10.1109/TAC.2023.3244884>
- [121] S. Kanta, M. Tavassoli, V. T. Chirkuri, V. A. Kumar, S. B. Punati, P. Damacharla, and S. Katyara, “Toward General Physical Intelligence for Resilient Agile Manufacturing Automation,” arXiv:2508.11960, 2025. [Online]. Available: <https://arxiv.org/abs/2508.11960>
- [122] W. Zhao, F. Li, and C. Liu, “Implicit Safe Set Algorithm for Provably Safe Reinforcement Learning,” arXiv:2405.02754, 2025. [Online]. Available: <https://arxiv.org/abs/2405.02754>
- [123] Y. Yin, Z. Wang, Y. Sharma, D. Niu, T. Darrell, and R. Herzig, “In-Context Learning Enables Robot Action Prediction in LLMs,” arXiv:2410.12782, 2025. [Online]. Available: <https://arxiv.org/abs/2410.12782>
- [124] S. Longpre, K. Klyman, R. E. Appel, S. Kapoor, R. Bommasani, M. Sahar *et al.*, “In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai,” arXiv preprint arXiv:2503.16861, 2025. [Online]. Available: <https://arxiv.org/abs/2503.16861>
- [125] D. Hendrycks, “Introduction to AI Safety, Ethics, and Society,” arXiv:2411.01042, 2024. [Online]. Available: <https://arxiv.org/abs/2411.01042>
- [126] A. Jiang, Y. Gao, Y. Wang, Z. Sun, S. Wang, Y. Heng *et al.*, “IRL-VLA: Training a Vision-Language-Action Policy via Reward World Model,” arXiv:2508.06571, 2025. [Online]. Available: <https://arxiv.org/abs/2508.06571>
- [127] Z. Zhu, X. Wang, W. Zhao, C. Min, B. Li, N. Deng *et al.*, “Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond,” arXiv:2405.03520, 2025. [Online]. Available: <https://arxiv.org/abs/2405.03520>
- [128] X. Lu, Z. Chen, X. Hu, Y. Zhou, W. Zhang, D. Liu, L. Sheng, and J. Shao, “IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks,” arXiv:2506.16402, 2025. [Online]. Available: <https://arxiv.org/abs/2506.16402>
- [129] A. Najib, S. Depeweg, and P. Swazinna, “Iterative Batch Reinforcement Learning via Safe Diversified Model-based Policy Search,” arXiv:2411.09722, 2024. [Online]. Available: <https://arxiv.org/abs/2411.09722>
- [130] T. Wulff, R. S. Maharjan, X. Chi, and A. Cangelosi, “Joint Action Language Modelling for Transparent Policy Execution,” arXiv preprint arXiv:2504.10055, 2025. [Online]. Available: <https://arxiv.org/abs/2504.10055>
- [131] C. Kennington, “Language Acquisition is Embodied, Interactive, Emotive: a Research Proposal,” arXiv:2105.04633, 2021. [Online]. Available: <https://arxiv.org/abs/2105.04633>
- [132] X. Yao, H. Zhou, O. Mees, Y. Meng, T. Xiao, Y. Bisk *et al.*, “Bridging Language and Action: A Survey of Language-Conditioned Robot Manipulation,” arXiv:2312.10807, 2025. [Online]. Available: <https://arxiv.org/abs/2312.10807>
- [133] X. Hu and Z. Xu, “Large language and vision-language models for robot: safety challenges, mitigation strategies and future directions,” *Industrial Robot: the international journal of robotics research and application*, 2025. [Online]. Available: <https://www.emerald.com/ir/article/doi/10.1108/IR-02-2025-0074/1269979/Large-language-and-vision-language-models-for>
- [134] W. Liang, R. Zhou, Y. Ma, B. Zhang, S. Li, Y. Liao, and P. Kuang, “Large Model Empowered Embodied AI: A Survey on Decision-Making and Embodied Learning,” arXiv:2508.10399, 2025. [Online]. Available: <https://arxiv.org/abs/2508.10399>
- [135] R. Shao, W. Li, L. Zhang, R. Zhang, Z. Liu, R. Chen, and L. Nie, “Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey,” arXiv:2508.13073, 2025. [Online]. Available: <https://arxiv.org/abs/2508.13073>
- [136] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng *et al.*, “Latent Action Pretraining from Videos,” arXiv:2410.11758, 2025. [Online]. Available: <https://arxiv.org/abs/2410.11758>
- [137] S. Zhou, Y. Du, Y. Yang, L. Han, P. Chen, D.-Y. Yeung, and C. Gan, “Learning 3D Persistent Embodied World Models,” arXiv:2505.05495, 2025. [Online]. Available: <https://arxiv.org/abs/2505.05495>
- [138] L. Berducci, S. Yang, R. Mangharam, and R. Grosu, “Learning Adaptive Safety for Multi-Agent Systems,” *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. [Online]. Available: <https://doi.org/10.1109/ICRA57147.2024.10611037>
- [139] J. C. Kiemel and T. Kröger, “Learning Collision-free and Torque-limited Robot Trajectories based on Alternative Safe Behaviors,” *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022. [Online]. Available: <https://doi.org/10.1109/Humanoids53995.2022.10000077>
- [140] K. Srinivasan, B. Eysenbach, S. Ha, J. Tan, and C. Finn, “Learning to be Safe: Deep RL with a Safety Critic,” arXiv:2010.14603, 2020. [Online]. Available: <https://arxiv.org/abs/2010.14603>
- [141] B. Lim, A. Reichenbach, and A. Cully, “Learning to walk autonomously via reset-free quality-diversity,” *Proceedings of the Genetic and Evolutionary Computation Conference*, 2022. [Online]. Available: <https://doi.org/10.1145/3512290.3528715>
- [142] A. HasanzadeZonuzi, A. Bura, D. Kalathil, and S. Shakkottai, “Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdp,” *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: <https://arxiv.org/abs/2008.00311>
