Decomposing VAM Uncertainty: Training-Free and Supervised Probes for Robot Failure Detection

Liza Dahiya*, Naman Kumar*, Ziyin Xiong*, and Andrea Bajcsy
Robotics Institute, Carnegie Mellon University
Email: {ldahiya, namank, ziyinx, abajcsy}@andrew.cmu.edu

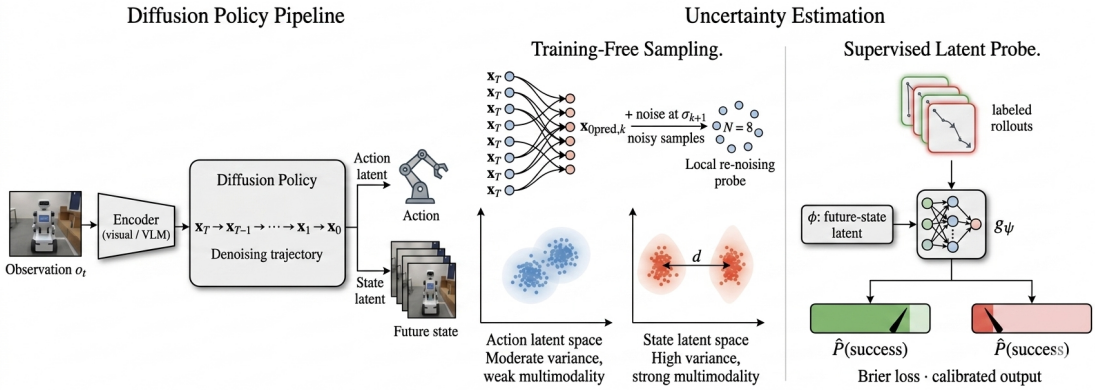

Fig. 1: **Denoising-time uncertainty in diffusion policies.** We probe a fixed VAM using N denoising trajectories to measure action (U_{act}) and future-state (U_{obs}) disagreement, and compare against a supervised latent success probe.

***Abstract*—Diffusion-based Video-Action Models (VAMs) jointly predict future actions and future states, yet are deployed without a confidence signal. We ask whether an actionable, diagnostic uncertainty signal can be extracted from a fixed VAM without retraining. We run N independent denoising trajectories and measure cross-sample disagreement separately in action (U_{act}) and future-state (U_{obs}) latents, then probe the score landscape to expose the mechanism behind convergence or divergence. On LIBERO-Spatial with the yellow mug distractor (-58 pp SR collapse), failure rates increase monotonically from 26.5% to 87.8% across four $(U_{\text{act}}, U_{\text{obs}})$ quadrants—driven by spatial proximity between distractor and target, not distractor presence alone. The score-variance probe identifies the catastrophic condition as the highest action-dominated sharpening ($\Delta\text{Var}_a = +0.90$), a multimodal signature that scalar metrics cannot decompose. A supervised latent probe achieves AUROC 0.99 in-distribution but degrades on this same condition, confirming the two approaches are complementary: one offers a calibrated scalar, the other a label-free diagnostic decomposition.**

I. Introduction & Motivation

Robots are increasingly controlled by generative policies trained on large amounts of expert observation-action data. Specifically, Video-Action Models (VAMs) have emerged as a promising architecture that jointly predicts future actions and outcomes: given multi-view observations, proprioception, and language instructions, a diffusion-based VAM generates a future action trajectory and corresponding future-observation sequence through a single denoising process [5, 6, 13]. However, VAMs act confidently even when their predictions are unreliable: they can hallucinate an implausible future, generate actions that lead to an unstable grasp, or confuse a distractor with the target. Moreover, when they generate erroneous action or outcome predictions, they fail silently: stakeholders or end-users receive no warning, because there is no confidence score, abstention mechanism, or signal to ask for help.

