Perturbation-Based Uncertainty for Failure Detection in Vision-Language-Action Models

Yousung Lee Dongsoo Har
Korea Advanced Institute of Science and Technology (KAIST)

***Abstract*—Vision-Language-Action (VLA) models have shown strong performance in robotic manipulation, but reliable uncertainty quantification remains challenging, particularly under distribution shift. Unlike autoregressive policies, many modern VLA models generate continuous actions through regression or flow-based generation, where explicit predictive probabilities are unavailable. Moreover, existing approaches often rely on stochastic action sampling or supervised failure labels, limiting their applicability across diverse pretrained VLA models. In this work, we propose a label-free and model-agnostic framework for inference-time uncertainty estimation through hidden activation perturbations, motivated by Bayesian perspectives on local model variations. Specifically, we inject Gaussian perturbations into transformer hidden activations and estimate epistemic signals from disagreement across perturbed action predictions. Experiments on LIBERO and LIBERO-PRO show that perturbation-based uncertainty consistently improves failure detection under distribution shift compared to sampling-based uncertainty, providing a practical uncertainty signal for VLA models.**

I. Introduction

Vision-Language-Action (VLA) models have recently demonstrated strong generalization capabilities across robotic manipulation tasks by mapping visual observations and language instructions to continuous control actions [13, 14, 2]. However, reliable deployment of VLA systems remains challenging, since even small action errors can lead to catastrophic physical failures in real-world environments. Recent studies have shown that VLA models can exhibit overconfident failures under distribution shift, sometimes ignoring language instructions and relying on memorized or shortcut-like manipulation patterns [25, 23]. Such failures are particularly problematic because the model may continue executing incorrect actions despite violating the language instruction, while exhibiting low uncertainty under stochastic action sampling.

Reliable uncertainty estimation is therefore critical for safe deployment and failure detection in VLA systems. In Bayesian learning, predictive uncertainty is commonly decomposed into aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty captures inherent stochasticity in observations or action generation and is generally irreducible, whereas epistemic uncertainty arises from limited model knowledge and can in principle be reduced with additional data or improved coverage of the environment [7, 8, 18, 22]. In safety-critical robotic systems, epistemic uncertainty is particularly important under distribution shift, where policies may encounter unfamiliar or ambiguous situations.

However, uncertainty estimation in VLA models is fundamentally challenging. Unlike autoregressive token-based VLA models [3, 13], many recent VLA models generate actions through regression or flow-based generation [14, 2], where explicit predictive probabilities are unavailable. As a result, uncertainty estimation methods developed for autoregressive language models are not directly applicable to VLA models.

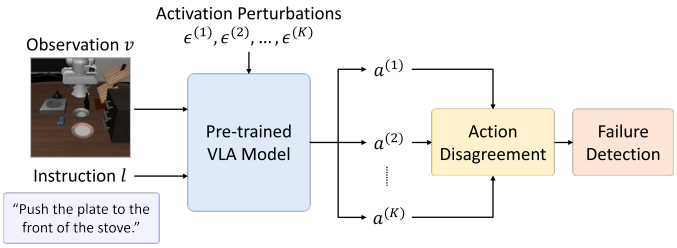


Fig. 1: Overview of perturbation-based uncertainty estimation for VLA failure detection. Hidden activation perturbations induce multiple action predictions, and disagreement across these predictions is used for failure detection.

Recent approaches often estimate uncertainty through stochastic sampling under a fixed pretrained model [21]. In generative models such as $\pi_{0.5}$ [2], multiple trajectories can be generated from the same observation to estimate uncertainty. However, such approaches primarily capture stochastic variability within a single model, which is more closely related to aleatoric uncertainty and may not fully capture epistemic uncertainty arising from insufficient model knowledge [11, 19]. Consequently, epistemic uncertainty estimation in pretrained VLA systems remains relatively underexplored. Furthermore, existing VLA failure detection methods often rely on supervised failure labels or trajectory-level supervision, architecture-specific stochastic generation mechanisms, or access to output token probabilities and perplexity estimates [9, 6, 21, 10], limiting their practicality and applicability across diverse VLA models and deployment settings.

To address these limitations, we propose a label-free and model-agnostic approach for inference-time epistemic uncertainty estimation in pretrained VLA models through hidden activation perturbations. Motivated by Bayesian model uncertainty, we interpret hidden activation perturbations as tractable approximations of local parameter variations around pretrained weights. As shown in Fig. 1, we inject Gaussian noise into hidden activations of the vision or vision-language backbone, depending on the VLA architecture, and quantify epistemic uncertainty as action disagreement across perturbations. Experiments on LIBERO [17] and LIBERO-PRO [25] demonstrate that the proposed perturbation-based epistemic

uncertainty effectively detects failures under distribution shift and consistently outperforms sampling-based uncertainty for VLA failure detection.

