ROBUSTNESS MARGINS FOR LOCAL VISUAL PERTURBATIONS IN VISION–LANGUAGE–ACTION POLICIES WITH FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies often remain effective under visual disturbances such as changes in lighting, but it is not well understood why their behavior is insensitive to some perturbations and highly sensitive to others. As a first step toward understanding this robustness, we study a more local question: given a fixed observation and a perturbation to its visual conditioning, how much can the distribution of executed action chunks change? We formulate this question probabilistically and derive a local bound on the shift of the action distribution under small perturbations. The bound depends on the magnitude and direction of the input perturbation together with the sensitivity of the policy's flow to its conditioning variable, characterized through the Jacobian. This provides a way to relate perturbations in observation space to changes in the distribution of sampled actions without making assumptions about downstream task success. We empirically study this relationship in LIBERO using several classes of visual perturbations and compare the predicted sensitivity with the observed shifts in action distributions. Our results provide an initial framework for analyzing VLA robustness through the geometry of how conditioning perturbations propagate through the generative action policy.

1 INTRODUCTION

Vision-language-action (VLA) policies combine visual and language understanding with continuous robot control, demonstrating increasingly broad manipulation capabilities across tasks, environments, and embodiments (Kim et al., 2024; Ghosh et al., 2024; Black et al., 2026; Intelligence et al., 2025; NVIDIA et al., 2025). Their deployment makes robustness to observation changes a central concern. In policies that execute action chunks open-loop between observations, a visual disturbance can affect several successive commands before the policy receives new feedback (Black et al., 2026). Systematic evaluations have revealed substantial performance degradation under changes in viewpoint, lighting, and other input conditions (Fei et al., 2025) and motivated robust training and test-time mitigation strategy (Guo et al., 2026; Kwok et al., 2026). These empirical advances leave a complementary question: *under what conditions can we establish a provable robustness margin for a trained VLA?* A useful starting point is the local case: for a fixed observation and a specified natural visual perturbation, can we rigorously bound the change in the policy's action distribution and the probability of retaining its typical behavior?

Establishing such a bound is difficult because action generation is inherently probabilistic. Flow-matching VLAs such as $\pi_{0.5}$, GR00T N1, and SmolVLA generate continuous action chunks by iteratively transforming random initial noise, conditioned on visual observations, language, and robot state (Intelligence et al., 2025; NVIDIA et al., 2025; Shukor et al., 2025). A visual perturbation first changes the latent context produced by the vision-language backbone and then modifies the velocity field throughout the action-generating flow (Lipman et al., 2023). Its effect therefore depends on both the perturbation direction and the sampled flow trajectory. Standard neural-network robustness certificates bound deterministic outputs over input uncertainty sets, for example to establish classification invariance (Wong & Kolter, 2018; Gowal et al., 2019). Here, we must additionally account for how conditioning changes propagate through iterative sampling and alter an entire action distribution. Bounding only the magnitude of the context change loses directional information, while a

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.

bound covering every direction in a context ball must account for sensitive directions that may not arise from the specified visual perturbation.

In this paper, we derive a rigorous, direction-dependent bound on the Wasserstein-2 distance between clean and perturbed action distributions under local smoothness and curvature assumptions. We fix the observation, language instruction, and robot state, and measure the context displacement by evaluating the vision-language backbone on the clean and perturbed images. Coupling the two samplers through shared initial noise lets us separate the action displacement into a first-order response, propagated through clean-trajectory Jacobians, and a nonlinear remainder. We bound this remainder by recursively constructing tubes around the first-order prediction and controlling curvature within them, without evaluating perturbed flow trajectories. The construction retains the measured perturbation direction while accounting for uncertainty in the resulting state displacement. We then combine the distribution-shift bound with a standard optimal-transport probability bound (Blanchet & Murthy, 2017) to lower-bound the probability that perturbed actions remain in a region typical of the clean policy. Accounting for the predicted mean translation can tighten this probability bound. The resulting guarantee concerns local action behavior; connecting it to closed-loop task success requires additional analysis.

We evaluate the bounds on a $\pi_{0.5}$ policy finetuned on LIBERO (Liu et al., 2023) using six visual perturbation families: global brightness and contrast, and patch-wise noise, blur, saturation and warping. We compare numerical estimates of our bounds with shared-noise action displacements, empirical Wasserstein-2 distances, and observed coverage of a clean-action region. Perturbations of similar context magnitude produce substantially different action shifts: at one matched magnitude, patch noise produces more than 6 times the shift of contrast. For small context changes, the tube-based estimate closely tracks the measured shift, whereas unrestricted curvature estimates are substantially larger. Translation adjustment also gives tighter estimates of clean-region retention. At larger magnitudes, nonlinear effects make the estimates increasingly conservative, sometimes uninformative even when observed retention remains high. These computations use sampled trajectories and numerical curvature searches, so they evaluate the theoretical bounds' practical behavior without constituting certified numerical guarantees.

Our contributions are threefold. First, we derive a directional Wasserstein-2 bound using clean flow trajectories and recursive curvature control (Section 4). Second, we bound clean-action retention probability, with a refinement accounting for predicted translation (Section 4.3). Third, we evaluate six visual perturbation families, comparing numerical bounds with observed action shifts and retention to identify their accuracy and limitations (Sections 5 and 6).

2 Related Work

Data scale and generalization in robot policies. A major driver of recent progress in robot learning has been the increasing scale and diversity of training data, following earlier developments in language and vision (Kaplan et al., 2020; Huh et al., 2016). Open X-Embodiment aggregates data across robot tasks and embodiments (Collaboration et al., 2025), while DROID collects in-the-wild manipulation data across diverse scenes and tasks (Khazatsky et al., 2025). Models such as π_0, $\pi_{0.5}$, and GR00T N1 further combine heterogeneous robot data with pretrained visual and language representations (Black et al., 2026; Intelligence et al., 2025; NVIDIA et al., 2025). Broader training coverage can support generalization, but data scale alone does not characterize how a trained policy responds to a particular observation change.

Improving robustness through training data and adaptation. LIBERO evaluates knowledge transfer across language-conditioned manipulation tasks, including changes in spatial configurations, objects, and goals (Liu et al., 2023). LIBERO-Plus expands its environmental and observational diversity through systematic perturbations, providing both a more realistic evaluation benchmark and an augmented training dataset (Fei et al., 2025). Its experiments expose robustness gaps under these changes and show that training on the expanded data substantially improves performance. Robustness also depends on how a policy is trained and adapted. Guo et al. (2026) introduce RobustVLA, which modifies flow-matching training to improve robustness to input perturbations and worst-case action noise. Li et al. (2026) improve viewpoint robustness through lightweight adaptation of visual tokens or the visual encoder, keeping the remaining policy fixed. Our work com-

plements these approaches by analyzing how a measured conditioning change propagates through an already trained action model.

Generative action policies and approximation bounds. Diffusion and flow matching models learn to transform samples from a simple noise distribution into samples from a target data distribution (Ho et al., 2020; Lipman et al., 2023). Diffusion Policy adapts this approach to robotic control by generating action sequences conditioned on visual observations (Chi et al., 2024). Models such as π_0, $\pi_{0.5}$, GR00T N1, and SmolVLA use flow matching to generate continuous action chunks (Black et al., 2026; Intelligence et al., 2025; NVIDIA et al., 2025; Shukor et al., 2025). Theoretical work examines how accurately these generative models approximate their target distributions. For diffusion models with stochastic sampling, convergence analyses relate score-estimation and discretization errors to error in the generated distribution, establishing polynomial guarantees under suitable assumptions (Lee et al., 2023; Benton et al., 2024a). For deterministic flow-matching sampling, Benton et al. (2024b) derive Wasserstein-2 bounds in terms of velocity-field approximation error and regularity assumptions. Our work instead studies how the generated distribution changes at finite sampling steps at inference time when the conditioning of a trained model is perturbed.

