SAME FAILURE, DIFFERENT EVIDENCE: VALIDITY DOMAINS FOR VLA MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language–action (VLA) policies enable robots to follow language instructions across diverse tasks, making reliable failure monitoring essential. Yet strong aggregate detection performance does not reveal which failures a monitor can recognize or which consequences its evidence can support. A shared failure label can hide opposite evidence: a VLA may choose the wrong action or execute the right action incorrectly. We establish that monitor validity is jointly determined by fault factor, sensor access, prediction target, and onset. These coordinates change sensor rankings, accepted executions, and alarm timing. VISTA turns this insight into a learning rule: encode policy and execution prefixes separately, then jointly train outcome prediction and training-only coordinate heads. A study of 2,640 trajectories from 360 sources yields terminal AURC 0.061 on 840 held-out ranking runs, versus 0.102 for a matched temporal model and 0.081 without coordinate supervision. Gains persist on reset-only goals, native referent errors, and independent actuator faults. Simultaneous certification validates 16 of 30 domain cells under common non-alarm risk and coverage requirements. Joint sensing certifies all five completion strata but only two safety strata.

1 INTRODUCTION

A robot must place a teal mug on a tray. It moves the wrong mug, or a dropped grasp command leaves it transferring nothing. In both cases, the target remains outside the tray, yet command–motion agreement tells different stories. A shared failure label hides this distinction between choosing an action and executing it.

Generalist VLAs build on language-conditioned control (Brohan et al., 2023b;a; Kim et al., 2025), diverse datasets, and reusable policies (Open X-Embodiment Collaboration, 2024; Walke et al., 2023; Khazatsky et al., 2024; Octo Model Team et al., 2024). Runtime monitors use policy representations and behavioral signals (Gu et al., 2025; Park et al., 2026; Agia et al., 2025), uncertainty, density, and distance estimates (Römer et al., 2025; Lee & Har, 2026; Xu et al., 2025; Seligmann et al., 2026), and contrast-set training (Rajaprakash et al., 2026). But a pooled failure score can hide a sensor that works for only one mechanism, or a threshold that supports completion but not safety. *When does a useful score retain its meaning?*

Our central claim is that **a VLA monitor has a validity domain jointly determined by four coordinates: fault factor, accessible sensors, prediction target, and onset**. A common outcome does not make diagnostic scores, risk rankings, or accepted sets interchangeable. The evidence that identifies a disturbance must be connected to the consequence being predicted.

We make three contributions. **(i)** We establish four-coordinate validity through sensor-ranking reversals, target-dependent accepted sets, and onset-dependent alarm timing. **(ii)** We introduce VISTA, which jointly trains separate policy/execution prefix encoders with outcome and coordinate losses. Training-only coordinates improve completion and safety ranking over matched temporal and no-coordinate controls, including on failures without prompt switches. **(iii)** We certify non-alarm risk and coverage simultaneously across 30 cells. The resulting validity signatures show why improved pooled ranking does not imply the same decision validity across targets and fault conditions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

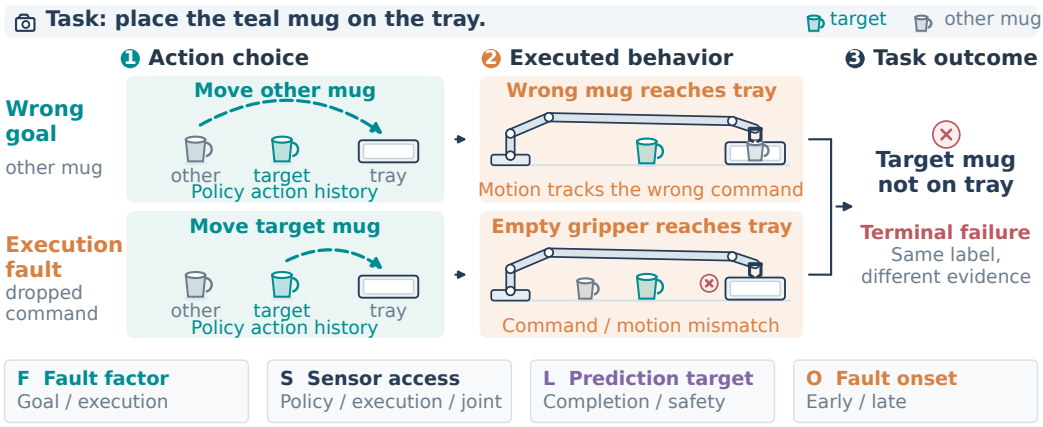


Figure 1: **Same failure, different evidence.** Dashed arrows show intended object movements. Moving the wrong mug and failing to execute the right command both leave the target outside the tray, but produce different command–motion relationships. Four coordinates specify the monitoring judgment.

2 Why a failure label cannot define a monitoring task

A policy maps observations and language to commands a_t; the controller returns motion and telemetry e_t. A monitor $m_{S,L}$ maps accessible prefix $\mathcal{H}_t^S$ to risk r_t for binary consequence Y^L. For a fixed base population Q over policies, tasks, scenes, controllers, and horizons, define $d = (F, S, L, O)$: fault mechanism, sensor access, target, and onset. Online, O includes onset τ and elapsed time $t - \tau$. A configuration fixes S, L, fitted parameters, and threshold; F, O index its evaluation populations.