- [143] H. Cheng, E. Xiao, Y. Wang, C. Yu, M. Sun, Q. Zhang *et al.*, “Manipulation Facing Threats: Evaluating Physical Vulnerabilities in End-to-End Vision Language Action Models,” arXiv:2409.13174, 2025. [Online]. Available: <https://arxiv.org/abs/2409.13174>
- [144] B. H’aon, K. Stocking, I. Chuang, and C. Tomlin, “Mechanistic interpretability for steering vision-language-action models,” arXiv preprint arXiv:2509.00328, 2025. [Online]. Available: <https://arxiv.org/abs/2509.00328>
- [145] J. Rothfuss, C. Koenig, A. Rupenyan, and A. Krause, “Meta-Learning Priors for Safe Bayesian Optimization,” arXiv:2210.00762, 2023. [Online]. Available: <https://arxiv.org/abs/2210.00762>
- [146] J. Klotz and S. K. Nayar, “Minimalist Vision with Freeform Pixels,” *Computer Vision – ECCV 2024*, 2024. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-73039-9_19
- [147] P. Lancaster, N. Hansen, A. Rajeswaran, and V. Kumar, “MoDem-V2: Visuo-Motor World Models for Real-World Robot Manipulation,” arXiv:2309.14236, 2024. [Online]. Available: <https://arxiv.org/abs/2309.14236>
- [148] G. Liu, J. De Winter, B. Vanderborght, A. Nowé, and D. Steckelmacher, “MoveRL: To a Safer Robotic Reinforcement Learning Environment,” *Benelux Conference on Artificial Intelligence*, 2021. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-93842-0_14
- [149] Z. Feng, R. Xue, L. Yuan, Y. Yu, N. Ding, M. Liu, B. Gao, J. Sun, X. Zheng, and G. Wang, “Multi-agent embodied ai: Advances and future directions,” arXiv preprint arXiv:2505.05108, 2025. [Online]. Available: <https://arxiv.org/abs/2505.05108>
- [150] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, “Multi-Modal Masked Autoencoders for Medical Vision-and-

- Language Pre-Training,” arXiv:2209.07098, 2022. [Online]. Available: <https://arxiv.org/abs/2209.07098>
- [151] R. Kaushik, K. Chatzilygeroudis, and J.-B. Mouret, “Multi-objective Model-based Policy Search for Data-efficient Learning with Sparse Rewards,” arXiv:1806.09351, 2018. [Online]. Available: <https://arxiv.org/abs/1806.09351>
 - [152] A. Kinose, M. Okada, R. Okumura, and T. Taniguchi, “Multi-view dreaming: Multi-view world model with contrastive learning,” *Advanced Robotics*, 2023. [Online]. Available: <https://arxiv.org/abs/2203.11024>
 - [153] G. Bakirtzis, A. A. Tubella, A. Theodorou, D. Danks, and U. Topcu, “Navigating the sociotechnical labyrinth: dynamic certification for responsible embodied AI,” *Bi-directionality in Human-AI Collaborative Systems*, 2025. [Online]. Available: <https://doi.org/10.1016/b978-0-44-340553-2.00019-8>
 - [154] P. Mineault, N. Zanichelli, J. Z. Peng, A. Arkhipov, E. Bingham, J. Jara-Ettinger *et al.*, “NeuroAI for AI Safety,” arXiv:2411.18526, 2025. [Online]. Available: <https://arxiv.org/abs/2411.18526>
 - [155] Q. Yang, J. A. Stork, and T. Stoyanov, “Null Space Based Efficient Reinforcement Learning with Hierarchical Safety Constraints,” 2021 European Conference on Mobile Robots (ECMR), 2021. [Online]. Available: <https://doi.org/10.1109/ECMR50962.2021.9568848>
 - [156] R. Sapkota and M. Karkee, “Object Detection with Multimodal Large Vision-Language Models: An In-depth Review,” arXiv:2508.19294, 2025. [Online]. Available: <https://arxiv.org/abs/2508.19294>
 - [157] X. Wu, S. Chakraborty, R. Xian, J. Liang, T. Guan, F. Liu, B. M. Sadler, D. Manocha, and A. S. Bedi, “On the Vulnerability of LLM/VLM-Controlled Robotics,” arXiv:2402.10340, 2025. [Online]. Available: <https://arxiv.org/abs/2402.10340>
 - [158] A. Balakrishna, B. Thananjeyan, J. Lee, F. Li, A. Zahed, J. E. Gonzalez, and K. Goldberg, “On-Policy Robot Imitation Learning from a Converging Supervisor,” arXiv:1907.03423, 2019. [Online]. Available: <https://arxiv.org/abs/1907.03423>
 - [159] J. Bruce, N. Suenderhauf, P. Mirowski, R. Hadsell, and M. Milford, “One-Shot Reinforcement Learning for Robot Navigation with Interactive Replay,” arXiv:1711.10137, 2017. [Online]. Available: <https://arxiv.org/abs/1711.10137>
 - [160] W. Gao, Z. Deng, Z. Niu, F. Rong, C. Chen, Z. Gong *et al.*, “OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue,” arXiv:2306.12174, 2023. [Online]. Available: <https://arxiv.org/abs/2306.12174>
 - [161] Y. Xu, “Opinion: A Unified World Model is the cornerstone for integrating perception, reasoning, and decision-making in embodied AI,” *NeurIPS 2025 Workshop on Embodied World Models for Decision Making*, 2025. [Online]. Available: <https://openreview.net/forum?id=WjmfSw2eg9>
 - [162] A. Lykov, J. Sam, H. K. Nguyen, V. Kozlovskiy, Y. Mahmoud, V. Serpiva, M. A. Cabrera, M. Konenkov, and D. Tsetserukou, “PhysicalAgent: Towards General Cognitive Robotics with Foundation World Models,” arXiv:2509.13903, 2025. [Online]. Available: <https://arxiv.org/abs/2509.13903>
 - [163] F. Den Hengst, V. François-Lavet, M. Hoogendoorn, and F. van Harmelen, “Planning for potential: efficient safe reinforcement learning,” *Machine Learning*, 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s10994-022-06143-6>