3 PRELIMINARIES

Conditional Flow Matching. Let $p_{\mathrm{data}}(x, c)$ denote a data distribution over samples x and conditioning variables c. Conditional flow matching (Lipman et al., 2023) learns a time-dependent velocity field $u_t^\theta(z_t \mid c)$ that transports samples from a reference distribution p_0 toward the conditional data distribution $p_{\mathrm{data}}(x \mid c)$. Using the linear probability path,

$$z_t = (1-t)z_0 + tx, \qquad z_0 \sim p_0, \qquad t \sim \mathcal{U}[0,1], \tag{1}$$

the target velocity is $x - z_0$, yielding the standard objective

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{\substack{(x,c)\sim p_{\mathrm{data}},\\ z_0\sim p_0,\, t\sim\mathcal{U}[0,1]}} \left[\left\| u_t^\theta(z_t \mid c) - (x - z_0) \right\|_2^2 \right]. \tag{2}$$

At inference time, the learned velocity field defines the conditional ordinary differential equation $dz_t/dt = u_t^\theta(z_t \mid c)$, initialized with $z_0 \sim p_0$. We consider an N-step Euler discretization with $t_k = k/N$, $h = 1/N$, and define the one-step map

$$\psi_k(z \mid c) \coloneqq z + h\, u_{t_k}^\theta(z \mid c). \tag{3}$$

The flow sampling trajectory is then $z_{k+1} = \psi_k(z_k \mid c)$ for $k = 0, \ldots, N-1$. Along a clean flow trajectory, denote the state and context Jacobians of the Euler map by

$$\boldsymbol{J}_k^z \coloneqq \nabla_{\boldsymbol{z}} \psi_k(\boldsymbol{z}_k \mid \boldsymbol{c}) = \boldsymbol{I} + h \nabla_{\boldsymbol{z}} u_{t_k}^\theta(\boldsymbol{z}_k \mid \boldsymbol{c}), \tag{4}$$

$$\boldsymbol{J}_k^c \coloneqq \nabla_{\boldsymbol{c}} \psi_k(\boldsymbol{z}_k \mid \boldsymbol{c}) = h \nabla_{\boldsymbol{c}} u_{t_k}^\theta(\boldsymbol{z}_k \mid \boldsymbol{c}). \tag{5}$$

For a twice continuously differentiable vector-valued function f, let $\nabla^2 f(\boldsymbol{x})[\boldsymbol{v}, \boldsymbol{w}]$ be the Hessian of f evaluated along directions $\boldsymbol{v}$ and $\boldsymbol{w}$. For vector inputs and outputs, its ith component is $\left(\nabla^2 f(\boldsymbol{x})[\boldsymbol{v}, \boldsymbol{w}]\right)_i = \boldsymbol{v}^\top \nabla^2 f_i(\boldsymbol{x}) \boldsymbol{w}$ for $i = 1, \ldots, m$. The same derivative notation applies to tensor inputs and outputs.

Vision-Language-Action Policy. We consider a VLA policy following the architecture of $\pi_{0.5}$ (Intelligence et al., 2025). An observation $o = (I^{1:M}, \ell, q)$ contains camera images $I^{1:M}$, a language instruction ℓ, and robot proprioception q. The vision–language backbone maps this multimodal input to the context supplied to the flow-matching action head, $\boldsymbol{c} = \mathrm{VLM}(o)$. For $\pi_{0.5}$, $\boldsymbol{c}$ consists of the layerwise key-value representations of the multimodal prefix.

The flow-matching policy head generates an action chunk by transporting an initial noise sample $\boldsymbol{z}_0$ through the learned velocity field conditioned on the context $\boldsymbol{c}$. The flow states $\boldsymbol{z}_k$, $k < N$, represent intermediate stages of the generative transport, while the terminal state $\boldsymbol{z}_N$ is the generated action chunk sampled from the learned conditional action distribution. Only a subset of the terminal chunk is executed. Let S denote the corresponding linear selection operator that extracts the executed time steps and action coordinates from $\boldsymbol{z}_N$. The executed action is $\boldsymbol{a} = S\boldsymbol{z}_N$.

4 ROBUSTNESS BOUND

We characterize the local effect of a specified visual perturbation on the executed-action distribution of a flow-matching VLA. With language and proprioception fixed, a clean observation $\boldsymbol{o}$ and its perturbed counterpart $\boldsymbol{o}'$ induce the contexts

$$\boldsymbol{c} = VLM(\boldsymbol{o}), \qquad \boldsymbol{c}' = VLM(\boldsymbol{o}') = \boldsymbol{c} + \Delta\boldsymbol{c}. \tag{6}$$

Our analysis bounds the induced distribution shift using the measured context displacement $\Delta\boldsymbol{c}$, the clean flow trajectory, and local curvature bounds, without evaluating the perturbed flow. We separate the directional first-order response, obtained by propagating $\Delta\boldsymbol{c}$ through the linearized clean flow, from the nonlinear remainder, which captures higher-order effects of the finite displacement. The resulting bound yields a probabilistic guarantee of clean-action retention.

4.1 PERTURBATION PROPAGATION AND DISTRIBUTION SHIFT

Couple the clean and perturbed samplers through *shared initial noise* $\boldsymbol{z}'_0 = \boldsymbol{z}_0$:

$$\boldsymbol{z}_{k+1} = \psi_k(\boldsymbol{z}_k \mid \boldsymbol{c}), \qquad \boldsymbol{z}'_{k+1} = \psi_k(\boldsymbol{z}'_k \mid \boldsymbol{c}'). \tag{7}$$

Denote the step-k displacement as $\boldsymbol{\delta}_k \coloneqq \boldsymbol{z}'_k - \boldsymbol{z}_k$. A joint Taylor expansion about the clean state-context pair $(\boldsymbol{z}_k, \boldsymbol{c})$ gives the exact recursion

$$\boldsymbol{\delta}_{k+1} = \boldsymbol{J}^z_k \boldsymbol{\delta}_k + \boldsymbol{J}^c_k \Delta\boldsymbol{c} + \boldsymbol{e}_k \tag{8}$$

where $\boldsymbol{e}_k$ is the joint state-context nonlinear remainder. The term $\boldsymbol{J}^z_k \boldsymbol{\delta}_k$ propagates accumulated displacement, whereas $\boldsymbol{J}^c_k \Delta\boldsymbol{c}$ injects the context perturbation. All Jacobians are evaluated along the clean trajectory. For $j < k$, define the chronological transition operator

$$\boldsymbol{T}_{k,j} = \boldsymbol{J}^z_{k-1} \boldsymbol{J}^z_{k-2} \cdots \boldsymbol{J}^z_j, \qquad \boldsymbol{T}_{j,j} = \boldsymbol{I}. \tag{9}$$

Repeated substitution of equation 8, together with $\boldsymbol{\delta}_0 = 0$, yields

$$\boldsymbol{\delta}_k = \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{J}^c_\ell \Delta\boldsymbol{c} + \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{e}_\ell, \tag{10}$$

which exactly decomposes the step-k displacement into the directional first-order response to $\Delta\boldsymbol{c}$ and the nonlinear remainder, both propagated through the clean-trajectory Jacobians.

From pointwise displacement to distribution shift. The random initial noise $\boldsymbol{z}_0$ induces distributions P_k and Q_k over the clean and perturbed states $\boldsymbol{z}_k$ and $\boldsymbol{z}'_k$, respectively. Trajectories sharing the same initial noise defines a coupling of these distributions, and yields

$$W_2(P_k, Q_k) = \inf_{\gamma \in \Pi(P_k, Q_k)} \left(\mathbb{E}_{(\boldsymbol{z}, \boldsymbol{z}') \sim \gamma} \left[\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2\right]\right)^{1/2} \leq \left(\mathbb{E}_{\boldsymbol{z}_0} \left[\|\boldsymbol{\delta}_k\|_2^2\right]\right)^{1/2}. \tag{11}$$

Substituting the exact decomposition in (10) and applying the triangle inequality gives

$$W_2(P_k, Q_k) \leq \left(\mathbb{E}\left[\left\|\sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{J}^c_\ell \Delta\boldsymbol{c}\right\|_2^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left\|\sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{e}_\ell\right\|_2^2\right]\right)^{1/2}. \tag{12}$$

All quantities on the right-hand side of (12) are determined by the clean-trajectory Jacobians and the measured $\Delta\boldsymbol{c}$ except for $\boldsymbol{e}_\ell$, which depends on the perturbed states. We next bound these unknown terms through local curvature, without evaluating the perturbed flow.

4.2 BOUNDING THE NONLINEAR REMAINDER

The nonlinear remainder in (12) depends on the unknown perturbed states. To bound it without evaluating the perturbed flow, we require a reference displacement determined entirely by the clean trajectory and $\Delta\boldsymbol{c}$. We therefore define the first-order displacement prediction

$$\hat{\boldsymbol{\delta}}_k = \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{J}^c_\ell \Delta\boldsymbol{c}, \qquad \hat{\boldsymbol{\delta}}_0 = \boldsymbol{0}. \tag{13}$$

This displacement decomposition identifies the prediction error with the accumulated nonlinear remainder, $\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k = \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{e}_\ell$, whose L^2 norm is the unresolved term in (12).

To control this contribution, we bound each local remainder $\boldsymbol{e}_\ell$ using the velocity field's joint state-context Hessian over tubes around the first-order perturbation segments. We first define a radius-dependent curvature bound, then recursively construct tubes containing the true Taylor segments.