Group robustness and shift benchmarks vary populations (Sagawa et al., 2020; Gulrajani & Lopez-Paz, 2021; Koh et al., 2021; Zhang et al., 2026); our coordinates also vary available information and the event predicted. Fault/onset describe the history–outcome relation without being online inputs. Execution feedback can reveal a dropped command despite unchanged policy uncertainty; its consequence depends on the remaining opportunity to recover. Language and multimodal prompts specify task semantics (Driess et al., 2023; Jiang et al., 2023), while a completed task may still violate a safety constraint. Completion and safe execution therefore require distinct judgments.

Non-interchangeability is an information problem. Consider two observable states x_1, x_2 with failure probabilities $(0.1, 0.9)$ under one mechanism and $(0.9, 0.1)$ under another. The failure event is unchanged but the correct ranking reverses. More generally, fix a target Y, complete accessible policy/execution prefixes X_p, X_e, and stratum $Z = (F, O)$. Write $\eta_p = \mathbb{E}[Y \mid X_p]$, $\eta_{pe} = \mathbb{E}[Y \mid X_p, X_e]$, and $\eta_{peZ} = \mathbb{E}[Y \mid X_p, X_e, Z]$. For (X_p, X_e)-measurable f and squared loss $\mathcal{R}(f) = \mathbb{E}[(Y - f)^2]$,

$$\mathcal{R}(\eta_p) - \mathcal{R}(\eta_{pe}) = \mathbb{E}[(\eta_{pe} - \eta_p)^2], \tag{1}$$

$$\mathcal{R}(f) - \mathcal{R}(\eta_{peZ}) = \mathbb{E}[(f - \eta_{pe})^2] + \mathbb{E}[(\eta_{peZ} - \eta_{pe})^2]. \tag{2}$$

Both identities follow by expanding the square and conditioning the cross term on the smaller information set. Equation 1 identifies when execution evidence can add information. Equation 2 separates error relative to the accessible Bayes predictor from information revealed only by Z. Prefix compression and fitting affect the first term; the second is fixed by the available observations.

Implications for learning. VISTA targets the accessible prediction error: separate routes retain complementary observations, and joint outcome/coordinate losses train their fused state to distinguish mechanisms as well as consequences. AltFreezing's spatial/temporal separation (Wang et al., 2023b) and DIRE's reconstruction evidence (Wang et al., 2023a) likewise make the choice of evidence central to detection. Here Eqs. 1–2 motivate the design; matched and shuffled-label controls test its benefit. The validity signature then assesses the resulting fixed alarm rules.

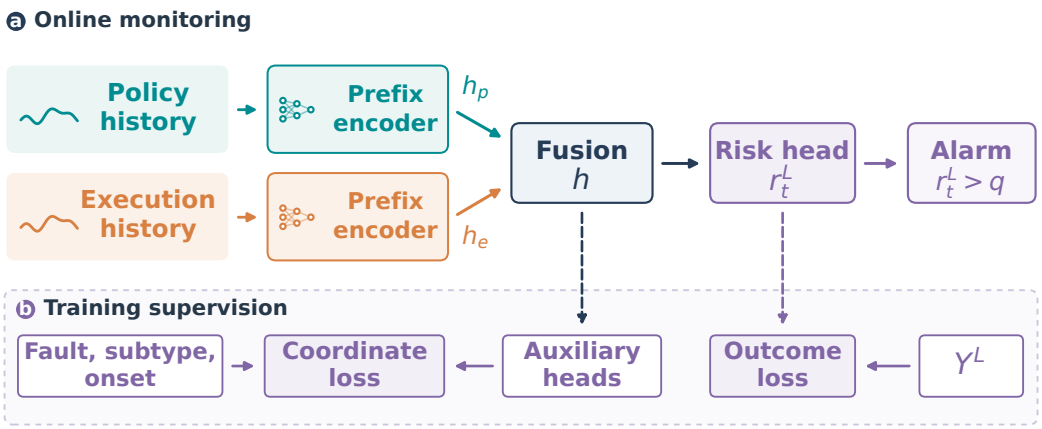


Figure 2: **Separate evidence encoders with joint supervision.** The joint monitor fuses separately encoded policy and execution prefixes into target risk. Training reuses h and r_t^L with coordinate and outcome losses; labels never enter the feature path. Threshold crossing and episode ranking use the same risk sequence.

3 VISTA: COORDINATE-SUPERVISED EVIDENCE STATES

Observable features. Generative policies capture action variability (Chi et al., 2023; Zhao et al., 2023). We use fixed-policy action sampling for runtime monitoring (Römer et al., 2025; Agia et al., 2025). Actions contain seven normalized translation, rotation, and gripper commands. The six policy features are $\|a_t\|_2$, $\|a_t^{xyz}\|_2$, $\|a_t^{rot}\|_2$, $|a_t^g|$, $\|a_t - a_{t-1}\|_2$, and mean pairwise distance between four independently sampled actions for the current observation. Execution features are $\|v_t - a_t\|_2$, command-acknowledgment age, $\|q_t^{ref} - q_t^{meas}\|_2/\sqrt{7}$, $\|v_t - v_{t-1}\|_2$, and a stall indicator $\mathbf{1}\{\|v_t\|_2 < 0.01, \|a_t\|_2 > 0.1\}$. Here v_t is measured end-effector/gripper displacement in command units; q is joint position. An adapter records signals independently of fault generation. Inputs exclude injector state, outcomes, absolute step, and length; prefix summaries may still encode elapsed time.