 - [164] A. Bennett, D. Misra, and N. Kallus, “Provable Safe Reinforcement Learning with Binary Feedback,” arXiv:2210.14492, 2022. [Online]. Available: <https://arxiv.org/abs/2210.14492>
 - [165] J. Thumm and M. Althoff, “Provably Safe Deep Reinforcement Learning for Robotic Manipulation in Human Environments,” arXiv:2205.06311, 2022. [Online]. Available: <https://arxiv.org/abs/2205.06311>
 - [166] D. Zhang, J. Sun, C. Hu, X. Wu, Z. Yuan, R. Zhou, F. Shen, and Q. Zhou, “Pure Vision Language Action (VLA) Models: A Comprehensive Survey,” arXiv:2509.19012, 2025. [Online]. Available: <https://arxiv.org/abs/2509.19012>
 - [167] V. Serpiva, A. Lykov, A. Myshlyayev, M. H. Khan, A. A. Abdulkarim, O. Sautenkov, and D. Tsetserukou, “RaceVLA: VLA-based Racing Drone Navigation with Human-like Behaviour,” arXiv:2503.02572, 2025. [Online]. Available: <https://arxiv.org/abs/2503.02572>
 - [168] W. Song, J. Chen, W. Li, X. He, H. Zhao, C. Cui *et al.*, “RationalVLA: A Rational Vision-Language-Action Model with Dual System,” arXiv:2506.10826, 2025. [Online]. Available: <https://arxiv.org/abs/2506.10826>
 - [169] T. Motoda, K. Makiyama, R. Nakajo, H. Oh, K. Shirai, R. Hanai *et al.*, “Recipe for Vision-Language-Action Models in Robotic Manipulation: A Survey,” *Authorea Preprints*, 2025. [Online]. Available: <https://arxiv.org/abs/2507.10672>
 - [170] A. Martinetti, P. K. Chemweno, K. Nizamis, and E. Fosch-Villaronga, “Redefining safety in light of human-robot interaction: A critical review of current standards and regulations,” *Frontiers in chemical engineering*, 2021. [Online]. Available: <https://www.frontiersin.org/journals/chemical-engineering/articles/10.3389/fceng.2021.666237/full>
 - [171] J. Thumm, G. Pelat, and M. Althoff, “Reducing Safety Interventions in Provably Safe Reinforcement Learning,” arXiv:2303.03339, 2023. [Online]. Available: <https://arxiv.org/abs/2303.03339>
 - [172] U. Berdica, M. Jackson, N. E. Veronese, J. Foerster, and P. Maiolino, “Reinforcement Learning Controllers for Soft Robots Using Learned Environments,” 2024 IEEE 7th International Conference on Soft Robotics (RoboSoft), 2024. [Online]. Available: <http://dx.doi.org/10.1109/RoboSoft60065.2024.10522003>
 - [173] F. Yang, W. Zhou, Z. Liu, D. Zhao, and D. Held, “Reinforcement Learning in a Safety-Embedded MDP with Trajectory Optimization,” arXiv:2310.06903, 2024. [Online]. Available: <https://arxiv.org/abs/2310.06903>
 - [174] K. Majd, G. Clark, G. Fainekos, and H. B. Amor, “Repairing Neural Networks for Safety in Robotic Systems using Predictive Models,” arXiv:2411.04408, 2024. [Online]. Available: <https://arxiv.org/abs/2411.04408>
 - [175] W. Niu, H. Ma, and Y. Zhai, “Research on Task Planning Methods for Embodied Intelligence Based on Large Language Models,” 2025 International Symposium on Intelligent Robotics and Systems (ISIRS), 2025. [Online]. Available: <https://doi.org/10.1109/ISIRS65690.2025.11168047>
 - [176] K. Rana, M. Xu, B. Tidd, M. Milford, and N. Sünderhauf, “Residual Skill Policies: Learning an Adaptable Skill-based Action Space for Reinforcement Learning for Robotics,” arXiv:2211.02231, 2022. [Online]. Available: <https://arxiv.org/abs/2211.02231>
 - [177] M. Lai, K. Go, Z. Li, T. Kroger, S. Schaal, K. Allen, and J. Scholz, “RoboBallet: Planning for Multi-Robot Reaching with Graph Neural Networks and Reinforcement Learning,” arXiv:2509.05397, 2025. [Online]. Available: <https://arxiv.org/abs/2509.05397>
 - [178] S. Li, J. Wang, R. Dai, W. Ma, W. Y. Ng, Y. Hu, and Z. Li, “RoboNurse-VLA: Robotic Scrub Nurse System based on Vision-Language-Action Model,” arXiv:2409.19590, 2024. [Online]. Available: <https://arxiv.org/abs/2409.19590>
 - [179] Y. Shang, X. Zhang, Y. Tang, L. Jin, C. Gao, W. Wu, and Y. Li, “RoboScape: Physics-informed Embodied World Model,” arXiv preprint arXiv:2506.23135, 2025. [Online]. Available: <https://arxiv.org/abs/2506.23135>
 - [180] F. Capuano, C. Pascal, A. Zouitine, T. Wolf, and M. Aractingi, “Robot Learning: A Tutorial,” arXiv:2510.12403, 2025. [Online]. Available: <https://arxiv.org/abs/2510.12403>
 - [181] P. Liu, D. Tateo, H. B. Ammar, and J. Peters, “Robot Reinforcement Learning on the Constraint Manifold,” *Proceedings of the 5th Conference on Robot Learning*, 2022. [Online]. Available: <https://proceedings.mlr.press/v164/liu22c.html>
 - [182] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, “Robotic Control via Embodied Chain-of-Thought Reasoning,” arXiv:2407.08693, 2025. [Online]. Available: <https://arxiv.org/abs/2407.08693>
 - [183] M. Hussing, J. A. Mendez, A. Singrodia, C. Kent, and E. Eaton, “Robotic Manipulation Datasets for Offline Compositional Reinforcement Learning,” arXiv:2307.07091, 2023. [Online]. Available: <https://arxiv.org/abs/2307.07091>
 - [184] R. Sakagami, F. S. Lay, A. D’omel, M. J. Schuster, A. Albu-Schäffer, and F. Stulp, “Robotic world models—conceptualization, review, and engineering best practices,” *Frontiers in Robotics and AI*, 2023. [Online]. Available: <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1253049/full>