Recursive curvature bound. For a candidate radius $r \geq 0$, consider the tube formed by segments from the clean pair $(\boldsymbol{z}_k, \boldsymbol{c})$ to a candidate endpoint $(\boldsymbol{z}_k + \boldsymbol{d}, \boldsymbol{c} + \Delta\boldsymbol{c})$, where $\|\boldsymbol{d} - \hat{\boldsymbol{\delta}}_k\|_F \leq r$. Define

$$\overline{\mathcal{M}}_k(r) \coloneqq \sup_{\substack{s\in[0,1],\\ \|\boldsymbol{d}-\hat{\boldsymbol{\delta}}_k\|_F \leq r}} \left\| \nabla^2 u^\theta_{t_k}(\boldsymbol{z}_k + s\boldsymbol{d} \mid \boldsymbol{c} + s\Delta\boldsymbol{c})[(\boldsymbol{d}, \Delta\boldsymbol{c}), (\boldsymbol{d}, \Delta\boldsymbol{c})] \right\|_F, \tag{14}$$

Thus, $\overline{\mathcal{M}}_k$ bounds the curvature over an explicitly specified region, without reference to an unknown perturbed state. Starting from $r_0 = 0$, we have:

$$r_{k+1} = \|\boldsymbol{J}^z_k\|_{\mathrm{op}} r_k + \frac{h}{2}\overline{\mathcal{M}}_k(r_k), \qquad k = 0, \ldots, N-1. \tag{15}$$

Set $\overline{M}_k = \overline{\mathcal{M}}_k(r_k)$. We now show that these radii certify $\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F \leq r_k$. The claim holds initially because $\delta_0 = \hat{\delta}_0 = 0$. If it holds at step k, the true displacement δ_k is admissible in M_k. The integral Taylor remainder therefore gives $\|\boldsymbol{e}_k\|_F \leq h\overline{M}_k/2$, and

$$\|\delta_{k+1} - \hat{\delta}_{k+1}\|_F = \|J^z_k(\delta_k - \hat{\delta}_k) + \boldsymbol{e}_k\|_F \leq \|J^z_k\|_{\mathrm{op}} r_k + \tfrac{h}{2}\overline{M}_k = r_{k+1}. \tag{16}$$

Thus, induction establishes containment at every flow step.

Theorem 1 (Wasserstein-2 bound with tube curvature) *Suppose the velocity fields are twice continuously differentiable on neighborhoods of the tubes defined above, with finite $\overline{M}_\ell$ for every $\ell < N$ almost surely. Assume that the clean and perturbed state distributions have finite second moments. For every $k \leq N$,*

$$W_2(P_k, Q_k) \leq \left(\mathbb{E}[\|\hat{\boldsymbol{\delta}}_k\|_F^2]\right)^{1/2} + \frac{h}{2}\left(\mathbb{E}\left[\left(\sum_{\ell=0}^{k-1} \|\boldsymbol{T}_{k,\ell+1}\|_{\mathrm{op}}\overline{M}_\ell\right)^2\right]\right)^{1/2}. \tag{17}$$

The expectation is over the shared initial noise $\boldsymbol{z}_0$, on which the clean trajectory, predictions, radii, and curvature quantities depend. A proof is given in Appendix A.1.1.

4.3 From Bounded Distribution Shift to Probability Mass within a Region

Let P and Q denote the clean and perturbed *executed-action* distributions, respectively (the laws of Sz_N and Sz'_N). For executed actions, Theorem 1 applies with $\hat{\boldsymbol{\delta}}_N$ replaced by $S\hat{\boldsymbol{\delta}}_N$ and $\boldsymbol{T}_{N,\ell+1}$ replaced by $S\boldsymbol{T}_{N,\ell+1}$ in the terminal bound. The full-state tube construction remains unchanged. Denote the resulting bound by B, so that

$$W_2(P, Q) \leq B. \tag{18}$$

Since W_2^2 is the minimum expected squared transport cost between two distributions, B^2 gives a transport budget for the perturbation. We use this budget to control the probability that perturbed actions remain within a prescribed clean-action region.

Clean-action region. For a measurable scalar score $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and threshold $\tau \in \mathbb{R}$, define

$$A_\tau \coloneqq \{\boldsymbol{a} \in \mathbb{R}^n : f(\boldsymbol{a}) \leq \tau\}. \tag{19}$$

In our work, we consider f as the negative log of a density estimate fitted to clean action projections onto a fixed subspace. Our experiments use density estimates of clean principal-component scores, so retention measures typicality under the chosen projected score.

For any action $\boldsymbol{a} \in \mathbb{R}^n$, define its distance to the complement of the region,

$$d(\boldsymbol{a}) = d(\boldsymbol{a}, A_\tau^c) = \inf_{\tilde{\boldsymbol{a}} \notin A_\tau} \|\boldsymbol{a} - \tilde{\boldsymbol{a}}\|_2. \tag{20}$$

Thus $d(\boldsymbol{a}) = 0$ whenever $\boldsymbol{a} \notin A_\tau$. For $\boldsymbol{a} \in A_\tau$, transporting $\boldsymbol{a}$ outside the region incurs squared cost at least $d(\boldsymbol{a})^2$. Consequently, these margins, together with the transport budget B^2, yield the following retention guarantee.

Theorem 2 (Worst-case probability bound under W_2 budget) *If $W_2(P, Q) \le B$, then*

$$Q(A_\tau) \ge \sup_{\lambda \ge 0} \left\{ \mathbb{E}_{\boldsymbol{a} \sim P} \left[\min \left\{ 1, \lambda d(\boldsymbol{a})^2 \right\} \right] - \lambda B^2 \right\}. \tag{21}$$

This bound is standard and follows from the worst-case probability reformulation of Blanchet & Murthy (2017), specialized to the type-2 Wasserstein distance. It converts the shift budget and clean-action margins into a lower bound on perturbed-policy coverage.

Accounting for systematic translation. The scalar budget B^2 bounds transport cost but does not specify the directions in which mass moves. Theorem 2 therefore allows transport in the directions that most efficiently reduce coverage of A_τ. This can be conservative when the displacement contains a predictable translation component. We account for this component using the predicted mean executed-action shift,

$$m \coloneqq \mathbb{E}_{\boldsymbol{z}_0} \left[S \hat{\boldsymbol{\delta}}_N \right] = \mathbb{E}_{\boldsymbol{z}_0} \left[\sum_{\ell=0}^{N-1} S \boldsymbol{T}_{N,\ell+1} \boldsymbol{J}_\ell^c \Delta \boldsymbol{c} \right]. \tag{22}$$

This is the mean first-order prediction, rather than the exact mean action shift. Let $Q' = Q - m$ denote the distribution obtained by subtracting m from each perturbed action. The shared-noise coupling and the same remainder bound give $W_2(P, Q') \le b$, where

$$b = \left(\mathbb{E}_{\boldsymbol{z}_0} [\| S \hat{\boldsymbol{\delta}}_N - m \|_F^2] \right)^{1/2} + \frac{h}{2} \left(\mathbb{E}_{\boldsymbol{z}_0} \left[\left(\sum_{\ell=0}^{N-1} \| S \boldsymbol{T}_{N,\ell+1} \|_{\mathrm{op}} \overline{M}_\ell \right)^2 \right] \right)^{1/2}. \tag{23}$$

The first term captures variation of the predicted executed-action displacement around its mean, while the second bounds the accumulated nonlinear remainder. Thus b bounds the residual distributional discrepancy after subtracting m. We apply the probability bound to this residual budget while retaining the translation in the region geometry.

Corollary 1 (Probability bound after translation) *Let P and Q be distributions on $\mathbb{R}^n$ with finite second moments, and let A_τ be a measurable region. For a specified translation $m \in \mathbb{R}^n$, let $Q' = Q - m$ denote the distribution obtained by subtracting m from each sample drawn from Q. If, for some $b \ge 0$,*

$$W_2(P, Q') \le b, \tag{24}$$

then

$$Q(A_\tau) \ge \sup_{\lambda \ge 0} \left\{ \mathbb{E}_{\boldsymbol{a} \sim P} \left[\min \left\{ 1, \lambda d(\boldsymbol{a} + m)^2 \right\} \right] - \lambda b^2 \right\}. \tag{25}$$

A proof is given in Appendix A.1.2.

5 Experiments

Our experiments address three questions: (1) How does perturbation direction affect action-distribution shifts at matched context magnitudes? (2) How closely does the numerical estimate of our direction-dependent bound track these shifts, and how does it compare with oracle and worst-direction curvature calculations? (3) What do the resulting probabilistic lower bounds tell us about how likely perturbed actions will stay within a region typical of the clean policy?

Model, observations, and visual perturbations. We analyze a $\pi_{0.5}$ policy finetuned on the LIBERO dataset (Liu et al., 2023). The main experiments use observations from two LIBERO tasks shown in Figure 1. Fixing the observation allows us to examine the effects of perturbation direction and magnitude on the action distribution. The observation contains language, robot proprioception and isometric and wrist-view images, preprocessed to 224×224 with pixel values in $[-1, 1]$. We perturb the wrist-view image while holding the isometric view fixed and evaluate the VLM on the clean and perturbed observations to measure $\Delta \boldsymbol{c}$. All reported action distances use the selected executed-action chunk in normalized policy coordinates, with the Frobenius norm.