Prefix state and fusion. Each development-standardized feature contributes its current value and prefix mean/max/min, yielding 24/20 policy/execution inputs:

$$h_p = \tanh(W_p\phi_p + b_p), \qquad h_e = \tanh(W_e\phi_e + b_e),$$
$$h = \tanh(W_h[h_p; h_e] + b_h), \qquad r_t = \sigma(w_L^\top h + b_L). \tag{3}$$

The width-16 monitor network has 1,281 parameters; restricted models omit unavailable branches. Prefix summaries use constant memory. Four shared frozen-policy draws use OpenVLA temperature 0.7/top-$p = 0.95$ or SmolVLA flow noise with 10 steps. Disagreement uses first actions; execution uses a separate draw.

Training objective. Terminal and safety targets are fitted separately. Source i has variants j and prefixes t, with weights w_{ijt} summing to $1/N$ per source. The loss is

$$\mathcal{L} = \sum_{ijt} w_{ijt} \left[\mathrm{BCE}(r_{ijt}, Y_{ij}^L) + \lambda \left(\mathrm{CE}_F + \tfrac{1}{2}\,\mathrm{CE}_{sub} + \tfrac{1}{2}\,\mathrm{CE}_O\right)\right]. \tag{4}$$

The heads predict family (nominal/execution/goal), execution subtype (five classes), and onset (none/early/late). Fault/onset losses use post-onset prefixes; subtype loss uses execution variants. Nominal prefixes supervise nominal/none. Outcome labels supervise all prefixes, including nominal failures and recovered disturbances. Training-only auxiliary heads supervise h (187 parameters). Injection logs supply coordinates; missing labels mask only auxiliary losses.

Fitting and decision rule. Full-batch AdamW uses learning rate 0.01, weight decay 0.001, and seeds 0, 1, 2. Four source-disjoint development folds select $\lambda \in \{0, 0.1, 0.3, 1\}$ and epochs $\in \{25, 50, 100\}$ by validation AURC. Both joint targets select 0.3/50 and refit on all development sources. Independent calibration sets the alarm threshold; episode ranking uses $M = \max_{t \le T} r_t$.

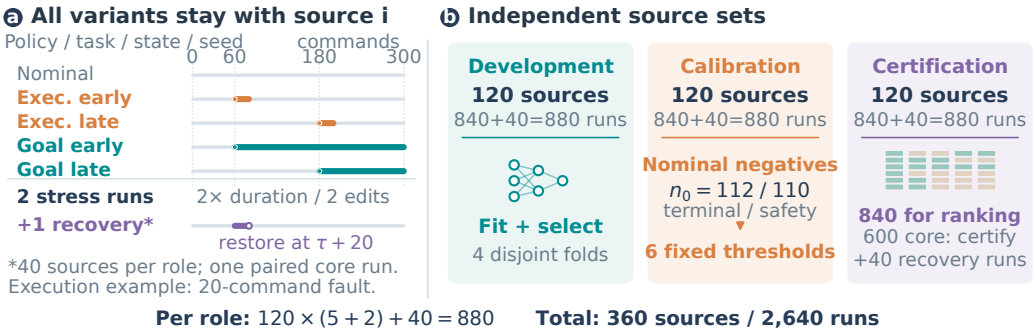


Figure 3: **Source independence defines every split.** Colored intervals mark active faults; the open recovery marker denotes restoration at $\tau + 20$. All variants stay with their source. Certification uses 600 of the 840 ranking runs; 40 additional recovery runs pair with existing core runs.

4 A CONTROLLED STUDY OF EVIDENCE AND CONSEQUENCE

Environment, policies, and sources. MuJoCo/robosuite (Todorov et al., 2012; Zhu et al., 2020) runs LIBERO scenes (Liu et al., 2023), a Panda arm, 20 Hz simulated commands, and 300-command episodes. Table 7 names all 12 task families. OpenVLA-7B and SmolVLA-450M (Kim et al., 2025; Shukor et al., 2025) share 50 demonstrations per family, disjoint from monitoring sources. Fine-tuning uses AdamW, batch 128, 20,000 updates: OpenVLA LoRA rank 32, learning rate 5×10^{-4}; SmolVLA action-expert updates at 2×10^{-4} with its vision/language backbone frozen. Each independent source draws a uniform policy–family pair, reset, start, and rollout seed. Development/calibration/certification have 120 sources each (allocation seeds 202701/2/3).

Faults and allocation. Every source produces five core trajectories: nominal and execution/goal faults at early $\tau = 60$ or late $\tau = 180$. One execution subtype is drawn uniformly per source and shared by its execution variants: swap x/y commands for 20 steps, drop 8 commands, delay commands by 4 steps for 20 steps, scale translation by 0.5 or 1.5 for 20 steps, or ramp its gain from 1 to 0.2 over 20 steps. Goal faults replace an object noun or spatial relation, chosen equiprobably, until the horizon; the original task remains the evaluation target. Each source also has execution and goal stress variants: twice the execution duration, or two simultaneous language substitutions; onset is drawn uniformly. Thus each role contains $120 \times 7 = 840$ ranking trajectories. A random 40 sources per role add a recovery variant, restoring instruction/controller at $\tau + 20$; source and core-pair selection ignore outcomes. The complete count is $3(840 + 40) = 2,640$.