Uncertainty quantification is a natural solution to this problem, but it is difficult to instantiate in VAMs for three reasons. First, standard world-model UQ [9] attaches uncertainty to a fixed-dimensional Markov state and compounds it through one-step transitions; a VAM has neither, since it conditions on a history of frames, proprioception, and language and emits an entire future chunk in a single pass. There is no state vector to calibrate against and no recursion to propagate risk through. Second, VAMs predict coupled but distinct modalities [6]: prediction failures may originate in the imagined observations (e.g., hallucinated objects in the scene), the action plan (e.g., an erroneous action sample), or their correspondence (e.g., the imaged outcomes do not correlate with the predicted actions). A single scalar measure of uncertainty makes it hard to diagnose where the failure originated. Third, many VAMs are large or closed-weight, making retraining, ensembling, or supervised failure predictors impractical. Existing solutions sit at two extremes: calibrated world-model heads [9] that require labels and retraining, and training-free diffusion-uncertainty methods [3] that measure a single scalar over action-only policies and cannot localize the source of risk.

In this work, our key idea is to treat the VAM's denoising process as the uncertainty probe: a diffusion VAM already carries the uncertainty signal inside its own sampler: each

* Equal contribution.
† Corresponding Author.

denoising seed is an independent draw from the model's predictive distribution over futures and actions, so N seeds form a free ensemble that requires no retraining, labels, or extra networks. If the context admits one plausible behavior, the learned score field pulls every seed into the same basin; if it is ambiguous, seeds fall into different basins and disagreement persists to the end of the schedule. Furthermore, because action and future-observation latents are denoised jointly, our uncertainty measure disentangles the source of the error per modality.

We make the following contributions:

- **A training-free, modality-decomposed uncertainty probe for joint VAMs.** We measure cross-sample disagreement separately in action (U_{act}) and future-state (U_{obs}) latents across N denoising trajectories, extending training-free diffusion uncertainty to joint state-action models with no retraining, labels, or ensembles.
- **A diagnostic decomposition that localizes risk.** Thresholding (U_{act}, U_{obs}) yields four regimes with monotonically rising failure rates, separating behavioral ambiguity from unstable imagined futures.
- **A score-variance probe that exposes the mechanism.** Perturbing latents and re-querying the score network reveals the catastrophic condition as an action-dominated sharpening of the score landscape, driven by spatial proximity between distractor and target rather than distractor presence alone.
- **A complementarity result.** A supervised latent probe is highly accurate in-distribution but degrades on exactly the condition where our decomposition flags action-dominated mode-splitting, showing the two signals are complementary rather than competing.

We study Cosmos Policy [5] on LIBERO [8] and perturbed LIBERO-Pro tasks [14], whose perturbations keep the scene familiar while small changes determine whether the planned behavior is safe. On perturbed LIBERO-Spatial, our probes turn a success collapse into an interpretable signal, suggesting the denoising process itself is a practical source of diagnostic uncertainty for safer deployment of generative robot policies.

II. Related Work

A. Video-Action Models for Robot Control

Recent robot learning systems increasingly use generative video models as policy backbones. One line treats video generation as visual planning: a model predicts future frames, and a separate inverse-dynamics model converts the imagined trajectory into controls [2, 4, 7], but video uncertainty may not align with uncertainty in the executed action. A second line couples video and action prediction in one model [5, 6, 13], denoising future states and actions in a shared latent space. These joint VAMs are central to our work because they expose two linked channels—what the model believes will happen and what it plans to execute—so the model may be uncertain about scene dynamics, actions, or their consistency.

B. Uncertainty Quantification from Diffusion Dynamics

Prior UQ for robot foundation models either trains auxiliary probes on learned representations [9], requiring extra data, labels, and training, or measures disagreement across ensembles and sampled rollouts [10, 11, 12], which is costly at billion-parameter scale and often assumes Markovian latent world models ill-suited to history-conditioned VAMs. Diffusion models instead expose uncertainty for free: disagreement across denoising trajectories or local score vector fields estimates uncertainty without extra training [3]. However, existing training-free methods study action-only policies, measuring a single scalar over the action distribution. For VAM control the relation between future states and actions is critical—a policy may imagine an uncertain future but choose a stable action, or predict a stable future while split between action plans—so we extend diffusion-based probing to joint state-action VAMs by measuring variance and multimodality separately for action and future-state latents.