 - [185] C. E. Mower, Y. Wan, H. Yu, A. Grosnit, J. Gonzalez-Billandon, M. Zimmer *et al.*, “ROS-LLM: A ROS framework for embodied AI with task feedback and structured reasoning,” arXiv:2406.19741, 2024. [Online]. Available: <https://arxiv.org/abs/2406.19741>
 - [186] R. Dang, Y. Yuan, Y. Mao, K. Li, J. Liu, Z. Wang, X. Li, F. Wang, and D. Zhao, “RynnEC: Bringing MLLMs into Embodied World,” arXiv:2508.14160, 2025. [Online]. Available: <https://arxiv.org/abs/2508.14160>
 - [187] S. Guist, J. Schneider, H. Ma, L. Chen, V. Berenz, J. Martus *et al.*, “Safe & Accurate at Speed with Tendons: A Robot Arm for Exploring Dynamic Motion,” arXiv:2307.02654, 2024. [Online]. Available: <https://arxiv.org/abs/2307.02654>
 - [188] H. Kwon, G. Lee, J. Lee, and S. Oh, “Safe CoR: A Dual-Expert Approach to Integrating Imitation Learning and Safe Reinforcement Learning Using Constraint Rewards,” 2024 IEEE/RSJ International

- Conference on Intelligent Robots and Systems (IROS), 2024. [Online]. Available: <https://doi.org/10.1109/iros58592.2024.10801577>
- [189] H. Cao, Z. Cai, H. Wei, W. Lu, L. Zhang, and H. Xiong, "Safe Reinforcement Learning for a Robot Being Pursued but with Objectives Covering More Than Capture-avoidance," arXiv:2207.00842, 2022. [Online]. Available: <https://arxiv.org/abs/2207.00842>
- [190] J. Ramirez and W. Yu, "Safe Reinforcement Learning for Learning from Human Demonstrations," 2023. [Online]. Available: <https://doi.org/10.21203/rs.3.rs-3195158/v1>
- [191] S. Huh and I. Yang, "Safe reinforcement learning for probabilistic reachability and safety specifications: A Lyapunov-based approach," arXiv:2002.10126, 2020. [Online]. Available: <https://arxiv.org/abs/2002.10126>
- [192] F. Khan, W. Feng, Z. Wang, T. Huang, X. Liu, Y. Cui, and W. Wang, "Safe reinforcement learning for vision-based robotic manipulation in human-centered environments," *International Journal of Intelligent Robotics and Applications*, 2025. [Online]. Available: <https://www.researchsquare.com/article/rs-6736564/v1>
- [193] L. Kovač and I. Farkaš, "Safe Reinforcement Learning in a Simulated Robotic Arm," *Lecture Notes in Computer Science*, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-44207-0_53
- [194] P. Liu, K. Zhang, D. Tateo, S. Jauhri, Z. Hu, J. Peters, and G. Chaltatzaki, "Safe Reinforcement Learning of Dynamic High-Dimensional Robotic Tasks: Navigation, Manipulation, Interaction," 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023. [Online]. Available: <https://doi.org/10.1109/icra48891.2023.10161548>
- [195] P. Liu, H. Bou-Ammar, J. Peters, and D. Tateo, "Safe Reinforcement Learning on the Constraint Manifold: Theory and Applications," *IEEE Transactions on Robotics*, 2025. [Online]. Available: <https://doi.org/10.1109/tro.2025.3567477>
- [196] Y. Emam, G. Notomista, P. Glotfelter, Z. Kira, and M. Egerstedt, "Safe Reinforcement Learning Using Robust Control Barrier Functions," *IEEE Robotics and Automation Letters*, 2025. [Online]. Available: <https://doi.org/10.1109/lra.2022.3216996>
- [197] Z. Chen, Z. Zhao, T. He, B. Chen, X. Zhao, L. Gong, and C. Liu, "Safe reinforcement learning via hierarchical adaptive chance-constraint safeguards," 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024. [Online]. Available: <https://arxiv.org/abs/2310.03379>
- [198] H. Zhang and Y. Guo, "Safe Reinforcement Learning with Contrastive Risk Prediction," arXiv:2209.09648, 2022. [Online]. Available: <https://arxiv.org/abs/2209.09648>
- [199] X. Lou, J. Zhang, Z. Wang, K. Huang, and Y. Du, "Safe Reinforcement Learning with Free-form Natural Language Constraints and Pre-Trained Language Models," arXiv:2401.07553, 2024. [Online]. Available: <https://arxiv.org/abs/2401.07553>
- [200] T.-Y. Yang, M. Hu, Y. Chow, P. J. Ramadge, and K. Narasimhan, "Safe Reinforcement Learning with Natural Language Constraints," arXiv:2010.05150, 2020. [Online]. Available: <https://arxiv.org/abs/2010.05150>
- [201] F. Schmoeller Roza, H. Rasheed, K. Roscher, X. Ning, and S. Günemann, "Safe Robot Navigation Using Constrained Hierarchical Reinforcement Learning," 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022. [Online]. Available: <https://doi.org/10.1109/icmla55696.2022.00123>
- [202] Q. Gu, Y. Ju, S. Sun, I. Gilitschenski, H. Nishimura, M. Itkina, and F. Shkurti, "SAFE: Multitask Failure Detection for Vision-Language-Action Models," arXiv:2506.09937, 2025. [Online]. Available: <https://arxiv.org/abs/2506.09937>
- [203] W. Zhang, X. Kong, T. Braunl, and J. B. Hong, "SafeEmbodAI: a Safety Framework for Mobile Robots in Embodied AI Systems," arXiv:2409.01630, 2024. [Online]. Available: <https://arxiv.org/abs/2409.01630>
- [204] I. Domkundwar, M. N. S. I. Bhola, and R. Kochhar, "Safeguarding AI Agents: Developing and Analyzing Safety Architectures," arXiv:2409.03793, 2025. [Online]. Available: <https://arxiv.org/abs/2409.03793>
- [205] R. Chen, Y. Sun, J. Wang, M. Lv, Q. Zhang, and Y. Zeng, "SafeMind: Benchmarking and Mitigating Safety Risks in Embodied LLM Agents," arXiv:2509.25885, 2025. [Online]. Available: <https://arxiv.org/abs/2509.25885>