We construct six perturbation families: brightness, contrast, patch noise, patch blur, patch saturation, and patch warping. Each family defines a direction by normalizing the difference between the clean

input and a reference-transformed input jointly over both cameras. We evaluate 10 magnitudes along this direction, clipping image values to $[-1, 1]$ and recording the achieved displacement. We compare families at matched measured context magnitudes $\|\Delta \boldsymbol{c}\|_F$. Figure 1 illustrates the resulting image changes at selected magnitudes of change in context.

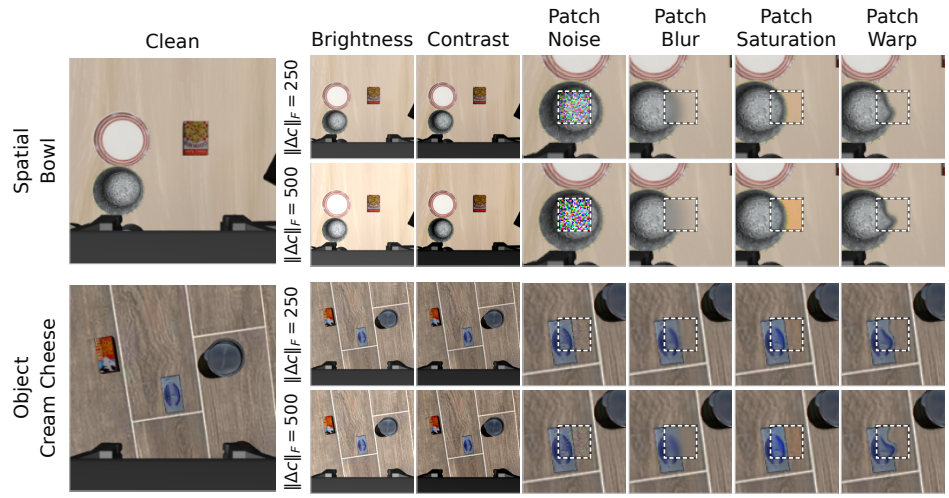

Figure 1: Visual perturbations on the LIBERO dataset at context-displacement magnitudes of $\|\Delta \boldsymbol{c}\|_F = 250$ and 500. The tasks are: "pick up the black bowl between the plate and the ramekin and place it on the plate" (top) and "pick up the cream cheese and place it in the basket" (bottom).

Empirical comparison of the derived Wasserstein-2 bound. We evaluate how closely the numerical estimate of Theorem 1 tracks executed-action distribution shifts. The calculation uses the measured context change, clean flow trajectories, and curvature evaluations within the numerically constructed tubes and does not require knowledge of perturbed flow trajectories. We then compare our method in an ablation study with an oracle curvature calculation that uses the actual displacement direction and less informed curvature calculations. This ablation uses a separate suite of global brightness, pixel-noise, contrast, zoom, Gaussian-blur, and saturation perturbations applied to both cameras, whereas the main study perturbs only one camera. Numerical curvature procedures are described in Section A.2.1 of the Appendix.

For evaluation, we run clean and perturbed samplers using shared Gaussian initial-noise draws. We compute the root mean square between paired executed-action samples and compute Wasserstein-2 distance between the empirical action distributions by optimal assignment.

We plot these quantities against $\|\Delta \boldsymbol{c}\|_F$ for each perturbation family. Comparisons at matched context magnitudes reveal directional differences in action sensitivity, while comparisons between the numerical bound, paired RMS, and empirical Wasserstein-2 distance assess the accuracy of the bound estimate and the cost of the shared-noise coupling. The curvature searches and Monte Carlo calculations are numerical approximations rather than certified upper bounds.

Probabilistic Action Retention. We evaluate the probability bound of Section 4.3 using a region $A_\tau = \{\boldsymbol{a} : f(\boldsymbol{a}) \leq \tau\}$ constructed from clean action samples. We hold this region fixed under visual perturbations, so $Q(A_\tau)$ measures how often the perturbed policy generates actions within the selected clean-action region.

Using evaluation samples, we estimate the clean and perturbed coverage by $\widehat{P}(A_\tau) = M^{-1} \sum_{i=1}^{M} \mathbb{1}\{\boldsymbol{a}_i \in A_\tau\}$ and $\widehat{Q}(A_\tau) = M^{-1} \sum_{i=1}^{M} \mathbb{1}\{\boldsymbol{a}'_i \in A_\tau\}$, respectively. For the translation-adjusted calculation in Corollary 1, we replace d_i by $d(\boldsymbol{a}_i + \widehat{m}, A_\tau^c)$ and B by the estimated residual budget $\widehat{b}$ from Equation 23. We compare these estimates with clean and perturbed coverage across perturbation families and magnitudes. Because both the budgets and expectations are estimated numerically, the reported values are estimates of the theoretical probability bounds.

6 Results and Analysis

Wasserstein-2 Bound on Perturbed Action Chunks. Table 1 shows that the numerical estimate of Theorem 1 closely tracks the empirical Wasserstein-2 distance at small context displacements. At $\|\Delta c\|_F = 100$, the estimates are within a factor of approximately 1.8 of the empirical distances across all six perturbation families and both observations. Figure 2 further illustrates this agreement and the increasing conservatism at larger displacements.

Table 1 also shows that context-displacement magnitude alone does not determine the action-distribution shift. For the black-bowl observation at $\|\Delta c\|_F = 500$, patch noise produces an empirical Wasserstein-2 distance of 1.2, compared with 0.20 for contrast. This ordering reverses for the cream-cheese observation, where contrast produces a larger shift than patch noise. These comparisons highlight the importance of accounting for perturbation direction in the robustness bound.

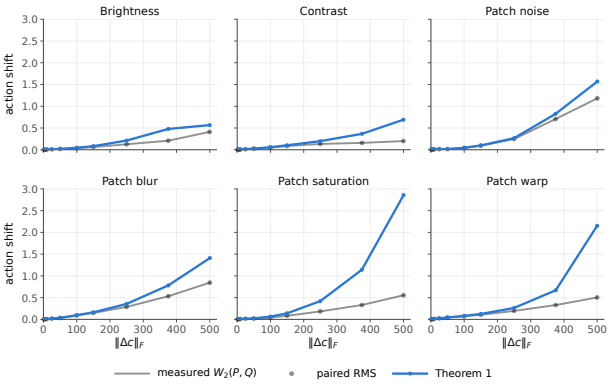

Figure 2: **Numerical Wasserstein-2 bound.** Empirical action-distribution shifts and Theorem 1 across perturbation magnitudes for the black bowl pick and place task.

As context displacement increases, larger tubes may be needed to control the nonlinear remainder. These tubes admit a broader range of state displacements over which curvature is bounded, and the resulting remainder estimates are amplified through subsequent flow steps. This can make the estimates increasingly conservative, even when the observed action shift remains small. The results support the construction's usefulness in the local regime while showing that its numerical estimates can become uninformative at larger displacements.

Numerical Curvature Ablation Study. Figure 3 compares curvature estimates that retain different amounts of directional information. All curvature variants are compared under the same ablation protocol, detailed in Appendix A.2.1. Our tube-based estimate (Theorem 1) closely tracks the empirical Wasserstein-2 shift and the oracle estimate (green) using the true displacement at small context displacements, without requiring perturbed flow trajectories. Even with the true displacement direction, the oracle bounds the magnitude of the nonlinear remainder and its propagation, so its estimate can remain conservative relative to the empirical shift. The closeness of our estimate to the oracle suggests that controlling the unknown displacement through the tube adds little conservatism in the local regime.

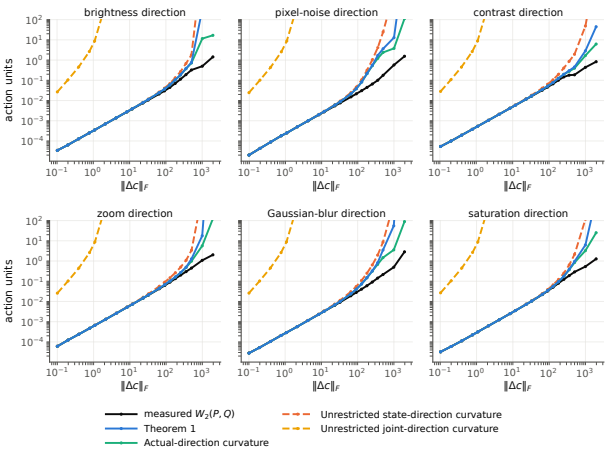

Figure 3: **Curvature Ablation.** Numerical approximation of the Wasserstein-2 bound from Theorem 1 and comparison with alternative curvature calculations.

In contrast, allowing arbitrary state or joint state-context directions yields substantially looser estimates that grow more rapidly with perturbation magnitude. The unrestricted joint-direction calculation searches over both context and flow-state directions and uses the largest curvature value found by the iterative search. This comparison highlights the benefit of preserving the measured context direction and controlling state-displacement uncertainty around the first-order prediction.