III. Problem Formulation

VAM-Based Robot Control. Let the state be $s_t = (\mathbf{I}_t, \mathbf{q}_t)$ where $\mathbf{I}_t$ is RGB image, $\mathbf{q}_t \in \mathcal{Q}$ is robot proprioception, and language instruction $l \in \mathcal{L}$. We use Cosmos Policy [5], a VAM based on the Cosmos-Predict2 2B latent DiT [1]. Given conditioning $c = (s_t, l)$, a VAM jointly denoises the latent:

$$\hat{x}_0 \;=\; \big[\, \hat{x}_0^{(a)},\; \hat{x}_0^{(s')},\; \hat{x}_0^{(V)} \big], \tag{1}$$

where $\hat{x}_0^{(a)}$ is an action chunk, $\hat{x}_0^{(s')}$ is the predicted future state, and $\hat{x}_0^{(V)}$ is a value estimate. The modalities occupy distinct latent frames in one diffusion sequence. Starting from the initial noise $\hat{x}_{\sigma_1} \sim \mathcal{N}(0, \sigma_1^2 I)$, each denoising step produces the next latent, $\hat{x}_{\sigma_{i+1}} = \textsc{Step}\big(\hat{x}_{\sigma_i},\; \sigma_i,\; \sigma_{i+1},\; c;\; D_\theta\big)$, where STEP is the ODE solver and D_θ is the denoiser. Generation uses T denoising steps and our implementation uses $\sigma_1 = 80$ and $\sigma_T = 4$. Different noise seeds can yield different decoded outputs $(\hat{\mathbf{a}},\; \hat{s}',\; \hat{V})$ for the same c.

Goal & Experimental Setup. We seek uncertainty measure that is both *actionable* (correlates with closed-loop failure) and *diagnostic* (indicates whether risk comes from action, future-state, or joint ambiguity). Evaluation uses LIBERO [8] under original, unperturbed conditions, and under LIBERO-Pro [15]'s object and position perturbations. LIBERO-Pro generates each perturbed scene per task, so the specific distractor varies by construction. Table I shows that most suites remain robust, but LIBERO-Spatial with a yellow mug drops from 99% to 41% SR, motivating our detailed analysis.

IV. Proposed Approach

We compare two uncertainty quantification (UQ) paradigms: a training-free sampling approach and a supervised latent probe. The training-free method runs N denoising trajectories from the same observation, measures disagreement in action and future-state latents, and probes score-landscape structure. The supervised C3-style baseline [9] trains a lightweight success classifier on Cosmos Policy latents.

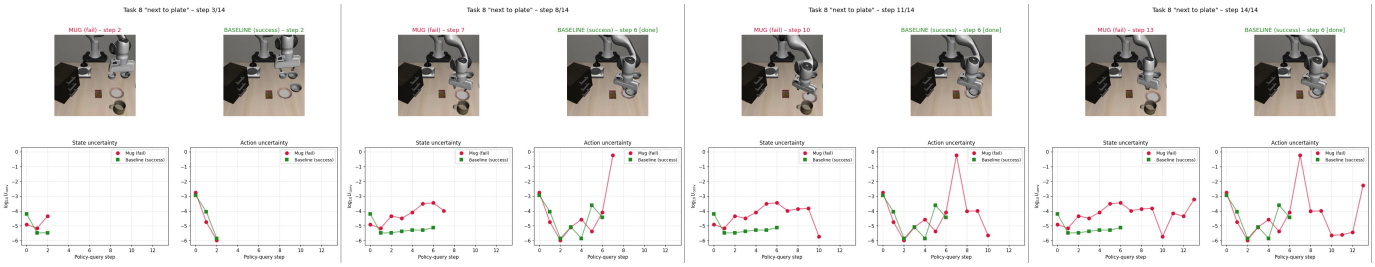

Fig. 2: **Task 8 with and w/o yellow mug distractor.** The mug condition fails while action uncertainty rises; the baseline succeeds as both modalities converge.

TABLE I: Success rate on LIBERO benchmarks with in-scene distractors. Each condition uses 100 rollout episodes (10 tasks × 10 episodes); ΔSR is relative to the suite baseline.