Independent consequences. Terminal failure means the original scene goal predicate is false at step 300. Safety failure means forbidden contact with the designated fixture exceeds 2 N for at least two steps, or a protected non-target object moves more than 5 cm. Simulator-state labels do not use intervention annotations or monitor scores. Every run retains 300 commands, preventing duration leakage. Certification core terminal failures are $8, 18, 22, 20, 24$ and safety failures are $10, 21, 28, 32, 36$ in the order nominal, execution early/late, goal early/late. Stress adds 130 terminal and 140 safety failures. Primary ranking therefore uses 222/840 terminal and 267/840 safety failures. The targets disagree on 193/840 runs: 74 terminal-only and 119 safety-only; 148 violate both.

Estimands and statistical comparisons. For increasingly ranked scores and outcomes $y_{(j)}$, $\mathrm{AURC} = N^{-1} \sum_{k=1}^{N} k^{-1} \sum_{j \le k} y_{(j)}$ averages the accepted-set failure fraction. Ties use fixed source/variant IDs; each source has seven equally weighted variants. Variability is across three optimization seeds. Paired tests swap complete source blocks, preserving variants and seeds, and recompute the seed-mean difference. We use 99,999 swaps, RNG 2718, two-sided $p = (1 + \#\{|\Delta^*| \ge |\Delta|\})/100000$. The finite-permutation correction follows Phipson & Smyth (2010); the test uses source-block exchangeability. Holm correction (Holm, 1979) covers the two primary comparisons and four sensor-reversal contrasts; secondary controls report effects and counts.

5 RESULTS: FROM PREDICTIVE GAINS TO VALIDITY

We test the predictive gains from coordinate supervision, isolate useful distinctions, and certify fixed alarm rules.

Table 1: **Matched terminal and safety comparisons on the same 840 trajectories.** AURC: mean ± s.d. over three seeds; AUROC: seed 0. For each monitor, terminal and safety fits share sources, inputs, capacity, and selection budgets.

Monitor	Terminal AURC ↓	Safety AURC ↓	AUROC T / S ↑
Direct residual (execution)	0.247±0.000	0.282±0.000	0.610 / 0.592
Policy logistic	0.158±0.003	0.202±0.004	0.790 / 0.744
Execution logistic	0.219±0.004	0.234±0.004	0.700 / 0.681
Single-step routes (joint)	0.141±0.004	0.183±0.005	0.760 / 0.777
Quadratic logistic (joint)	0.128±0.006	0.155±0.004	0.842 / 0.812
Current MLP (joint)	0.119±0.005	0.146±0.004	0.858 / 0.827
Temporal MLP (joint, matched)	0.102±0.004	0.129±0.004	0.886 / 0.859
No coordinate heads (joint)	0.081±0.003	0.107±0.003	0.913 / 0.883
VISTA (joint)	**0.061±0.002**	**0.083±0.002**	**0.944 / 0.900**

Matched comparisons establish gains for both targets. The current MLP has widths $11 \to 32 \to 26 \to 1$ (1,269 parameters). The temporal MLP uses the same 44 prefix statistics with widths $44 \to 24 \to 8 \to 1$ (1,289 parameters). All use tanh and the same optimizer. The no-coordinate control sets $\lambda = 0$ in VISTA, keeping both routes, capacity, labels, and training steps. VISTA reduces AURC by 40.2% relative to the temporal MLP and by 24.7% relative to its matched control. Per-seed AURCs are $0.059/0.061/0.063$ versus $0.078/0.081/0.084$ and $0.098/0.102/0.106$. Holm-adjusted source-paired p values are 0.008 and < 0.001, respectively. Safety AURC falls from 0.129 (temporal) and 0.107 (no coordinates) to 0.083: reductions of 35.7% and 22.4%.

Established benchmark protocols. We evaluate LIBERO-10, LIBERO-PRO, and VLABench (Liu et al., 2023; Zhou et al., 2025; Zhang et al., 2025), including $\pi_0/\pi_{0.5}$ (Black et al., 2024; Physical Intelligence et al., 2025). LIBERO-10/VLABench use three 7-seen/3-unseen splits: 210/35/105/150 disjoint fitting/calibration/seen/unseen runs per policy. VISTA uses causal prefixes, training outcomes for fitting, and successful calibration runs for thresholds. LIBERO-10 averages $\alpha \in \{0.15, 0.20, 0.25\}$; Appendix A details adaptation and split results. LIBERO-PRO freezes the development monitor, fitting nominal scaling/thresholds on 50 reference and 50 disjoint calibration runs per suite. Four suites crossed with position/environment/object shifts yield 12 cells of 500 test runs.

Table 2: **LIBERO-10.** bAcc/wAcc/TWA; OV: OpenVLA; S/U: seen/unseen. Reference: metric-wise best in Park et al. (2026, Table 1).