- [206] A. V. S. Neto, J. B. Camargo, J. R. Almeida, and P. S. Cugnasca, "Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review on the State of the Art and Guidelines for Future Work," *IEEE Access*, 2022. [Online]. Available: <https://doi.org/10.1109/access.2022.3229233>
- [207] A. Goemans, M. D. Buhl, J. Schuett, T. Korbak, J. Wang, B. Hilton, and G. Irving, "Safety case template for frontier AI: A cyber inability argument," arXiv:2411.08088, 2024. [Online]. Available: <https://arxiv.org/abs/2411.08088>
- [208] Y. Qi, G. Kyebambo, S. Xie, W. Shen, S. Wang, B. Xie, B. He, Z. Wang, and S. Jiang, "Safety Control of Service Robots with LLMs and Embodied Knowledge Graphs," arXiv:2405.17846, 2024. [Online]. Available: <https://arxiv.org/abs/2405.17846>
- [209] L. Zhang, Z. Yan, L. Shen, S. Li, X. Wang, and D. Tao, "Safety Correction from Baseline: Towards the Risk-aware Policy in Robotics via Dual-agent Reinforcement Learning," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. [Online]. Available: <https://doi.org/10.1109/iros47612.2022.9981831>
- [210] D. Kim, Y. Kim, K. Lee, and S. Oh, "Safety Guided Policy Optimization," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. [Online]. Available: <https://doi.org/10.1109/iros47612.2022.9981030>
- [211] A. N. Abbas, S. Mehak, G. C. Chasparis, J. D. Kelleher, M. Guilfoyle, M. C. Leva, and A. K. Ramasubramanian, "Safety-Driven Deep Reinforcement Learning Framework for Cobots: A Sim2Real Approach," 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), 2024. [Online]. Available: <https://doi.org/10.1109/codit62066.2024.10708345>
- [212] M. Samson, B. Muraccioli, and F. Kanehiro, "Scalable, Training-Free Visual Language Robotics: a modular multi-model framework for consumer-grade GPUs," 2025 IEEE/SICE International Symposium on System Integration (SII), 2025. [Online]. Available: <https://arxiv.org/abs/2502.01071>
- [213] F. khalifa said ALSaidi, S. Ali AL'Abd AL-Busaidi, and B. Sundaravadivazhagan, "Securing Embodied AI: Addressing Cybersecurity Challenges in Physical Systems," *Building Embodied AI Systems: The Agents, the Architecture Principles, Challenges, and Application Domains*, 2025. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-68256-8_21
- [214] S. Neupane, S. Mitra, I. A. Fernandez, S. Saha, S. Mittal, J. Chen, N. Pillai, and S. Rahimi, "Security Considerations in AI-Robotics: A Survey of Current Methods, Challenges, and Opportunities," arXiv:2310.08565, 2024. [Online]. Available: <https://arxiv.org/abs/2310.08565>
- [215] A. Eftekhari, K.-H. Zeng, J. Duan, A. Farhadi, A. Kembhavi, and R. Krishna, "Selective Visual Representations Improve Convergence and Generalization for Embodied AI," arXiv:2311.04193, 2024. [Online]. Available: <https://arxiv.org/abs/2311.04193>
- [216] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware Adaptive Knowledge Distillation for Sensor-to-Vision Action Recognition," arXiv:2009.00210, 2020. [Online]. Available: <https://arxiv.org/abs/2009.00210>
- [217] A. Raffin, J. Kober, and F. Stulp, "Smooth Exploration for Robotic Reinforcement Learning," arXiv:2005.05719, 2020. [Online]. Available: <https://arxiv.org/abs/2005.05719>
- [218] B. Wu and C. S. Suh, "State-of-the-art in Robot Learning for Multi-Robot Collaboration: A Comprehensive Survey," arXiv:2408.11822, 2024. [Online]. Available: <https://arxiv.org/abs/2408.11822>
- [219] Y. Son, M. Kim, S. Kim, S. Han, J. Kim, D. Jang, Y. Yu, and C. Park, "Subtle Risks, Critical Failures: A Framework for Diagnosing Physical Safety of LLMs for Embodied Decision Making," arXiv:2505.19933, 2025. [Online]. Available: <https://arxiv.org/abs/2505.19933>
- [220] H. Li, Y. Chen, W. Cui, W. Liu, K. Liu, M. Zhou, Z. Zhang, and D. Zhao, "Survey of Vision-Language-Action Models for Embodied Manipulation," arXiv:2508.15201, 2025. [Online]. Available: <https://arxiv.org/abs/2508.15201>
- [221] T. Gupta, W. Gong, C. Ma, N. Pawlowski, A. Hilmkil, M. Scetbon *et al.*, "The Essential Role of Causality in Foundation World Models for Embodied AI," arXiv:2402.06665, 2024. [Online]. Available: <https://arxiv.org/abs/2402.06665>
- [222] I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal *et al.*, "The Ethics of Advanced AI Assistants," arXiv:2404.16244, 2024. [Online]. Available: <https://arxiv.org/abs/2404.16244>
- [223] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The Evolution of First Person Vision Methods: A Survey," arXiv:1409.1484, 2014. [Online]. Available: <https://arxiv.org/abs/1409.1484>
- [224] B. Hill, "The Physical Basis of Prediction: World Model Formation in Neural Organoids via an LLM-Generated Curriculum," arXiv:2509.04633, 2025. [Online]. Available: <https://arxiv.org/abs/2509.04633>
- [225] Y. Qi, Z. Pan, Y. Hong, M.-H. Yang, A. v. d. Hengel, and Q. Wu, "The Road to Know-Where: An Object-and-Room Informed Sequential

- BERT for Indoor Vision-Language Navigation,” arXiv:2104.04167, 2021. [Online]. Available: <https://arxiv.org/abs/2104.04167>