Table 1: Comparison of the empirical Wasserstein-2 (W_2) estimate with our Theorem 1 bound under perturbation magnitudes $\|\Delta c\|_F = \{100, 250, 500\}$ on two separate tasks at a fixed observation.

	LIBERO Spatial: Black Bowl						**LIBERO Object: Cream Cheese**					
	$\|\Delta c\|_F = 100$		250		500		100		250		500	
Perturbation	W_2	**Thm. 1**	W_2	**Thm. 1**	W_2	**Thm. 1**	W_2	**Thm. 1**	W_2	**Thm. 1**	W_2	**Thm. 1**
Brightness	0.038	**0.045**	0.13	**0.21**	0.41	**0.57**	0.081	**0.11**	0.050	**0.063**	0.42	**1.6**
Contrast	0.051	**0.058**	0.13	**0.20**	0.20	**0.69**	0.12	**0.17**	0.29	**0.88**	0.54	**6.0**
Patch noise	0.041	**0.044**	0.25	**0.27**	1.2	**1.6**	0.020	**0.029**	0.076	**0.16**	0.25	**1.1**
Patch blur	0.092	**0.098**	0.29	**0.35**	0.84	**1.4**	0.023	**0.029**	0.089	**0.14**	0.35	**0.94**
Patch saturation	0.042	**0.064**	0.18	**0.42**	0.55	**2.9**	0.074	**0.090**	0.11	**0.17**	0.12	**79**
Patch warp	0.071	**0.080**	0.19	**0.26**	0.50	**2.2**	0.028	**0.050**	0.070	**9.5**	0.18	**120**

Probabilistic Action Retention. Beyond quantifying distribution shift, our Wasserstein-2 bound can be used to lower-bound the probability that perturbed actions remain within a region typical of the clean policy. Figure 4 illustrates this for a small brightness perturbation. We define $A_{0.95}$ as a clean-action region containing approximately 95% of clean samples. The perturbed policy retains 94.4% of its samples within this region, while Corollary 1 together with our distribution-shift bound yields a numerical lower-bound estimate of 82%. This provides one example of how a known perturbation direction can yield an informative numerical estimate of clean-action retention under a new visual condition. Appendix A.3 shows that translation adjustment gives tighter retention estimates, even when the unadjusted estimates fall to zero.

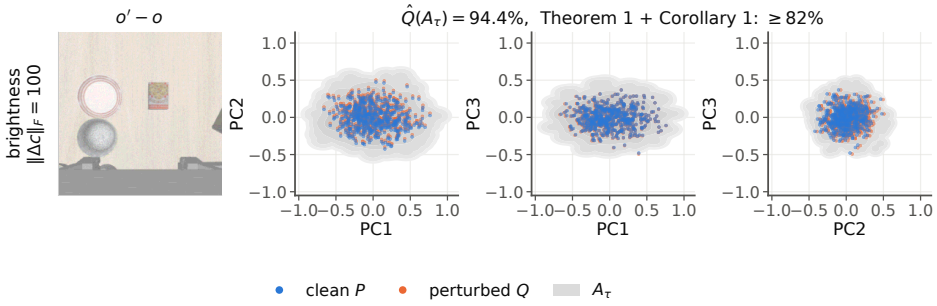

Figure 4: **Probabilistic Action Retention.** Clean and perturbed action samples projected onto the first three principal components under a brightness perturbation. $A_{0.95}$ denotes a region containing approximately 95% of clean actions. Observed perturbed coverage is 94.4%, while our Wasserstein-2 bound together with Corollary 1 yields a lower-bound estimate of 82%.

7 Conclusion and Future Work

We study how a specified visual perturbation changes the action distribution of a flow-matching VLA at fixed observations. Our analysis illustrates why perturbation direction matters in the high-dimensional conditioning space. Context changes of similar magnitude can produce substantially different action shifts, while bounds uniform over a context ball must also account for sensitive directions that may not arise from the specified visual change. Given the measured context change, our direction-dependent Wasserstein-2 bound provides a way to analyze this response without requiring perturbed flow trajectories.

This distributional perspective opens new possibilities for understanding VLA robustness. We explore one application by using a standard probability bound to estimate how likely the perturbed policy is to produce actions typical of the clean observation. Our experiments illustrate the potential of this approach for small perturbations, while highlighting the need for tighter numerical control of nonlinear effects at larger magnitudes. Extending action-level bounds across successive observations and connecting them to closed-loop task success are promising directions for future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for aiding and polishing writing, debugging code, and providing feedback on research methodology and mathematical claims. We have not used generative AI tools for generating synthetic datasets, help developing theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, implementing methods, supporting qualitative and thematic data analysis, interpreting results. AI assisting with translation, cleaning and reformatting datasets are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We used a pretrained $\pi_{0.5}$ model in this work finetuned on the LIBERO dataset. Weights for this model are publicly available at `https://huggingface.co/lerobot/pi05_libero_finetuned_v044`.

The main text identifies the evaluated tasks, and the appendix describes the numerical curvature calculations and probability estimates. Tabulated numerical results are provided in Appendix A.6.

REFERENCES

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization, 2024a. URL `https://arxiv.org/abs/2308.03686`.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods, 2024b. URL `https://arxiv.org/abs/2305.16860`.

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, 2026. URL `https://arxiv.org/abs/2410.24164`.

Jose Blanchet and Karthyek R. A. Murthy. Quantifying distributional model risk via optimal transport, 2017. URL `https://arxiv.org/abs/1604.01446`.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2024. URL `https://arxiv.org/abs/2303.04137`.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake,

Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025. URL https://arxiv.org/abs/2310.08864.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: In-depth robustness analysis of vision-language-action models, 2025. URL https://arxiv.org/abs/2510.13626.

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, 2024. URL https://arxiv.org/abs/2405.12213.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models, 2019. URL https://arxiv.org/abs/1810.12715.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, Qi Dou, Yaodong Yang, Xianglong Liu, Huijie Zhao, Weifeng Lv, and Simin Li. Robustvla: On robustness of vision-language-action model against multi-modal perturbations, 2026. URL https://arxiv.org/abs/2510.00037.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning?, 2016. URL https://arxiv.org/abs/1608.08614.

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, 2025. URL https://arxiv.org/abs/2504.16054.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

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, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, 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, 2025. URL https://arxiv.org/abs/2403.12945.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag 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, 2024. URL https://arxiv.org/abs/2406.09246.

Jacky Kwok, Xilun Zhang, Mengdi Xu, Yuejiang Liu, Azalia Mirhoseini, Chelsea Finn, and Marco Pavone. Scaling verification can be more effective than scaling policy learning for Vision-Language-Action alignment, 2026. URL https://arxiv.org/abs/2602.12281.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/lee23a.html.

Weiqi Li, Quande Zhang, Ruifeng Zhai, Liang Lin, and Guangrun Wang. Vla models are more generalizable than you think: Revisiting physical and spatial modeling, 2026. URL https://arxiv.org/abs/2512.02902.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

NVIDIA, :, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang,

Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025. URL https://arxiv.org/abs/2503.14734.

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, 2025. URL https://arxiv.org/abs/2506.01844.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295. PMLR, 2018. URL https://proceedings.mlr.press/v80/wong18a.html.

A Appendix

A.1 Proofs

A.1.1 Proof of the Wasserstein-2 Bound with Tube Curvature

Proof 1 *Fix a shared initial noise $\boldsymbol{z}_0$ and consider the clean and perturbed trajectories with contexts $\boldsymbol{c}$ and $\boldsymbol{c} + \Delta\boldsymbol{c}$. Their displacement is $\boldsymbol{\delta}_k = \boldsymbol{z}'_k - \boldsymbol{z}_k$. We first establish that the true displacement remains within the tube around its first-order prediction.*

Recall that $\psi_k(\boldsymbol{z} \mid \boldsymbol{c}) = \boldsymbol{z} + h\, u^\theta_{t_k}(\boldsymbol{z} \mid \boldsymbol{c})$. The identity term is linear, so the nonlinear residual arises entirely from the velocity field. Writing $\boldsymbol{x}_k = (\boldsymbol{z}_k, \boldsymbol{c})$ and $\Delta\boldsymbol{x}_k = (\boldsymbol{\delta}_k, \Delta\boldsymbol{c})$, the exact integral remainder gives

$$\boldsymbol{e}_k = h \int_0^1 (1-s)\nabla^2 u^\theta_{t_k}(\boldsymbol{x}_k + s\Delta\boldsymbol{x}_k)[\Delta\boldsymbol{x}_k, \Delta\boldsymbol{x}_k]\, ds. \tag{26}$$