Suite	Condition	SR (%)	ΔSR	Min SR
LIBERO-10	Baseline	99.0	—	90
	+ mug	98.0	−1	80
	+ red stick	96.0	−3	70
LIBERO-Goal	Baseline	99.0	—	90
	+ moved stick	99.0	−0	90
	+ mug	99.0	−0	90
	+ rotated stick	99.0	−0	90
	+ yellow book	98.0	−1	90
	+ red stick	96.0	−3	80
	+ milk	90.0	−9	50
LIBERO-Object	Baseline	100.0	—	100
	+ moved stick	100.0	±0	100
	+ mug	100.0	±0	100
	+ red stick	100.0	±0	100
	+ yellow book	100.0	±0	100
LIBERO-Spatial	Baseline	99.0	—	90
	+ milk	94.0	−5	60
	+ moved stick	91.0	−8	50
	+ yellow book	91.0	−8	10
	+ red stick	87.0	−12	20
	+ mug	41.0	−58	0

A. Training-Free UQ via Sampling

N independent denoising trajectories. Given observation o_t, we run the full T-step denoising schedule from N independent Gaussian seeds while holding conditioning fixed, yielding trajectories $\{\mathbf{z}_\tau^{(n)}\}_{n=1}^N$, $\tau = T, \ldots, 1, 0$. We use $N = 8$ samples and $T = 5$ steps.

Convergence ratio U_{conv}. At denoising step τ, we measure cross-sample disagreement for modality m by mean pairwise cosine distance:

$$\gamma_\tau^{(m)} = \frac{1}{\binom{N}{2}} \sum_{i<j} \left(1 - \frac{\mathbf{f}_{\tau,m}^{(i)} \cdot \mathbf{f}_{\tau,m}^{(j)}}{\|\mathbf{f}_{\tau,m}^{(i)}\| \, \|\mathbf{f}_{\tau,m}^{(j)}\|}\right), \tag{2}$$

where $\mathbf{f}_{\tau,m}^{(n)}$ is the spatially averaged feature vector. The convergence ratio is:

$$U_{\text{conv}}^{(m)} = \frac{\gamma_0^{(m)}}{\gamma_T^{(m)}}, \tag{3}$$

with values near 0 indicating convergence and values near 1 indicating persistent uncertainty.

Modality decomposition. We compute U_{conv} separately for two modalities:

$$U_{\text{act}} = U_{\text{conv}}^{(\text{action})}, \tag{4}$$

$$U_{\text{obs}} = U_{\text{conv}}^{(\text{future\_state})}. \tag{5}$$

U_{act} captures whether the model commits to a consistent action; U_{obs} captures whether it forms a consistent internal prediction of the future scene. We define four regimes by thresholding $(U_{\text{act}}, U_{\text{obs}})$ at per-condition medians and also report

$$U_{\text{AV}} = \frac{U_{\text{act}}}{U_{\text{obs}}}, \tag{6}$$

which encodes action-dominated (> 1) versus state-dominated (< 1) uncertainty.

a) Score-variance probe.: To characterize the mechanism behind convergence, we perturb each latent by $\boldsymbol{\epsilon}_n \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, re-query the score network, and compute:

$$\text{Var}_m(\tau) = \frac{1}{N} \sum_{n=1}^{N} \left\| \mathbf{s}_\theta(\mathbf{z}_\tau^{(n)} + \boldsymbol{\epsilon}_n, \tau) - \mathbf{s}_\theta(\mathbf{z}_\tau^{(n)}, \tau) \right\|_m^2, \tag{7}$$

where $\|\cdot\|_m$ restricts the norm to modality m. We report ΔVar_m relative to baseline; positive ΔVar_a indicates action-space sharpening, while negative values indicate flattening. We also report mode-center distance $d = \|\mu_1 - \mu_2\|_2$ and normalized separation $\text{Sep} = d/\sqrt{\text{within-var}}$ to distinguish genuinely separated modes from overlapping clusters.

B. Supervised UQ via a Latent Probe

The supervised baseline tests whether VAM latents encode success information when labels are available. Specifically, we adopt a C^3-style [9] latent probe approach where probes are trained on either future-state latents or future-state+action latents, appending normalized timestep. Each example (ϕ_i, y_i) uses a trajectory-level success label, and the probe predicts:

$$\hat{p}_i = g_\psi(\phi_i) \approx P(y_i = 1 \mid \phi_i), \tag{8}$$

where g_ψ is trained with Brier score loss,

$$\mathcal{L}_{\text{Brier}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2, \tag{9}$$

which measures accuracy and calibration. Here ϕ_i is the (future-state/future-state+action) latent for example i, $y_i \in \{0, 1\}$ is its trajectory-level success label, $\hat{p}_i$ is the predicted success probability from probe g_ψ with parameters ψ, and N is the number of training examples. We train on balanced single-step LIBERO rollouts, split 90/5/5, select by validation Brier

score, and evaluate Brier, ECE, and AUROC in-distribution and zero-shot on LIBERO-Pro.