- [226] C.-P. Huang, Y.-H. Wu, M.-H. Chen, Y.-C. F. Wang, and F.-E. Yang, “ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Planning,” arXiv:2507.16815, 2025. [Online]. Available: <https://arxiv.org/abs/2507.16815>
- [227] F. Martín, E. Soriano-Salvador, J. M. Guerrero, G. Guardiola Múzquiz, J. C. Manzanares, and F. J. Rodríguez, “Towards a robotic intrusion prevention system: Combining security and safety in cognitive social robots,” *Robotics and Autonomous Systems*, 2025. [Online]. Available: <http://dx.doi.org/10.1016/j.robot.2025.104959>
- [228] S. Salimpour, L. Fu, K. Rachwał, P. Bertrand, K. O’Sullivan, R. Jakob *et al.*, “Towards Embodied Agentic AI: Review and Classification of LLM- and VLM-Driven Robot Autonomy and Interaction,” arXiv:2508.05294, 2025. [Online]. Available: <https://arxiv.org/abs/2508.05294>
- [229] L. He, Q.-S. Jia, A. Li, H. Sang, L. Wang, J. Lu *et al.*, “Towards provable probabilistic safety for scalable embodied AI systems,” arXiv:2506.05171, 2025. [Online]. Available: <https://arxiv.org/abs/2506.05171>
- [230] X. Tan, B. Liu, Y. Bao, Q. Tian, Z. Gao, X. Wu *et al.*, “Towards Safe and Trustworthy Embodied AI: Foundations, Status, and Prospects,” *Open Review*, 2025. [Online]. Available: <https://openreview.net/pdf?id=Eu6Yt21Alv>
- [231] H. Yu, W. Xu, and H. Zhang, “Towards Safe Reinforcement Learning with a Safety Editor Policy,” *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/11afefdd848d1bc9ac9f1604d9f45817-Paper-Conference.pdf
- [232] Z. Wang, Z. Zhou, J. Song, Y. Huang, Z. Shu, and L. Ma, “VLATest: Testing and Evaluating Vision-Language-Action Models for Robotic Manipulation,” *Proceedings of the ACM on Software Engineering*, 2025. [Online]. Available: <http://dx.doi.org/10.1145/3729343>
- [233] V. Palit, R. Pandey, A. Arora, and P. P. Liang, “Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP,” arXiv:2308.14179, 2023. [Online]. Available: <https://arxiv.org/abs/2308.14179>
- [234] R. Zheng, Y. Liang, S. Huang, J. Gao, H. D. III, A. Kolobov, F. Huang, and J. Yang, “TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies,” arXiv:2412.10345, 2025. [Online]. Available: <https://arxiv.org/abs/2412.10345>
- [235] Z. Liu, Y. Gu, S. Zheng, Y. Fu, X. Xue, and Y.-G. Jiang, “TriVLA: A Triple-System-Based Unified Vision-Language-Action Model with Episodic World Modeling for General Robot Control,” arXiv:2507.01424, 2025. [Online]. Available: <https://arxiv.org/abs/2507.01424>
- [236] C. Chen, X. Gong, Z. Liu, W. Jiang, S. Q. Goh, and K.-Y. Lam, “Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations,” arXiv:2408.12935, 2025. [Online]. Available: <https://arxiv.org/abs/2408.12935>
- [237] J. Ding, Y. Zhang, Y. Shang, Y. Zhang, Z. Zong, J. Feng *et al.*, “Understanding World or Predicting Future? A Comprehensive Survey of World Models,” *ACM Comput. Surv.*, 2025. [Online]. Available: <https://doi.org/10.1145/3746449>
- [238] L. Santos, Z. Li, L. Peters, S. Bansal, and A. Bajcsy, “Updating Robot Safety Representations Online from Natural Language Feedback,” arXiv:2409.14580, 2024. [Online]. Available: <https://arxiv.org/abs/2409.14580>
- [239] L. Dubois, K. Schmidt, C. Wang, J.-H. Park, L. Wang, and S. Munoz, “Video Event Reasoning and Prediction by Fusing World Knowledge from LLMs with Vision Foundation Models,” arXiv:2507.05822, 2025. [Online]. Available: <https://arxiv.org/abs/2507.05822>
- [240] H. Wang, R. Zhou, L. Ding, T. Liu, Z. Zhang, P. Xu, H. Gao, and Z. Deng, “Whole-Body Constrained Learning for Legged Locomotion via Hierarchical Optimization,” *IEEE Robotics and Automation Letters*, 2025. [Online]. Available: <https://doi.org/10.1109/LRA.2025.3577526>
- [241] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur *et al.*, “World models and predictive coding for cognitive and developmental robotics: frontiers and challenges,” *Advanced Robotics*, 2023. [Online]. Available: <https://arxiv.org/abs/2301.05832>
- [242] Z. Zeng, C. Zhang, F. Liu, J. Sifakis, Q. Zhang, S. Liu, and P. Wang, “World models: The safety perspective,” *2024 IEEE 35th International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2024. [Online]. Available: <https://arxiv.org/abs/2411.07690>
- [243] J. Zhang, M. Jiang, N. Dai, T. Lu, A. Uzunoglu, S. Zhang *et al.*, “World-in-World: World Models in a Closed-Loop World,” arXiv:2510.18135, 2025. [Online]. Available: <https://arxiv.org/abs/2510.18135>
- [244] D. Hendrycks and M. Mazeika, “X-Risk Analysis for AI Research,” arXiv:2206.05862, 2022. [Online]. Available: <https://arxiv.org/abs/2206.05862>
- [245] T. Patwardhan, R. Dias, E. Proehl, G. Kim, M. Wang, O. Watkins *et al.*, “GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks,” arXiv:2510.04374, 2025. [Online]. Available: <https://arxiv.org/abs/2510.04374>
- [246] P. Fung, Y. Bachrach, A. Celikyilmaz, K. Chaudhuri, D. Chen, W. Chung *et al.*, “Embodied AI Agents: Modeling the World,” arXiv:2506.22355, 2025. [Online]. Available: <https://arxiv.org/abs/2506.22355>