At initialization, $\boldsymbol{\delta}_0 = \hat{\boldsymbol{\delta}}_0 = \mathbf{0}$ and $r_0 = 0$, so the prediction-error bound holds. Suppose that $\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F \le r_k$. Then $\boldsymbol{d} = \boldsymbol{\delta}_k$ is an admissible displacement in the supremum defining $\overline{M}_k$. The entire true interpolation segment therefore lies in the tube, and Equation 26 yields

$$\|\boldsymbol{e}_k\|_F \le h \int_0^1 (1-s)\big\|\nabla^2 u^\theta_{t_k}(\boldsymbol{x}_k + s\Delta\boldsymbol{x}_k)[\Delta\boldsymbol{x}_k, \Delta\boldsymbol{x}_k]\big\|_F\, ds \le h\overline{M}_k \int_0^1 (1-s)\, ds = \tfrac{h}{2}\overline{M}_k. \tag{27}$$

Subtracting the first-order prediction recursion from the exact displacement recursion gives

$$\boldsymbol{\delta}_{k+1} - \hat{\boldsymbol{\delta}}_{k+1} = \boldsymbol{J}^z_k \left(\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\right) + \boldsymbol{e}_k. \tag{28}$$

Consequently,

$$\|\boldsymbol{\delta}_{k+1} - \hat{\boldsymbol{\delta}}_{k+1}\|_F \le \|\boldsymbol{J}^z_k\|_{\mathrm{op}}\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F + \|\boldsymbol{e}_k\|_F \le \|\boldsymbol{J}^z_k\|_{\mathrm{op}} r_k + \tfrac{h}{2}\overline{M}_k = r_{k+1}. \tag{29}$$

By induction, the prediction-error bound holds for every $k \le N$, and $\|\boldsymbol{e}_k\|_F \le h\overline{M}_k/2$ for every $k < N$. Each radius r_k depends only on preceding steps, so this argument does not require knowledge of the perturbed trajectory to construct the tube.

Unrolling the prediction and error recursions gives

$$\hat{\boldsymbol{\delta}}_k = \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1}\boldsymbol{J}^c_\ell \Delta\boldsymbol{c}, \qquad \boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k = \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1}\boldsymbol{e}_\ell. \tag{30}$$

Thus,

$$\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F \le \frac{h}{2}\sum_{\ell=0}^{k-1} \|\boldsymbol{T}_{k,\ell+1}\|_{\mathrm{op}}\overline{M}_\ell. \tag{31}$$

The shared initial noise defines a coupling of P_k and Q_k. Using this coupling and the triangle inequality in L^2,

$$W_2(P_k, Q_k) \le \left(\mathbb{E}[\|\boldsymbol{\delta}_k\|_F^2]\right)^{1/2} \le \left(\mathbb{E}[\|\hat{\boldsymbol{\delta}}_k\|_F^2]\right)^{1/2} + \left(\mathbb{E}[\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F^2]\right)^{1/2}. \tag{32}$$

Substituting the expression for $\hat{\boldsymbol{\delta}}_k$ and Equation 31 proves Equation 17.

For executed actions, applying S to the terminal displacement decomposition gives $S\boldsymbol{\delta}_N = S\hat{\boldsymbol{\delta}}_N + \sum_{\ell=0}^{N-1} S\boldsymbol{T}_{N,\ell+1}\boldsymbol{e}_\ell$. The same coupling and triangle-inequality argument therefore gives the executed-action bound using $S\hat{\boldsymbol{\delta}}_N$ and $\|S\boldsymbol{T}_{N,\ell+1}\|_{\mathrm{op}}$.

A.1.2 Proof of Corollary 1

Proof 2 *Apply Theorem 2 to P and Q' with the region $A_\tau - m$. A sample drawn from Q' lies in $A_\tau - m$ exactly when the corresponding sample drawn from Q lies in A_τ. Thus $Q'(A_\tau - m) = Q(A_\tau)$. The distance from $\boldsymbol{a}$ to leaving $A_\tau - m$ is $d(\boldsymbol{a} + m)$, giving the stated lower bound.*

A.2 Details on Curvature Ablation Study

The following section provides additional details on the curvature ablation, including specifications of the new perturbations considered, and how each curvature constant is calculated numerically.

Figures 5 and 6 show the new classes of perturbations considered. Note that brightness and contrast are different global perturbations than the ones used in the main study because here they are applied to both agent view and wrist view images.

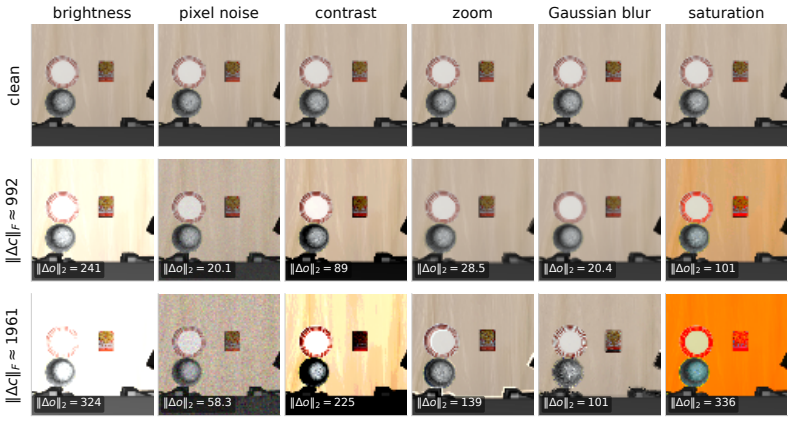

Figure 5: Image perturbations used for analysis for Libero-Spatial scene. Identical perturbations are also applied to the agent view camera. The same $\|\Delta c\|_F$ can be produced by vastly different ranges of image perturbations. Corresponding agent view visualizations of the same frame can be found in Figure 6.

A.2.1 Numerical Calculations of Curvature Constants

Figure 3 compares numerical curvature estimates that retain different amounts of perturbation-direction information. These comparisons examine how the choice of curvature calculation affects the estimated nonlinear remainder, with actual-direction curvature providing an oracle reference for the proposed tube construction. Below, we describe the numerical procedures used for each estimate. We use $N = 10$ Euler steps, with a generated action chunk of shape 50×32. The executed-action selection S retains the first 10 steps and 7 action coordinates. All calculations use float32 arithmetic. Below, $\boldsymbol{x}_k = (\boldsymbol{z}_k, \boldsymbol{c})$ and $\|(\boldsymbol{d}, \Delta\boldsymbol{c})\|_{\mathrm{joint}}^2 = \|\boldsymbol{d}\|_F^2 + \|\Delta\boldsymbol{c}\|_F^2$. Hessian directional products are evaluated using automatic differentiation without forming the full Hessian.

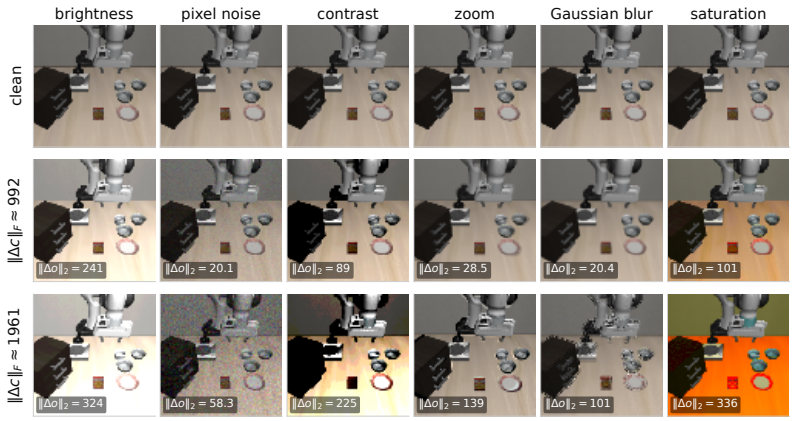

Figure 6: Ablation image perturbations used for analysis for the Libero-Spatial scene. The corresponding wrist-view perturbations are shown in Figure 5. The same $\|\Delta c\|_F$ can be produced by vastly different ranges of image perturbations.

Actual-direction curvature estimate. As an oracle comparison, we evaluate curvature along the actual joint displacement. For each shared initial-noise draw, we run the clean and perturbed samplers to obtain $\boldsymbol{\delta}_k = \boldsymbol{z}'_k - \boldsymbol{z}_k$ and define $\Delta \boldsymbol{x}_k = (\boldsymbol{\delta}_k, \Delta \boldsymbol{c})$ and $\boldsymbol{v}_k = \Delta \boldsymbol{x}_k / \|\Delta \boldsymbol{x}_k\|_{\text{joint}}$. For nonzero $\Delta \boldsymbol{x}_k$, the corresponding directional curvature is

$$M_{k,\boldsymbol{v}_k} = \sup_{s \in [0,1]} \left\| \nabla^2 u^{\theta}_{t_k}(\boldsymbol{x}_k + s\Delta \boldsymbol{x}_k)[\boldsymbol{v}_k, \boldsymbol{v}_k] \right\|_F . \tag{33}$$

We approximate this supremum using $J = 7$ equally spaced points, $\mathcal{S}_{\text{dir}} = \{j/(J-1) : j = 0, \ldots, J-1\}$, giving

$$\widehat{M}_{k,\boldsymbol{v}_k} = \max_{s \in \mathcal{S}_{\text{dir}}} \left\| \nabla^2 u^{\theta}_{t_k}(\boldsymbol{x}_k + s\Delta \boldsymbol{x}_k)[\boldsymbol{v}_k, \boldsymbol{v}_k] \right\|_F . \tag{34}$$

The oracle calculation uses 16 shared initial-noise draws. Since $\boldsymbol{v}_k$ is normalized, the magnitude-weighted curvature estimate used in the remainder calculation is

$$\widehat{\overline{M}}^{\text{oracle}}_k = \widehat{M}_{k,\boldsymbol{v}_k} \|\Delta \boldsymbol{x}_k\|^2_{\text{joint}}. \tag{35}$$

For $\Delta \boldsymbol{x}_k = \boldsymbol{0}$, we set this quantity to zero.