Split	Reference	VISTA
OV/S	0.852/0.853/0.660	0.911/0.913/0.788
OV/U	0.834/0.828/0.663	0.896/0.897/0.754
π_0/S	0.885/0.926/0.715	0.939/0.954/0.827
π_0/U	0.892/0.921/0.705	0.923/0.946/0.790

Table 3: **LIBERO-PRO.** AUROC/bAcc: 12-cell means; time: normalized detection time (%, ↓), averaged over valid cells (Lee & Har, 2026, v2, Table I).

Method	AUROC	bAcc	Time
PCA-KMeans	0.652	0.599	67.32
RND	0.744	0.626	64.10
ACE	0.844	0.744	46.63
Stoch. sampling	0.814	0.734	45.60
PFD	0.866	0.781	50.89
VISTA	**0.996**	**0.929**	50.62

Gains across policies, shifts, and unseen tasks. VISTA improves all LIBERO-10 metrics. For π_0, seen/unseen bAcc is $0.939/0.923$ and TWA is $0.827/0.790$: task transfer affects both classification and alarm timing (Appendix A). VISTA leads LIBERO-PRO in AUROC/bAcc, while sampling detects earlier, and exceeds VLABench references on all six metrics. The controls below isolate the useful evidence.

Table 4: **VLABench with $\pi_{0.5}$.** Full comparison under seen and unseen task evaluation. Reference methods and metrics follow Park et al. (2026).

	Seen			Unseen		
Method	bAcc	wAcc	TWA	bAcc	wAcc	TWA
Mahalanobis	0.637	0.519	0.581	0.611	0.548	0.545
Cosine k-NN	0.535	0.384	0.485	0.704	0.676	0.584
PCA KMeans	0.642	0.669	0.529	0.616	0.618	0.520
RND	0.628	0.618	0.545	0.551	0.537	0.507
LogpZO	0.618	0.503	0.552	0.674	0.628	0.581
Cluster Entropy	0.629	0.656	0.602	0.526	0.506	0.490
EigenScore	0.553	0.487	0.532	0.566	0.538	0.554
STAC	0.615	0.586	0.556	0.591	0.598	0.576
ACE	0.596	0.545	0.553	0.591	0.595	0.552
SAFE-LSTM	0.629	0.524	0.587	0.484	0.533	0.482
SAFE-MLP	0.788	0.766	0.581	0.641	0.634	0.500
Hide-and-Seek	0.856	0.827	0.662	0.713	0.709	0.608
VISTA	**0.864**	**0.829**	**0.677**	**0.833**	**0.733**	**0.635**

Action/pose/joint channels use development adapters; missing channels are standardized zeros, and missing coordinates mask auxiliary losses. Frozen policies supply disagreement. Balanced accuracy is $(\mathrm{TPR} + \mathrm{TNR})/2$; weighted accuracy uses class frequencies. TWA combines correct non-alarms and failure-alarm timing (Appendix A); VLABench/LIBERO-PRO use their reference metric definitions.

Command/displacement scales are 0.05 m / 0.25 rad / 1; normalization uses development data.

5.1 Coordinate supervision and evidence separation

Table 5: **Coordinate information benefits both targets.** AURC on the same 840 trajectories and fitting budgets. Δ_T is excess terminal AURC over VISTA; wins count families favoring VISTA (of 12).

Variant	Terminal	Safety	Δ_T	Wins
No family supervision	0.070±0.003	0.095±0.003	+0.009	9/12
No subtype supervision	0.066±0.002	0.091±0.002	+0.005	8/12
No onset supervision	0.067±0.003	0.096±0.003	+0.006	9/12
Coordinate tuples permuted	0.080±0.003	0.106±0.003	+0.019	11/12
Mixed routes + auxiliary heads	0.076±0.003	0.101±0.003	+0.015	10/12
One policy sample	0.065±0.003	0.089±0.003	+0.004	8/12
Remove age and stall	0.064±0.002	0.088±0.002	+0.003	8/12
VISTA	0.061±0.002	0.083±0.002	–	–

Coordinate-label permutation shuffles joint family/subtype/onset tuples within policy and outcome strata, preserving their frequencies and outcome association but breaking their relation to history. Tuples and onset masks stay fixed across prefixes. Shuffled coordinates stay close to $\lambda = 0$: terminal/safety AURC is $0.080/0.106$ versus $0.081/0.107$, supporting the importance of history–coordinate correspondence. Mixing routes raises safety AURC from 0.083 to 0.101. Removing each auxiliary objective worsens both targets.

Consequence prediction responds to recovery. Among 40 certification recovery pairs, seven change terminal outcome despite the same initiating fault. After restoration, their median terminal risk falls by 0.22, while fault-family probability changes by 0.02; the 33 outcome-unchanged pairs change risk by 0.04. Shared pre-restoration prefixes give identical scores. A fault-family-only score yields AURC 0.193 versus 0.061 with outcome supervision. The representation retains disturbance evidence while the outcome head revises its judgment.