- [247] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, “AI2-THOR: An Interactive 3D Environment for Visual AI,” arXiv preprint arXiv:1712.05474, 2017. [Online]. Available: <https://arxiv.org/abs/1712.05474>
- [248] J. Guo, Z. Zheng, D. Ying, M. Jin, S. Gu, C. Spanos, and J. Lavaei, “Don’t Trade Off Safety: Diffusion Regularization for Constrained Offline RL,” arXiv:2502.12391, 2025. [Online]. Available: <https://arxiv.org/abs/2502.12391>
- [249] T. Tomilin, M. Fang, and M. Pechenizkiy, “HASARD: A Benchmark for Vision-Based Safe Reinforcement Learning in Embodied Agents,” *ICLR* 2025, 2025. [Online]. Available: <https://arxiv.org/abs/2503.08241>
- [250] S. S. Zhan, P. Wang, Y. Liu, Y. Peng, Z. Wang, Q. Wang *et al.*, “SENTINEL: A Multi-Level Formal Framework for Safety Evaluation of Foundation Model-based Embodied Agents,” arXiv:2510.12985, 2025. [Online]. Available: <https://arxiv.org/abs/2510.12985>
- [251] W. Xiao, T.-H. Wang, and D. Rus, “ABNet: Attention BarrierNet for Safe and Scalable Robot Learning,” arXiv:2406.13025, 2024. [Online]. Available: <https://arxiv.org/abs/2406.13025>
- [252] M. Guerrier, H. Fouad, and G. Beltrame, “Learning Control Barrier Functions and their application in Reinforcement Learning: A Survey,” arXiv:2404.16879, 2024. [Online]. Available: <https://arxiv.org/abs/2404.16879>
- [253] N. Vertovec, F. B. Mathiesen, T. Badings, L. Laurenti, and A. Abate, “Scalable Verification of Neural Control Barrier Functions Using Linear Bound Propagation,” arXiv:2511.06341, 2025. [Online]. Available: <https://arxiv.org/abs/2511.06341>
- [254] J. Kiemel, L. Righetti, T. Kröger, and T. Asfour, “Safe Reinforcement Learning of Robot Trajectories in the Presence of Moving Obstacles,” arXiv:2411.05784, 2024. [Online]. Available: <https://arxiv.org/abs/2411.05784>
- [255] J. Lee, Y. Kim, S. Kim, Q. Nguyen, and Y. Heo, “Learning Fast, Tool aware Collision Avoidance for Collaborative Robots,” arXiv:2508.20457, 2025. [Online]. Available: <https://arxiv.org/abs/2508.20457>
- [256] S. Li, Z. Ma, F. Liu, J. Lu, Q. Xiao, K. Sun, L. Cui, X. Yang, P. Liu, and X. Wang, “Safe Planner: Empowering Safety Awareness in Large Pre-Trained Models for Robot Task Planning,” *AAAI* 2025, 2025. [Online]. Available: <https://arxiv.org/abs/2411.06920>
- [257] L. Wang, Z. Ying, X. Yang, Q. Zou, Z. Yin, T. Li, J. Yang, Y. Yang, A. Liu, and X. Liu, “RoboSafe: Safeguarding Embodied Agents via Executable Safety Logic,” arXiv:2512.21220, 2025. [Online]. Available: <https://arxiv.org/abs/2512.21220>
- [258] K. Chen, Y. Lei, H. Cheng, H. Wu, W. Sun, and S. Zheng, “FREA: Feasibility-Guided Generation of Safety-Critical Scenarios with Reasonable Adversariality,” *CoRL* 2024, 2024. [Online]. Available: <https://arxiv.org/abs/2406.02983>
- [259] L. Manda, S. Chen, and M. Fazlyab, “Learning Performance-oriented Control Barrier Functions Under Complex Safety Constraints and Limited Actuation,” *CoRL* 2024, 2024. [Online]. Available: <https://arxiv.org/abs/2401.05629>
- [260] L. Brunke, Y. Zhang, R. Römer, J. Naimier, N. Staykov, S. Zhou, and A. P. Schoellig, “Semantically Safe Robot Manipulation: From Semantic Scene Understanding to Motion Safeguards,” arXiv:2410.15185, 2024. [Online]. Available: <https://arxiv.org/abs/2410.15185>
- [261] Y. Huang, J. Li, B. Ma, X. Dai, M. Xu, K. Xu, Y. Zhang, J. Wang, and X. Cheng, “Beyond Model Jailbreak: Systematic Dissection of the Ten Deadly Sins in Embodied AI Security,” arXiv:2512.06387, 2025. [Online]. Available: <https://arxiv.org/abs/2512.06387>
- [262] E. K. Jones, A. Robey, A. Zou, Z. Ravichandran, G. J. Pappas, H. Hassani, M. Fredrikson, and J. Z. Kolter, “Adversarial Attacks on Robotic Vision Language Action Models,” arXiv:2506.03350, 2025. [Online]. Available: <https://arxiv.org/abs/2506.03350>
- [263] H. Xu, Y. S. Koh, S. Huang, Z. Zhou, D. Wang, J. Sakuma, and J. Zhang, “Model-agnostic Adversarial Attack and Defense for

- Vision-Language-Action Models,” arXiv:2510.13237, 2025. [Online]. Available: <https://arxiv.org/abs/2510.13237>
- [264] T. Zhang, L. Wang, X. Zhang, Y. Zhang, B. Jia, S. Liang, S. Hu, Q. Fu, A. Liu, and X. Liu, “Visual Adversarial Attack on Vision-Language Models for Autonomous Driving,” arXiv:2411.18275, 2024. [Online]. Available: <https://arxiv.org/abs/2411.18275>
- [265] K. Hu, W. Yu, L. Zhang, A. Robey, A. Zou, C. Xu, H. Hu, and M. Fredrikson, “Transferable Adversarial Attacks on Black-Box Vision-Language Models,” arXiv:2505.01050, 2025. [Online]. Available: <https://arxiv.org/abs/2505.01050>
- [266] A. A. Khan, M. Andrev, M. A. Murtaza, S. Aguilera, R. Zhang, J. Ding, S. Hutchinson, and A. Anwar, “Safety Aware Task Planning via Large Language Models in Robotics,” arXiv:2503.15707, 2025. [Online]. Available: <https://arxiv.org/abs/2503.15707>
- [267] N. Akbarzadeh, Y. Adulyasak, and E. Delage, “Risk-Aware Decision Making in Restless Bandits: Theory and Algorithms for Planning and Learning,” arXiv:2410.23029, 2024. [Online]. Available: <https://arxiv.org/abs/2410.23029>