II. Related Work

A. Epistemic Uncertainty

Bayesian approaches commonly decompose predictive uncertainty into aleatoric and epistemic components, where epistemic uncertainty reflects uncertainty over model parameters under limited data or knowledge [11, 7, 19]. Practical approximations of epistemic uncertainty include Monte Carlo dropout, deep ensembles, and perturbation-based uncertainty estimation methods [7, 15, 20, 22, 18, 19]. Among these, deep ensembles provide strong approximations of posterior predictive uncertainty but are computationally prohibitive for large pretrained foundation models [20], while Monte Carlo dropout is not directly applicable to many pretrained VLA models that were not trained with dropout, as its approximate Bayesian interpretation relies on dropout-enabled training [7]. These limitations motivate post-hoc perturbation methods, which provide tractable approximations of local model uncertainty without retraining or maintaining multiple pretrained models [18, 22, 19].

B. Uncertainty Quantification in LLM and VLM

Prior work has explored uncertainty quantification through output variability. In large language models, one line of work estimates predictive uncertainty by sampling multiple outputs for a fixed input and measuring their semantic dispersion [16]. More recently, perturbation-based approaches have been studied across both language and vision-language models, where semantic-preserving input variations (e.g., prompt rephrasing or visual perturbations) are used to reveal instability in model predictions [24, 8, 12]. In addition, recent studies have shown that uncertainty signals can be used to detect unreliable generations, such as hallucinations, via semantic entropy or representation-level dispersion [5, 4]. However, these approaches have primarily focused on discrete text generation, and their applicability to VLA models with continuous action outputs remains largely underexplored.

C. Failure Detection in Vision-Language-Action Models

In VLA models, reliable execution requires detecting failure-relevant signals during policy execution. Prior approaches such as SAFE [9] and recent VLA calibration methods [6] rely on labeled success and failure trajectories. However, such labeled failure trajectories are difficult and often unsafe to collect, and are typically unavailable at deployment time. Recent work has explored failure detection through action uncertainty, temporal consistency, and representation-level signals. For example, Sentinel [1] monitors generative robot policies by combining temporal action consistency with VLM-based task progress assessment. FIPER [21] estimates uncertainty from stochastic action samples in generative policies. Other approaches leverage token-level entropy or perplexity from autoregressive action distributions [10]. In contrast, we propose label-free uncertainty estimation for VLAs through hidden activation perturbations in vision-language representations, without requiring failure labels, output token probabilities, or stochastic action sampling. To our knowledge, this is the first approach to use hidden activation perturbations as an uncertainty signal for failure detection in VLA models.

III. Methodology

A. Vision-Language-Action Policy

We consider a Vision-Language-Action policy π_θ parameterized by θ. Given observation $x_t = (v_t, \ell)$ consisting of visual observation v_t and language instruction ℓ, the policy predicts a chunk of future continuous actions:

$$\mathbf{a}_{t:t+H-1} = \pi_\theta(x_t), \tag{1}$$

where H denotes the action chunk horizon, d denotes the action dimension, and $\mathbf{a}_t \in \mathbb{R}^d$ denotes a continuous robot action at timestep t.

B. Uncertainty Decomposition

From a Bayesian perspective, model parameters $\theta \sim p(\theta|\mathcal{D})$ are treated as random variables conditioned on training data $\mathcal{D}$. Predictive uncertainty can then be decomposed into aleatoric uncertainty and epistemic uncertainty under the posterior distribution over model parameters [11, 7]:

$$\mathrm{Var}(\mathbf{a}|x_t, \mathcal{D}) = \underbrace{\mathbb{E}_\theta\big[\mathrm{Var}(\mathbf{a}|x_t, \theta)\big]}_{\text{Aleatoric}} + \underbrace{\mathrm{Var}_\theta\big[\mathbb{E}(\mathbf{a}|x_t, \theta)\big]}_{\text{Epistemic}}. \tag{2}$$

The epistemic term reflects uncertainty over plausible model parameters under the posterior distribution. However, exact posterior sampling over model parameters remains computationally intractable for large-scale pretrained foundation models [20, 18, 22, 19]. Therefore, we focus on inference-time approximations of local model variations without retraining the pretrained VLA model.