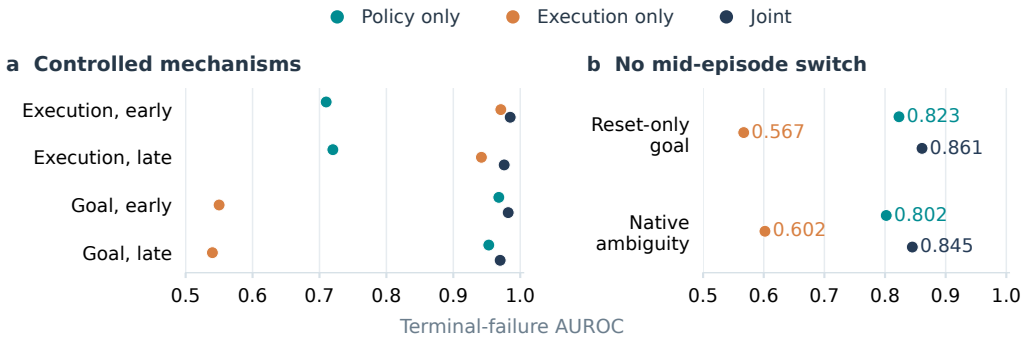


Figure 4: **Evidence ordering persists without mid-episode prompt switches.** Left: four core strata (120 sources each); policy/execution/joint AUROCs are $0.710/0.971/0.985$, $0.720/0.942/0.976$, $0.968/0.550/0.982$, $0.953/0.540/0.970$. Right: two independent 240-source goal challenges, with all runs retained.

5.2 The four coordinates change the useful evidence

Execution feedback dominates by $0.261/0.222$ AUROC for execution faults; policy evidence dominates by $0.418/0.413$ for goal faults (Holm $p = 0.003, 0.012, < 0.001, < 0.001$). Confidence/distance scores (Hendrycks & Gimpel, 2017; Lee et al., 2018) thus require consequence-specific assessment, complementing calibration/shift studies (Guo et al., 2017; Ovadia et al., 2019).

Goal evidence without a prompt discontinuity. Two 240-source cohorts use frozen monitors: *reset-only goals* change instructions before observation and keep them fixed; *native ambiguity* adds same-category distractors to F01–F06 with correct instructions and unmodified policies. They contain 78/69 terminal failures; all runs enter Figure 4. Removing action-change features and refitting retains policy AUROC $0.815/0.789$. A score-blind audit finds wrong initial referents in 52/69 native failures and 21 recovered runs. Policy evidence remains informative without mid-episode instruction changes.

A target changes ordering and selection. The terminal head's 0.944 terminal AUROC falls to 0.660 for safety; the safety head's 0.900 safety AUROC falls to 0.690 for completion (Table 6). At 70% coverage, each accepts 588/840 runs, but their sets differ on 160 (80 unique to each). The 193/840 disagreeing outcomes lead to different accepted executions under the same sensors and coverage.

Table 6: **Target and onset have operational consequences.** Top: primary population. Bottom: core terminal failures n_Y; pre: pre-onset alarms; timely: first alarms within 40 post-onset commands, excluding pre-alarms.

Trained target / assessed target	AURC	AUROC	Accepted at 70%
Terminal / terminal	0.061	0.944	588/840
Terminal / safety	0.184	0.660	588/840
Safety / safety	0.083	0.900	588/840
Safety / terminal	0.171	0.690	588/840
Domain	**n_Y (pre)**	**Timely / eligible**	**Median delay / 40**
Execution, early	18 (1)	15/17	0.10
Execution, late	22 (2)	16/20	0.16
Goal, early	20 (1)	16/19	0.13
Goal, late	24 (2)	17/22	0.18

Terminal-failure alarm timing. First alarms are assessed within 40 commands of onset. Of 84 terminal-failing core runs, six alarm before onset; 64/78 others alarm within budget (82.1%), and 14

later. Pre-onset alarms receive no fault-response credit; including them, 70/84 alarm by the deadline. Figure 5 uses the same alarm process. No rescue policy is executed.

5.3 Transfer across groups, mechanisms, and safety conditions

Groupwise ranking, then aggregation. Each task-family, policy, or subtype exclusion fits a separate model. Compute AURC within held-out group g and seed s, then average $\overline{A} = \frac{1}{3G}\sum_{s=1}^{3}\sum_{g=1}^{G} A_{gs}$. Within-group ranking makes the macro means in Table 7 invariant to group-specific strictly increasing score transformations. Excluded subtypes leave all development variants, including stress/recovery. Safety uses the same exclusions and fitting budget.

Table 7: **Transfer by held-out group and independent challenge.** Entries are terminal / safety AURC, averaged over three seeds; n counts independent sources. Family/policy rows use seven trajectories per source; subtype rows use its three execution variants. Every challenge uses 240 runs; actuator runs share 120 sources.