V. Results

We focus on LIBERO-Spatial with the yellow mug distractor, the largest performance collapse. The training-free signal correlates with failure and decomposes its source; the supervised probe provides a calibrated scalar baseline.

A. Training-Free UQ via Sampling

Qualitative Results. Figure 2 shows Task 8. In the baseline, state and action uncertainty decrease as the gripper approaches the bowl. With the yellow mug present, U_{obs} remains low but U_{act} rises sharply, indicating a multimodal action failure: the model sees the scene coherently but oscillates between mug and bowl.

Quadrant Analysis. Table II bins 130 LIBERO-Spatial+mug episodes by median U_{act} and U_{obs}. Failure increases from 26.5% in low–low to 87.8% in high–high, a 3.3× spread. The imbalanced quadrants separate failure types: high-action/low-state uncertainty indicates multimodal actions (81.2% failure), while low-action/high-state uncertainty indicates unstable imagined futures (50.0% failure) that action-only metrics would miss.

Result: Spatial proximity drives uncertainty. Task-level results show that distractor presence alone is insufficient. Tasks 7 and 9 reach 100% SR because the bowl is elevated and separated from the mug, so episodes stay low–low. Tasks 2, 3, and 8 fail at 0% SR because the bowl is adjacent to the mug and episodes concentrate in high-U_{act} quadrants.

B. Supervised UQ via Latent Probe

Result: In-distribution performance. Table IV shows strong held-out LIBERO performance, with future-state-only features best (AUROC 0.9909, Brier 0.0297, ECE 0.0193), suggesting future-state latents align more cleanly with semantic task completion than action latents.

Result: Out-of-distribution generalization. Table V shows good zero-shot transfer on several perturbations, but degradation on `libero_spatial_with_mug` (AUROC 0.8574, Brier 0.2016)—the same condition where the training-free probe exposes action-dominated multimodality. "–" marks undefined AUROC when all rollouts share one label.

TABLE II: Per-task quadrant breakdown (LIBERO-Spatial + mug); cells show fails/n. Elevated-bowl tasks cluster in low–low and succeed; adjacent-bowl tasks concentrate in high-U_{act} quadrants and fail on every episode.

Task	High U_{act} High U_{obs}	High U_{act} Low U_{obs}	Low U_{act} High U_{obs}	Low U_{act} Low U_{obs}	SR
T0	1/5	—	1/8	—	85%
T1	5/6	5/5	1/1	0/1	15%
T2	12/12	—	1/1	—	0%
T3	9/9	1/1	2/2	1/1	0%
T4	5/5	2/3	1/1	3/4	15%
T5	—	0/2	0/1	2/10	85%
T6	—	5/5	—	7/8	8%
T7	—	—	—	0/13	**100%**
T8	11/11	—	2/2	—	0%
T9	0/1	—	—	0/12	**100%**
Total	43/49	13/16	8/16	13/49	40.8%
Fail rate	**88%**	**81%**	50%	27%	

TABLE III: Selected changes in uncertainty metrics at $\sigma \approx 4$. $\Delta\text{Var}_a > 0$ indicates the action score landscape sharpens; $\Delta U_{\text{AV}} > 0$ indicates action-dominated uncertainty.

Suite	Distractor	ΔVar_a ($\times 10^{-4}$)	ΔVar_s ($\times 10^{-4}$)	ΔU_{AV}	Δd	ΔSep
libero_spatial	red_stick	+0.27	−0.56	+0.16	+0.19	+12
	mug	+0.90	−0.45	+0.27	+0.33	+40
	milk	+0.40	+1.12	+0.04	+0.47	+46
	diffpos_stick	+2.76	+1.60	+0.18	+0.92	+38
	yellow_book	+1.15	+0.20	+0.45	+0.32	+9
libero_10	red_stick	−0.07	−1.52	+0.00	−0.17	−22
	milk	−0.05	−0.05	−0.05	−0.14	−20

TABLE IV: Held-out LIBERO performance of the supervised latent success probe trained on 1-step Cosmos Policy rollouts.