C. Perturbation-Based Epistemic Uncertainty

Following prior work on perturbation-based uncertainty estimation [18, 22], we approximate local parameter variations around the pretrained parameters θ^* using Gaussian noise perturbations:

$$\theta^{(k)} = \theta^* + \epsilon^{(k)}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, \sigma^2 I). \tag{3}$$

Direct parameter perturbation in large-scale pretrained foundation models is computationally expensive. Instead, we approximate local parameter variations through Gaussian noise injections applied to MLP hidden activations shared across transformer layers [18]. Consider a linear transformation $y = Wh + b$, where W, b, and h denote the weight matrix, bias vector, and hidden activation, respectively. Applying hidden activation perturbations yields $y' = W(h + \epsilon) + b = Wh + (b + W\epsilon)$, showing that perturbing hidden activations can induce local variations in the output representation. These perturbations induce perturbed action chunks:

$$\mathbf{a}^{(k)}_{t:t+H-1} = \pi_{\theta^{(k)}}(x_t). \tag{4}$$

TABLE I: Failure detection AUC for $\pi_{0.5}$ on LIBERO-PRO. Both sampling and perturbation-based uncertainty use $K = 5$ samples. Perturbation-based uncertainty uses hidden activation perturbations with $\sigma = 0.25$. Each benchmark suite is evaluated using 500 rollout episodes. Early AUC is computed using only the first 70% of the maximum episode length.

Environment	Success (%)	Sampling		Perturbation	
		Full AUC	Early AUC (70%)	Full AUC	Early AUC (70%)
LIBERO-PRO Spatial-Pos	45.0	0.935	0.795	**0.963**	**0.844**
LIBERO-PRO Object-Pos	20.4	0.482	0.464	**0.651**	**0.511**
LIBERO-PRO Goal-Pos	32.4	0.861	0.824	**0.934**	**0.900**
LIBERO-PRO 10-Pos	10.6	0.869	0.830	**0.909**	**0.855**
LIBERO-PRO Spatial-Env	39.4	0.816	0.750	**0.845**	**0.770**
LIBERO-PRO Object-Env	75.4	0.750	0.685	**0.788**	**0.725**
LIBERO-PRO Goal-Env	52.8	0.736	0.681	**0.865**	**0.828**
LIBERO-PRO 10-Env	69.1	0.878	0.803	**0.911**	**0.850**

We then quantify uncertainty as disagreement across perturbed action chunks:

$$u_t = \frac{1}{Hd} \sum_{h=0}^{H-1} \sum_{i=1}^{d} \mathrm{Var}_k \left[a_{t+h,i}^{(k)} \right], \tag{5}$$

where $a_{t+h,i}^{(k)}$ denotes the i-th action dimension of the predicted action at future step h under perturbation sample k. Variance is computed across perturbation samples and averaged over chunk steps and action dimensions.

For each perturbation sample, we reuse the same Gaussian noise tensor at every timestep throughout the rollout, rather than resampling it independently at each timestep. This allows each perturbation sample to induce a temporally consistent local model variation. Intuitively, high epistemic uncertainty can indicate that the perturbed model variants produce inconsistent action chunks under the current observation, suggesting insufficient model knowledge for the current situation. Conversely, low uncertainty suggests that the predicted action chunks remain stable under perturbations.

D. Temporal Aggregation for Failure Detection

To obtain a stable uncertainty estimate for the current timestep, we aggregate recent chunk-level uncertainty scores using a temporal sliding window [21]:

$$U_t = \frac{1}{T_w} \sum_{\tau=t-T_w+1}^{t} u_\tau, \tag{6}$$

where T_w denotes the sliding window size. This temporal aggregation captures persistent disagreement patterns rather than transient action fluctuations during policy execution.

IV. Experiments

A. Models and Perturbation Design

We evaluate two modern VLA policies: OpenVLA-OFT [14] and $\pi_{0.5}$ [2]. OpenVLA-OFT predicts continuous action chunks through a deterministic regression head, whereas $\pi_{0.5}$ uses a flow-matching-based generative policy that supports stochastic action sampling. We perturb hidden activations within the visual or vision-language backbone of each model. Specifically, we inject Gaussian noise into intermediate MLP activations of the dual vision encoder in OpenVLA-OFT and into the vision-language backbone of $\pi_{0.5}$ before the action expert module. In both models, the perturbations are applied before the final projection layer of transformer MLP blocks. This design captures uncertainty in the model's understanding of the current situation and its effect on action predictions, separately from action-generation stochasticity.