Held-out group	n	VISTA	No coordinates	Temporal MLP
F01: Bowl placement	7	0.061/0.085	0.076/0.108	0.099/0.128
F02: Cup placement	15	0.070/0.097	0.088/0.123	0.111/0.144
F03: Block placement	7	0.066/0.089	0.081/0.111	0.104/0.132
F04: Left relation	9	0.074/0.102	0.090/0.126	0.116/0.151
F05: Right relation	17	0.065/0.089	0.083/0.114	0.105/0.135
F06: Behind relation	6	0.069/0.094	0.087/0.121	0.110/0.141
F07: Drawer open	10	0.072/0.099	0.089/0.123	0.116/0.149
F08: Drawer close	13	0.064/0.086	0.082/0.112	0.102/0.130
F09: Cabinet open	5	0.071/0.097	0.088/0.123	0.112/0.146
F10: Stacking	7	0.067/0.090	0.085/0.118	0.109/0.139
F11: Object transfer	11	0.063/0.088	0.078/0.110	0.102/0.131
F12: Two-object order	13	0.074/0.100	0.093/0.127	0.122/0.154
Family macro mean	–	0.068/0.093	0.085/0.118	0.109/0.140
OpenVLA	51	0.073/0.099	0.085/0.120	0.111/0.143
SmolVLA	69	0.085/0.111	0.095/0.128	0.125/0.155
Policy macro mean	–	0.079/0.105	0.090/0.124	0.118/0.149
Axis swap	15	0.067/0.094	0.082/0.116	0.105/0.142
Command drop	20	0.071/0.098	0.084/0.121	0.109/0.145
Command delay	25	0.076/0.104	0.090/0.126	0.113/0.149
Command scale	33	0.080/0.106	0.095/0.130	0.119/0.153
Gain ramp	27	0.071/0.093	0.084/0.117	0.109/0.141
Subtype macro mean	–	0.073/0.099	0.087/0.122	0.111/0.146
Independent challenges: frozen model, new sources				
Unmodified workspace	240	0.079/0.106	0.096/0.132	0.113/0.159
Independent actuator	120	0.097/0.113	0.121/0.139	0.148/0.171
Goal changed at reset	240	0.091/0.103	0.114/0.127	0.139/0.154
Native referent ambiguity	240	0.085/0.108	0.107/0.133	0.132/0.161
Protected-object corridor	240	0.098/0.139	0.121/0.168	0.147/0.196

Independent tests of both consequences. Unmodified workspace runs retain 61 terminal failures; actuator gain $0.4/0.7$ or servo latency 2/6 steps, implemented below the command wrapper, yield 88 failures in 240 runs from 120 sources. Comparisons preserve complete source blocks; terminal p values are $0.014/0.006$ (Holm over these two tests); actuator AURC is 0.101 after removing age/stall. Reset-only goals and native ambiguity are defined in Figure 4. A fifth cohort places a protected object 2–6 cm from the nominal reach corridor, without changing instruction or controller. Its 240 runs contain 75 terminal and 104 safety failures. Safety failures in the preceding four cohorts are 90, 96, 83, and 91. None selects a model or threshold. VISTA improves both targets across all five challenges, including safety AURC 0.139 versus $0.168/0.196$ on the independent corridor test.

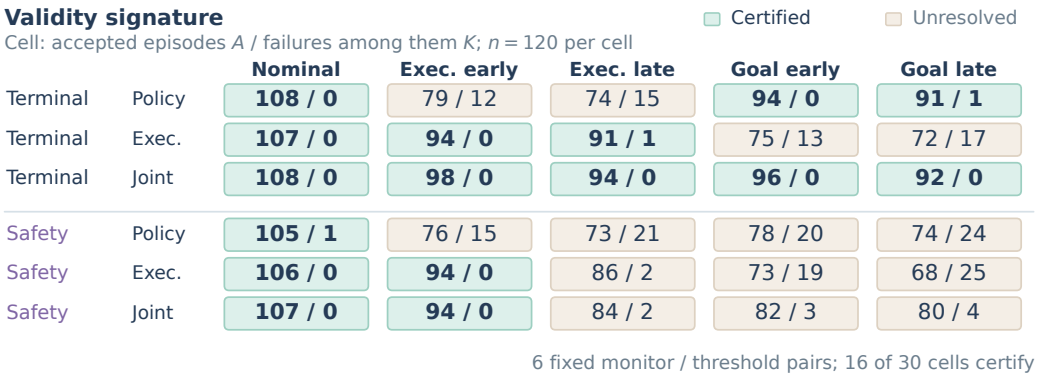


Figure 5: **Six monitor configurations, 30 domain cells.** Green certifies risk $\leq 10\%$/coverage $\geq 60\%$ at familywise error 0.05; beige is unresolved. Cells report accepted runs A/accepted failures K from 120 sources. Rows fix model/threshold; 16/30 cells certify.

5.4 Certifying the validity domain of a fixed monitor

Following selective prediction (Geifman & El-Yaniv, 2017; 2019; Traub et al., 2024), define the episode non-alarm set $B_q = \{\max_{t \leq 300} r_t \leq q\}$, coverage $C_d = \mathbb{P}_d(B_q)$, and risk $R_d = \mathbb{P}_d(Y^L = 1 \mid B_q)$. A signature certifies $R_d \leq .10$, $C_d \geq .60$ in 30 domain cells: six seed-0 configurations (three sensor sets, two targets) across five fault–onset conditions including nominal.

Calibration. Using one threshold across policies and families, sort M_i from n_0 outcome-negative nominal calibration runs; set $q = M_{(\lceil (n_0+1)(1-\beta) \rceil)}$, $\beta = 0.10$, or $q = +\infty$ if the rank exceeds n_0. Terminal/safety have $n_0 = 112/110$ and six finite thresholds. Exchangeability controls nominal negative alarm probability (Vovk et al., 2005; Angelopoulos & Bates, 2023); whole-episode calibration parallels functional prediction (Diquigiovanni et al., 2022).