Tube curvature estimate (Theorem 1). Our tube calculation uses the measured context change $\Delta \boldsymbol{c}$ and the first-order predictions $\hat{\boldsymbol{\delta}}_k$ along the clean trajectory. It does not require the actual state displacements.

Let $\widehat{r}_k$ denote the numerically propagated prediction-error radius. For each interpolation point $s \in \mathcal{S}_{\text{tube}}$, let $\mathcal{D}_{k,s}$ be a finite set of candidate state displacements satisfying $\|\boldsymbol{d} - \hat{\boldsymbol{\delta}}_k\|_F \le \widehat{r}_k$. The main perturbation study and the curvature ablation use different numerical searches over these admissible displacements, as specified below. We estimate the tube curvature by

$$\widehat{\overline{M}}_k = \max_{\substack{s \in \mathcal{S}_{\text{tube}} \\ \boldsymbol{d} \in \mathcal{D}_{k,s}}} \left\| \nabla^2 u^{\theta}_{t_k}\big(\boldsymbol{x}_k + s(\boldsymbol{d}, \Delta \boldsymbol{c})\big)[(\boldsymbol{d}, \Delta \boldsymbol{c}), (\boldsymbol{d}, \Delta \boldsymbol{c})] \right\|_F . \tag{36}$$

We use equally spaced interpolation grids $\mathcal{S}_{\text{tube}} = \{j/(J-1) : j = 0, \ldots, J-1\}$, with $J = 16$ for the Spatial observation and Cream Cheese observation.

Starting from $\widehat{r}_0 = 0$, we propagate

$$\widehat{r}_{k+1} = \widehat{\|\boldsymbol{J}^z_k\|}_{\text{op}} \widehat{r}_k + \frac{h}{2} \widehat{\overline{M}}_k, \tag{37}$$

where the operator norm is estimated by power iteration. The flow steps are processed sequentially for each initial-noise draw.

The operator norm is estimated by power iteration. The finite candidate search can underestimate the curvature supremum, and finite power iteration does not certify an upper bound on the operator norm. Because these estimates enter the radius recursion, the numerical tubes need not contain the true displacements. We therefore report numerical estimates of the theoretical construction, rather than certified upper bounds.

Unrestricted state-direction curvature. This calculation retains the measured context displacement $\Delta \boldsymbol{c}$ but discards the direction of the state displacement. We use $\|\hat{\boldsymbol{\delta}}_k\|_F + \widehat{r}_k$ as the numerical state-displacement envelope. Expanding the joint Hessian into context, mixed, and state contributions gives the following numerical curvature calculation, where the suprema are approximated by the searches described below:

$$\begin{aligned}\widehat{\overline{M}}_k^{\text{state}} \approx & \max_{s\in\mathcal{S}_{\text{dir}}} \left\|\nabla^2 u_{t_k}^{\theta}(\boldsymbol{z}_k \mid \boldsymbol{c}+s\Delta\boldsymbol{c})[(\boldsymbol{0},\Delta\boldsymbol{c}),(\boldsymbol{0},\Delta\boldsymbol{c})]\right\|_F \\ & + 2\big(\|\hat{\boldsymbol{\delta}}_k\|_F + \widehat{r}_k\big) \sup_{\|\boldsymbol{d}\|_F=1} \left\|\nabla^2 u_{t_k}^{\theta}(\boldsymbol{x}_k)[(\boldsymbol{d},\boldsymbol{0}),(\boldsymbol{0},\Delta\boldsymbol{c})]\right\|_F \\ & + \big(\|\hat{\boldsymbol{\delta}}_k\|_F + \widehat{r}_k\big)^2 \sup_{\|\boldsymbol{d}\|_F=1} \left\|\nabla^2 u_{t_k}^{\theta}(\boldsymbol{x}_k)[(\boldsymbol{d},\boldsymbol{0}),(\boldsymbol{d},\boldsymbol{0})]\right\|_F . \end{aligned} \tag{38}$$

Unrestricted joint-direction curvature. This calculation also discards the measured context direction. We approximate the magnitude-weighted curvature by searching over unit joint state-context directions:

$$\widehat{\overline{M}}_k^{\text{joint}} \approx \Big[\big(\|\hat{\boldsymbol{\delta}}_k\|_F + \widehat{r}_k\big)^2 + \|\Delta\boldsymbol{c}\|_F^2\Big] \sup_{\|\boldsymbol{v}\|_{\text{joint}}=1} \left\|\nabla^2 u_{t_k}^{\theta}(\boldsymbol{x}_k)[\boldsymbol{v},\boldsymbol{v}]\right\|_F . \tag{39}$$

At each flow step, we retain the largest unit-direction curvature encountered across the search and 16 initial-noise draws, and take its maximum with the unit state-direction curvature estimate from the preceding calculation. The search is evaluated at clean state-context pairs and is not guaranteed to find the largest directional curvature. It therefore provides a numerical comparison with unrestricted directions, rather than a certified bound over the tubes. Its large values illustrate the conservatism of this particular calculation.

The first-order term is estimated using 128 initial-noise draws, whereas empirical Wasserstein-2 distances use optimal assignment between 512 clean and perturbed samples generated with shared noise. These different sample counts can produce small discrepancies between the numerical bound estimate and the empirical distance, particularly when the curvature contribution is extremely small.

Computational cost. On an RTX 5090 GPU, a brightness perturbation at $\|\Delta\boldsymbol{c}\|_F = 250$ takes approximately 44 seconds to evaluate with a five-point interpolation grid and five state-displacement candidates.

A.3 Numerical Calculation of Probabilistic Bound and Additional Results

We define A_τ using the negative log of a Gaussian KDE fitted to clean principal-component scores. Figure 4 uses five components explaining 64.3% of clean variance, while Figure 7 uses three explaining 51.3%. Region membership depends only on these projected scores. We use separate sets of 2048 clean samples for fitting, threshold calibration to 95% coverage, and evaluation; perturbed evaluation samples share the clean samples' initial noise.

We use margin lower bounds $\underline{d}_i \le d(\boldsymbol{a}_i, A_\tau^c)$, obtained from a KDE curvature bound for the five-component region and a grid distance transform with interpolation correction for the three-component region. The plug-in probability estimate is

$$\widehat{L} = \max_{\lambda\ge 0}\left\{\frac{1}{M}\sum_{i=1}^{M}\min\{1,\lambda\underline{d}_i^2\} - \lambda\widehat{B}^2\right\}, \qquad M = 2048. \tag{40}$$

For Corollary 1, we use margins at $\boldsymbol{a}_i + \widehat{m}$ and the residual budget $\widehat{b}$. The measured-distance comparison recomputes optimal assignment after subtracting $\widehat{m}$ from perturbed samples. All reported probability bounds are numerical estimates without population-level confidence guarantees.

Figure 7 shows how Corollary 1 can improve the probability estimate by accounting for the predicted translation. The following results use the same global image perturbations applied to both cameras as the curvature ablation. For the comparison labeled "Measured W_2 + Corollary 1," we subtract $\widehat{m}$ from the perturbed action samples and recompute optimal assignment to obtain $W_2(\widehat{P}, \widehat{Q} - \widehat{m})$. This empirical residual distance replaces $\widehat{b}$ in the probability calculation. The comparison assesses the conservatism of the estimated residual budget, while still retaining the effects of the chosen region, margin lower bounds, and finite sampling.

As context magnitude increases, the probability estimates become more conservative. At $\|\Delta \boldsymbol{c}\|_F \approx 254$, translation-adjusted estimates remain positive for zoom, brightness, contrast, and saturation, whereas the corresponding unshifted estimates are zero or nearly zero. For Gaussian blur and pixel noise, both estimates are zero despite observed coverage near 95%. Thus, an uninformative probability estimate does not imply that typical clean actions are no longer retained. It indicates that the estimated distribution-shift budget and region margins no longer provide useful control of that probability.