- [268] A. Mangararis, V. Giammarino, A. H. Qureshi, and S. Jagannathan, “Formal Methods in Robot Policy Learning and Verification: A Survey on Current Techniques and Future Directions,” arXiv:2602.06971, 2026. [Online]. Available: <https://arxiv.org/abs/2602.06971>
- [269] B. Derajić, S. Bernhard, and W. Hönig, “CN-CBF: Composite Neural Control Barrier Function for Safe Robot Navigation in Dynamic Environments,” arXiv:2603.06921, 2026. [Online]. Available: <https://arxiv.org/abs/2603.06921>
- [270] J. Mootz and R. Akhavan, “Control Barrier Functions with Audio Risk Awareness for Robot Safe Navigation on Construction Sites,” arXiv:2602.12416, 2026. [Online]. Available: <https://arxiv.org/abs/2602.12416>
- [271] H. Jin, N. U. Shinde, S. Atar, H. Yu, D. Hirsch, S. Gao, M. C. Yip, and S. Herbert, “Learning to Nudge: A Scalable Barrier Function Framework for Safe Robot Interaction in Dense Clutter,” arXiv:2601.02686, 2026. [Online]. Available: <https://arxiv.org/abs/2601.02686>
- [272] K. Wong, M. Stölzle, W. Xiao, C. D. Santina, D. Rus, and G. Zardini, “Contact-Aware Safety in Soft Robots Using High-Order Control Barrier and Lyapunov Functions,” arXiv:2505.03841, 2025. [Online]. Available: <https://arxiv.org/abs/2505.03841>
- [273] M. C. Darling, A. H. Hesu, M. A. Mardikes, B. C. McGuigan, and R. M. Milewicz, “Toward Maturity-Based Certification of Embodied AI: Quantifying Trustworthiness Through Measurement Mechanisms,” arXiv:2601.03470, 2026. [Online]. Available: <https://arxiv.org/abs/2601.03470>
- [274] T. Chakraborty, U. Ghosh, A. E. Gongora, R. Glatt, Y. Dong, J. Li, A. K. Roy-Chowdhury, and C. Song, “Toward Autonomous Laboratory Safety Monitoring with Vision Language Models: Learning to See Hazards Through Scene Structure,” arXiv:2602.00414, 2026. [Online]. Available: <https://arxiv.org/abs/2602.00414>
- [275] H. Jiang, Y. Chaudhary, Y. Wang, Z. Wang, R. Sharma, M. Mehta *et al.*, “NavTrust: Benchmarking Trustworthiness for Embodied Navigation,” arXiv:2603.19229, 2026. [Online]. Available: <https://arxiv.org/abs/2603.19229>
- [276] S. Gao, J. Yao, H. Wen, Y. Guo, Z. Liu, and H. Huang, “HomeSafeBench: A Benchmark for Embodied Vision-Language Models in Free-Exploration Home Safety Inspection,” arXiv:2509.23690, 2025. [Online]. Available: <https://arxiv.org/abs/2509.23690>
- [277] S. Farrell, C. Li, H. Yu, R. Yoshimitsu, S. Gao, and H. I. Christensen, “Safe Human Robot Navigation in Warehouse Scenario,” arXiv:2503.21141, 2025. [Online]. Available: <https://arxiv.org/abs/2503.21141>
- [278] B. Ma, H. Guo, P. Lv, M. Xu, X. Dai, Y. Zhang, Y. Yang, and Y. Zhang, “What Breaks Embodied AI Security: LLM Vulnerabilities, CPS Flaws, or Something Else?” arXiv:2602.17345, 2026. [Online]. Available: <https://arxiv.org/abs/2602.17345>
- [279] A. Odínokov and R. Yavorskiy, “A Hazard-Informed Data Pipeline for Robotics Physical Safety,” arXiv:2603.06130, 2026. [Online]. Available: <https://arxiv.org/abs/2603.06130>
- [280] A. P. L. Torre, “Adversarial attacks against Modern Vision-Language Models,” arXiv:2603.16960, 2026. [Online]. Available: <https://arxiv.org/abs/2603.16960>
- [281] I. Jang, J. Park, S. Cho, C. E. Mballo, C. J. Tomlin, and H. J. Kim, “EigenSafe: A Spectral Framework for Learning-Based Probabilistic Safety Assessment,” arXiv:2509.17750, 2026. [Online]. Available: <https://arxiv.org/abs/2509.17750>
- [282] V. Mayoral-Vilches, U. Ayucar-Carbajo, O. Laflamme, R. Peng, M. Sanz-Gómez, F. Balassone, L. Apa, and E. Gil-Urriarte, “Cybersecurity AI: Hacking Consumer Robots in the AI Era,” arXiv:2603.08665, 2026. [Online]. Available: <https://arxiv.org/abs/2603.08665>
- [283] DeepMind, “Generating Robot Constitutions & Benchmarks for Semantic Safety,” arXiv:2503.08663, 2026. [Online]. Available: <https://arxiv.org/html/2503.08663v1>
- [284] K. Xie, C. Li, H. Zhang, and N. Y. Chong, “Multimodal Adversarial Quality Policy for Safe Grasping,” arXiv:2603.01479, 2026. [Online]. Available: <https://arxiv.org/abs/2603.01479>
- [285] A. Zhang, Z. Zhuang, C. Wang, S. S. Ge, F. Shi, and C. Xiang, “SAC-LoCo: Safe and Adjustable Compliant Quadrupedal Locomotion,” arXiv:2509.23223, 2025. [Online]. Available: <https://arxiv.org/abs/2509.23223>
- [286] L. Yang, B. Werner, M. de Sa, and A. D. Ames, “CBF-RL: Safety Filtering Reinforcement Learning in Training with Control Barrier Functions,” arXiv:2510.14959, 2025. [Online]. Available: <https://arxiv.org/abs/2510.14959>
- [287] B. Kim and H. J. Kim, “Deep QP Safety Filter: Model-free Learning for Reachability-based Safety Filter,” arXiv:2601.21297, 2026. [Online]. Available: <https://arxiv.org/abs/2601.21297>
- [288] G. N. Sue, Y. Choudhary, R. Desatnik, C. Majidi, J. Dolan, and G. Shi, “Q-learning-based Model-free Safety Filter,” arXiv:2411.19809, 2024. [Online]. Available: <https://arxiv.org/abs/2411.19809>
- [289] MIT, “From Chat Control to Robot Control: The Backdoors Left Open for the Sake of Safety,” arXiv:2601.02205, 2026. [Online]. Available: <https://arxiv.org/html/2601.02205v1>