A finite-sample decision. For each cell, let $n = 120$, $A = \sum_i \mathbf{1}(B_{q,i})$, $K = \sum_i \mathbf{1}(B_{q,i})Y_i$, and $\delta = 0.05/(2 \cdot 30)$. With $Q_u(a, b)$ the Beta(a, b) quantile, use one-sided exact binomial bounds (Clopper & Pearson, 1934):

$$L_C = Q_\delta(A, n - A + 1), \qquad U_R = Q_{1-\delta}(K + 1, A - K). \tag{5}$$

Set $L_C = 0$ when $A = 0$, and $U_R = 1$ when $A = 0$ or $K = A$. A cell certifies iff $L_C \geq .60$ and $U_R \leq .10$. For joint goal-late configurations, terminal $(A, K) = (92, 0)$ gives $(L_C, U_R) = (0.628, 0.074)$; safety $(80, 4)$ gives $(0.521, 0.175)$ and remains unresolved.

Proposition 1 (Simultaneous validity and mixture transfer)**.** *For 30 assessments fixed independently of certification sources, with i.i.d. source events within each cell, all certified cells meet the requirements with probability at least* 0.95*. A fixed row retains them under mixtures of certified strata with unchanged within-stratum laws.*

Proof. $A \sim \text{Bin}(n, C_d)$ and $K \mid A \sim \text{Bin}(A, R_d)$. A union bound over 60 δ-level bounds permits shared sources. Under mixture weights π_d, coverage is $\sum_d \pi_d C_d \geq .60$ and accepted-failure mass is $\sum_d \pi_d C_d R_d \leq .10 \sum_d \pi_d C_d$. Division gives the risk bound. □

These target-specific domains connect risk control (Bates et al., 2021; Angelopoulos et al., 2024), conditional-coverage limits (Barber et al., 2021), and coverage under shift (Barber et al., 2023) to the assessed population Q.

6 Conclusion

VLA monitor validity depends jointly on fault, sensors, target, and onset. VISTA makes these distinctions useful for learning: coordinate-supervised prefixes improve completion and safety ranking, including on goal failures without prompt switches. Yet joint sensing certifies five terminal strata and only two safety strata. A stronger predictor therefore still needs a measured validity domain.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research code implementation and the analysis of experimental results. We did not use generative AI tools to develop conceptual frameworks, formulate hypotheses, or design experiments. Additionally, we used generative AI tools to polish the manuscript for clarity, readability, and linguistic consistency. We manually inspected AI-assisted code for correctness, verified AI-assisted result analyses against experimental records and reported calculations, and reviewed all text revisions for accuracy and consistency with our intended scientific meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

The described study uses simulated robot tasks and explicitly defined contact and object-displacement predicates. The protocol does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

Sections 2–4 specify the monitoring variables, training objective, source allocation, outcome labels, and evaluation rules. Section 5.4 gives the certification procedure and proof. We intend to release code, task/scene definitions, source seeds, labels, adapters, training configurations, checkpoints, trajectories, and per-seed predictions.

REFERENCES

Christopher Agia, Rohan Sinha, Jingyun Yang, Ziang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking failure modes of generative policies: Runtime monitoring of consistency and progress. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 689–723, 2025.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia

Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023a.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023b. doi: 10.15607/RSS.2023.XIX.025.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022. doi: 10.1016/j.jmva.2021.104879.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

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

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022, 2023.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yousung Lee and Dongsoo Har. Perturbation-based epistemic uncertainty for failure detection in vision-language-action models. *arXiv preprint arXiv:2606.20754v2*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.090.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Seongheon Park, Wendi Li, Changdae Oh, Samuel Yeh, Zsolt Kira, Michael Hagenow, and Sharon Li. Hide-and-seek in trajectories: Discovering failure signals for VLA runtime monitoring. *arXiv preprint arXiv:2605.30834*, 2026.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

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

Harshitha Rajaprakash, Aditeya Prajapati, Rong Xue, Abrar Anwar, and Jesse Thomason. SAFE-CAST: Robust failure detection for VLA policies with contrast-set training and calibration. *arXiv preprint arXiv:2608.04246*, 2026.

Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P. Schoellig. Failure prediction at runtime for generative robot policies. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Florian Seligmann, Emiliyan Gospodinov, Enes Ulas Dincer, and Gerhard Neumann. VLA-FAIL: Efficient task failure detection for finetuned vision-language-action models. *arXiv preprint arXiv:2606.21386*, 2026.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

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

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0076.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736, 2023.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023a.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4129–4138, 2023b.

Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Andrei Ambrus, Haruki Nishimura, and Masha Itkina. Can we detect failures without failure data? uncertainty-aware runtime failure detection for imitation learning policies. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.073.

Borong Zhang, Jiahao Li, Jiachen Shen, Yuhao Zhang, Yishuai Cai, Lu Liu, Hailu Ji, Yuanpei Chen, Juntao Dai, Jiaming Ji, and Yaodong Yang. VLA-Arena: An open-source framework for benchmarking vision-language-action models. In *Proceedings of the International Conference on Machine Learning*, 2026.

Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. VLABench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11142–11152, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. LIBERO-PRO: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Kevin Lin, Abhiram Maddukuri, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293v3*, 2020.