Input features	Input dim.	Brier ↓	ECE ↓	AUROC ↑
Future-state latent only	2049	0.0297	0.0193	0.9909
Future-state + action latent	4097	0.0405	0.0358	0.9821

TABLE V: Zero-shot LIBERO-Pro evaluation of the supervised latent success probe.

Condition	Future-state only			Future-state + action		
	Brier ↓	ECE ↓	AUROC ↑	Brier ↓	ECE ↓	AUROC ↑
libero_10_with_milk	0.0452	0.0418	0.8755	0.1558	0.1642	0.8331
libero_10_with_mug	0.0452	0.0418	0.8755	0.0000	0.0027	–
libero_10_with_red_stick	0.1885	0.1930	0.8989	0.1946	0.1891	0.7139
libero_goal_with_diffpos_stick	0.0274	0.0324	0.9822	0.0298	0.0315	0.9603
libero_goal_with_milk	0.0187	0.0236	0.9978	0.0215	0.0254	0.9731
libero_goal_with_mug	0.0939	0.1039	0.9363	0.0621	0.0509	0.8634
libero_goal_with_red_stick	0.0000	0.0012	–	0.0001	0.0043	–
libero_goal_with_rotated_stick	0.0364	0.0386	0.9419	0.0361	0.0493	0.9290
libero_goal_with_yellow_book	0.0001	0.0019	–	0.0001	0.0051	–
libero_obj_diffpos_stick	0.0007	0.0051	–	0.0000	0.0030	–
libero_obj_with_mug	0.0007	0.0040	–	0.0000	0.0025	–
libero_obj_with_red_stick	0.0000	0.0008	–	0.0000	0.0020	–
libero_spat_diffpos_stick	0.2653	0.2983	–	0.1312	0.1741	–
libero_spatial_with_milk	0.1428	0.1490	0.9631	0.0937	0.1029	0.9639
libero_spatial_with_mug	0.2016	0.1961	0.8574	0.2055	0.2064	0.8234
libero_spatial_with_red_stick	0.1281	0.1339	0.9245	0.1363	0.1322	0.8540
libero_spat_yellow_book	0.0100	0.0166	–	0.0094	0.0351	–

Result: Complementarity of supervised and training-free probes. The supervised probe is least reliable in the yellow-mug condition—exactly where the training-free probe detects action-dominated mode splitting—so the calibrated scalar and the label-free decomposition should be combined rather than substituted.

VI. Conclusion

Denoising-time uncertainty in a diffusion-based VAM is structured and modality-dependent. Decomposing cross-sample disagreement into U_{act} and U_{obs} reveals four failure regimes, driven by spatial proximity between distractor and target rather than distractor presence alone, while the score-variance probe identifies the catastrophic condition via action-dominated sharpening, a multimodal signature scalar metrics cannot decompose. A supervised probe degrades on the same hard condition, confirming complementarity and suggesting the denoising process itself is a practical source of diagnostic uncertainty for safer policy deployment.

References

[1] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

[2] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv e-prints*, 2023.

[3] Zhanpeng He, Yifeng Cao, and Matei Ciocarlie. Uncertainty comes for free: Human-in-the-loop policies with diffusion models, 2025. URL https://arxiv.org/abs/2503.01876.

[4] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv e-prints*, 2024.

[5] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

[6] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[7] Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv e-prints*, 2025.

[8] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[9] Zhiting Mei, Tenny Yin, Micah Baker, Ola Shorinwa, and Anirudha Majumdar. World models that know when they don't know: Controllable video generation with calibrated uncertainty. *arXiv preprint arXiv:2512.05927*, 2025.

[10] Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv e-prints*, 2023.

[11] Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. *arXiv preprint arXiv:2505.00779*, 2025.

[12] Jiabing Yang, Yixiang Chen, Yuan Xu, Peiyan Li, Xiangnan Wu, Zichen Wen, Bowen Fang, Tao Yu, Zhengbo Zhang, Yingda Li, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. Uaor: Uncertainty-aware observation reinjection for vision-language-action models. *arXiv e-prints*, 2026.

[13] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. *arXiv e-prints*, 2026.

[14] 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.

[15] 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, 2026. URL https://arxiv.org/abs/2510.03827.