B. Environments

Experiments are conducted on LIBERO [17] and LIBERO-PRO [25]. LIBERO-PRO introduces controlled distribution shifts while preserving task semantics. We consider two shift categories: Position shifts, which modify object placements and relative spatial arrangements, and Environment shifts, which alter the visual scene context (e.g., replacing kitchen-table environments with living-room-table layouts) without changing the task itself. These distribution shifts frequently induce failures in pretrained VLA policies despite their strong performance on standard LIBERO, making LIBERO-PRO a challenging benchmark for evaluating uncertainty estimation and failure detection under distribution shift. We therefore consider evaluation under distribution shift essential for assessing failure detection, since many recent VLA models achieve near-perfect performance on standard LIBERO.

C. Evaluation Protocol

We evaluate perturbation-based uncertainty for trajectory-level failure detection using ground-truth success/failure labels only for evaluation purposes. The sampling baseline generates multiple action chunks from different initial noise samples, whereas our method fixes the initial sampling noise and perturbs only the hidden activations. We temporally aggregate per-timestep uncertainty scores for each rollout using the sliding-window formulation in Eq. (6) with window size $T_w = 5$. Early failure detection uses only the first 70% of the maximum episode length defined by the environment. We then compute a trajectory-level uncertainty score as the maximum temporally aggregated uncertainty:

$$s = \max_t U_t. \tag{7}$$

A rollout is classified as a failure when $s > \delta$, where δ denotes a decision threshold. We report the area under the

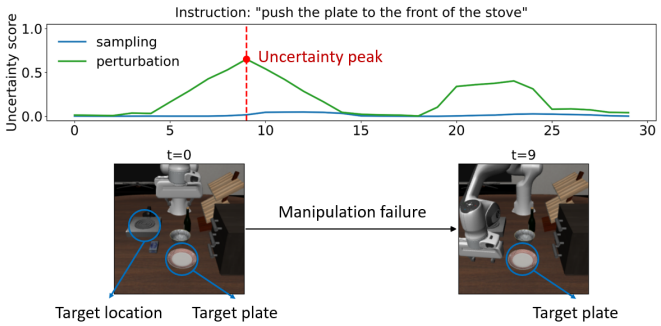
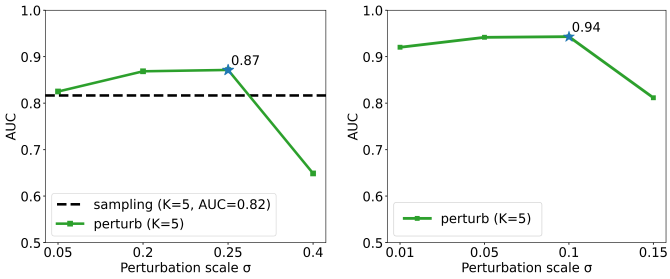


Fig. 2: Representative failure case from LIBERO-PRO Goal-Env. The red dashed line marks the peak perturbation-based uncertainty, corresponding to the right image. The curves show temporally windowed uncertainty scores, min–max normalized to $[0, 1]$ over all evaluation episodes for visualization.



(a) $\pi_{0.5}$ (b) OpenVLA-OFT

Fig. 3: Effect of perturbation strength σ on full-trajectory failure detection AUC under hidden activation perturbations on LIBERO-10 using perturbation samples $K = 5$.

receiver operating characteristic curve (AUC), a threshold-free metric that measures separability between successful and failed trajectories across all thresholds. Each benchmark suite includes 500 rollout episodes.

D. Main Results

We compare sampling-based and perturbation-based uncertainty for trajectory-level failure detection using $\pi_{0.5}$, which supports stochastic action sampling. Table I shows that perturbation-based uncertainty consistently improves failure detection performance under distribution shift across LIBERO-PRO benchmarks. The largest improvement is observed on LIBERO-PRO Object-Pos, where AUC increases from 0.482 to 0.651. Substantial gains are also observed on LIBERO-PRO Goal-Env (0.736 $\rightarrow$ 0.865) and Goal-Pos (0.861 $\rightarrow$ 0.934). These results suggest that perturbation-induced action disagreement captures failure-sensitive signals that are not fully reflected by stochastic sampling within a fixed pretrained policy. Notably, perturbation-based uncertainty also improves early failure detection performance, indicating that informative uncertainty signals emerge before task failure.

E. Qualitative Visualization

To better understand the performance gains observed in Table I, we visualize a representative failure case from LIBERO-PRO Goal-Env. As shown in Fig. 2, the policy repeatedly attempts to manipulate the target plate but fails to move it to the target location. Despite this failure, sampling-based uncertainty remains low throughout the rollout, indicating consistent action predictions across stochastic samples. In contrast, perturbation-based uncertainty increases during the failure-relevant manipulation phase, reflecting increased disagreement among the perturbed model variants.

F. Effect of Perturbation Parameters

We further analyze the effect of perturbation parameters on LIBERO-10 using $\pi_{0.5}$ and OpenVLA-OFT, where both VLA models achieve a task success rate of 94.0%. As shown in Fig. 3, both models achieve their best failure detection performance at moderate perturbation magnitudes, while excessively large perturbations degrade uncertainty quality. For $\pi_{0.5}$, performance peaks at $\sigma = 0.25$, whereas OpenVLA-OFT performs best at $\sigma = 0.1$. Interestingly, despite their different policy architectures, both models exhibit a similar trend where moderate perturbations achieve the best failure detection performance, suggesting that informative epistemic signals can be obtained from local variations that are sufficiently strong while remaining plausible around the pretrained policy.

TABLE II: Effect of the number of samples K on full-trajectory failure detection AUC for $\pi_{0.5}$ on LIBERO-10.

K	Sampling	Perturbation
2	0.831	0.869
3	0.808	0.875
4	0.823	0.876
5	0.817	0.871

Table II analyzes the effect of the number of samples on failure detection performance for $\pi_{0.5}$ on LIBERO-10. While sampling-based uncertainty does not consistently improve with larger sample counts, perturbation-based uncertainty remains effective even with a small number of perturbation samples. For $\pi_{0.5}$, using only $K = 2$ perturbations achieves an AUC of 0.869, which is comparable to the performance obtained with $K = 5$ (0.871). This suggests that reliable uncertainty estimates can be obtained with minimal additional computation.

V. Conclusion

In this work, we proposed a label-free, inference-time uncertainty estimation framework for pretrained VLA models through hidden activation perturbations. Motivated by Bayesian perspectives on local model variations, our approach captures epistemic uncertainty through disagreement across perturbed action predictions. Experiments on LIBERO and LIBERO-PRO showed that perturbation-based uncertainty consistently improves failure detection under distribution shift compared to sampling-based uncertainty. These results show that perturbation-based uncertainty can serve as a practical signal for VLA failure detection and runtime safety monitoring.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (Technology Development for Fire Response in Facilities and Components of Electric-based Mobility) (RS-2025-02613131, Development and on-site demonstration of an AI-based multi-sensor system for early detection and delayed spread of electric vehicle fires in charging facilities) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

REFERENCES

[1] Christopher Agia, Rohan Sinha, Jingyun Yang, Zi-ang Cao, Rika Antonova, Marco Pavone, and Jeannette Bohg. Unpacking failure modes of generative policies: Runtime monitoring of consistency and progress. In *Conference on Robot Learning*, 2024. URL https://arxiv.org/abs/2410.04640. arXiv:2410.04640.

[2] Kevin Black, Noah Brown, James Darpinian, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[4] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations (ICLR)*, 2024.

[5] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

[6] Shelly Francis-Meretzki, Mirco Mutti, Yaniv Romano, and Aviv Tamar. Temporal difference calibration in sequential tasks: Application to vision-language-action models. *arXiv preprint arXiv:2604.20472*, 2026.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 1050–1059, 2016.

[8] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2336–2346. Association for Computational Linguistics, 2024.

[9] 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 (NeurIPS)*, 2025.

[10] Ulas Berk Karli, Tetsu Kurumisawa, and Tesca Fitzgerald. Ask before you act: Token-level uncertainty for intervention in vision-language-action models. RSS 2025 Workshop on Out-of-Distribution Generalization in Robot Learning, 2025. URL https://openreview.net/forum?id=NX0euXAv98.

[11] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[12] Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[13] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[14] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[16] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

[17] 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.

[18] Litian Liu, Reza Pourreza, Sunny Panchal, Apratim Bhattacharyya, Yubing Jian, Yao Qin, and Roland Memisevic. Enhancing hallucination detection through noise injection. In *International Conference on Learning Representations (ICLR)*, 2026.

[19] Jun Nie, Yonggang Zhang, Tongliang Liu, Yiu ming Cheung, Bo Han, and Xinmei Tian. Epistemic uncertainty for generated image detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[20] Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39556–39586. PMLR, 2024. URL https://proceedings.mlr.press/v235/papamarkou24b.html.

[21] 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 (NeurIPS)*, 2025.

[22] Hanjing Wang and Qiang Ji. Epistemic uncertainty quantification for pre-trained neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[23] Youguang Xing, Xu Luo, Junlin Xie, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Shortcut learning in generalist robot policies: The role of dataset diversity and fragmentation. In *Conference on Robot Learning (CoRL)*, 2025.

[24] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

[25] 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.