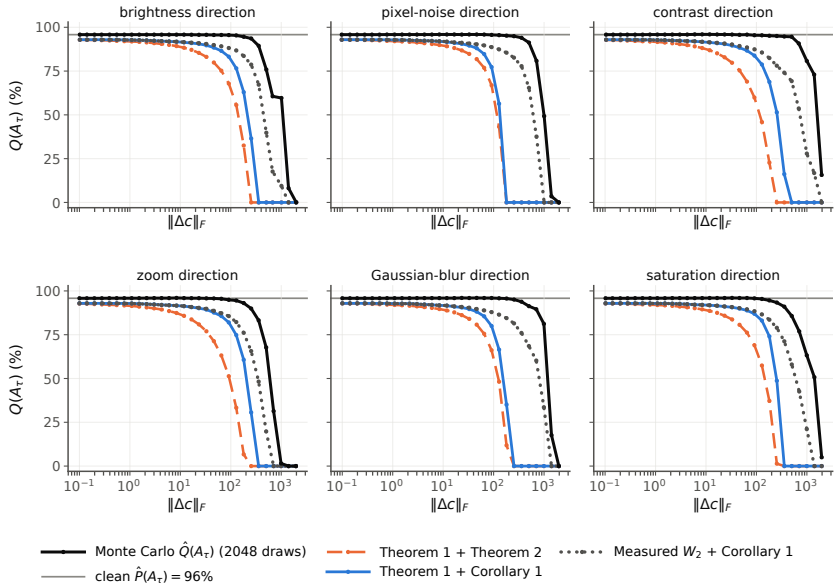

Figure 7: Probabilistic lower bound of $Q(A_{0.95})$. Corollary 1 empirically gives a tighter lower lower bound on $Q(A_{0.95})$ compared to theorem 2. Using the measured W_2 gives a much more accurate estimate after large changes in context.

Observation identifiers. We use `spatial_t0_step20` and `object_t1_step10`, where the identifier specifies the LIBERO suite, zero-based task index, and number of environment steps in the policy's rollout. Both rollouts start from LIBERO initial state 0 with reset seed 1000. We save the MuJoCo state and instruction at the selected step and re-render both camera views from that state.

A.4 Details on Flow Matching

Algorithm details for flow matching (Lipman et al., 2023) (supplemental to preliminaries in the main text) are provided in Algorithms 1 and 2. Algorithm 1 outlines the training procedure for flow matching. Note that for VLAs with flow matching heads such as $\pi_{0.5}$, this training procedure is integrated with the VLM. Additionally, $\pi_{0.5}$ uses a non-uniform distribution for sampling training times (Intelligence et al., 2025).

Algorithm 2 goes over a simple conditional sampling procedure at inference time for a given trained flow model. A sample is drawn from an initial distribution p_0 (Standard Gaussian in our work) and is iteratively updated with the Euler approximation of the flow.

Algorithm 1 Conditional Flow Matching Training with a Linear Probability Path (Lipman et al., 2023)

Require: Dataset $(x, c) \sim p_{\text{data}}$, reference distribution p_0, velocity field u_t^θ

1: **for** each mini-batch **do**
2: Sample $(x, c) \sim p_{\text{data}}$
3: Sample $z_0 \sim p_0$
4: Sample $t \sim \mathcal{U}[0, 1]$
5: Set $z_t = (1 - t)z_0 + tx$
6: Set target velocity $u_t^{\text{target}} = x - z_0$
7: Take a gradient step on $\|u_t^\theta(z_t \mid c) - u_t^{\text{target}}\|_2^2$
8: **end for**

Algorithm 2 Euler Sampling from a Conditional Flow Model

Require: Velocity field u_t^θ, conditioning context c, number of Euler steps N, reference distribution p_0

Sample $z_0 \sim p_0$
Set $h = \frac{1}{N}$
for $k = 0, \ldots, N - 1$ **do**
Set $t_k = \frac{k}{N}$
Update $z_{k+1} = z_k + h\, u_{t_k}^\theta(z_k \mid c)$
end for
return z_N

A.5 Accounting for Context Approximation in the Bound

The main bound assumes that the context change $\Delta \boldsymbol{c}$ is measured by evaluating the VLM on both clean and perturbed observations. Here we extend the analysis to bound action-distribution shifts directly from a specified visual perturbation, without requiring the measured context change. We use the first-order context prediction $\widehat{\Delta \boldsymbol{c}} = \boldsymbol{J}_{\text{VLM}}(\boldsymbol{o})\Delta \boldsymbol{o}$, where $\Delta \boldsymbol{o} = \boldsymbol{o}' - \boldsymbol{o}$ changes only the visual input. When $\Delta \boldsymbol{o} = \boldsymbol{0}$, the paired trajectories coincide and the distribution shift is zero. We therefore consider $\Delta \boldsymbol{o} \neq \boldsymbol{0}$ below.

Define

$$\boldsymbol{v}_o = \frac{\Delta \boldsymbol{o}}{\|\Delta \boldsymbol{o}\|_F}, \qquad M_{\boldsymbol{v}_o}^{\text{VLM}} = \sup_{s \in [0,1]} \left\| \nabla^2 VLM(\boldsymbol{o} + s\Delta \boldsymbol{o})[\boldsymbol{v}_o, \boldsymbol{v}_o] \right\|_F. \tag{41}$$

Assume the backbone is twice continuously differentiable on a neighborhood of this segment and $M_{\boldsymbol{v}_o}^{\text{VLM}}$ is finite. The integral Taylor remainder gives

$$\|\Delta \boldsymbol{c} - \widehat{\Delta \boldsymbol{c}}\|_F \leq \frac{1}{2} M_{\boldsymbol{v}_o}^{\text{VLM}} \|\Delta \boldsymbol{o}\|_F^2. \tag{42}$$

The first-order state prediction now uses the predicted context change,

$$\hat{\boldsymbol{\delta}}_0 = \boldsymbol{0}, \qquad \hat{\boldsymbol{\delta}}_{k+1} = \boldsymbol{J}_k^z \hat{\boldsymbol{\delta}}_k + \boldsymbol{J}_k^c \widehat{\Delta \boldsymbol{c}}. \tag{43}$$

Unlike the main construction, the curvature tube must account for uncertainty in both the state displacement and the context change. We use $\overline{M}_k$ for the corresponding magnitude-weighted curvature supremum over segments generated by joint displacements $(\boldsymbol{d}, \boldsymbol{\eta})$ satisfying $\|\boldsymbol{d} - \hat{\boldsymbol{\delta}}_k\|_F \leq r_k$ and $\|\boldsymbol{\eta} - \widehat{\Delta \boldsymbol{c}}\|_F \leq \frac{1}{2} M_{\boldsymbol{v}_o}^{\text{VLM}} \|\Delta \boldsymbol{o}\|_F^2$. Its definition otherwise follows Equation 14, with the candidate context displacement also varied in the supremum. At interpolation parameter s, the state and context radii are $s r_k$ and $\frac{s}{2} M_{\boldsymbol{v}_o}^{\text{VLM}} \|\Delta \boldsymbol{o}\|_F^2$, respectively.

Starting from $r_0 = 0$, propagate the prediction-error radius as

$$r_{k+1} = \|\boldsymbol{J}_k^z\|_{\text{op}} r_k + \frac{1}{2} \|\boldsymbol{J}_k^c\|_{\text{op}} M_{\boldsymbol{v}_o}^{\text{VLM}} \|\Delta \boldsymbol{o}\|_F^2 + \frac{h}{2} \overline{M}_k. \tag{44}$$

Subtracting Equation 43 from the exact displacement recursion introduces the additional term $\boldsymbol{J}_k^c(\Delta \boldsymbol{c} - \widehat{\Delta \boldsymbol{c}})$. The same induction as in the main proof therefore gives $\|\boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k\|_F \leq r_k$ and $\|\boldsymbol{e}_k\|_F \leq \frac{h}{2} \overline{M}_k$.

Under the same smoothness and moment assumptions as Theorem 1, and assuming the expectations below are finite, the shared-noise coupling yields

$$W_2(P_k, Q_k) \le \left(\mathbb{E}[\|\hat{\boldsymbol{\delta}}_k\|_F^2]\right)^{1/2} + \frac{1}{2} M_{\boldsymbol{v}_o}^{\mathrm{VLM}} \|\Delta \boldsymbol{o}\|_F^2 \left(\mathbb{E}\left[\left\| \sum_{\ell=0}^{k-1} \boldsymbol{T}_{k,\ell+1} \boldsymbol{J}_\ell^c \right\|_{\mathrm{op}}^2 \right] \right)^{1/2} + \frac{h}{2} \left(\mathbb{E}\left[\left(\sum_{\ell=0}^{k-1} \|\boldsymbol{T}_{k,\ell+1}\|_{\mathrm{op}} \overline{M}_\ell \right)^2 \right] \right)^{1/2}. \tag{45}$$

The expectation is over the shared initial noise. The context approximation error is the same at every flow step, allowing its propagation operators to be summed before taking the operator norm in the second term.

The three terms capture the first-order observation response, propagated VLM approximation error, and accumulated nonlinear flow remainder. For executed actions, use $S\hat{\boldsymbol{\delta}}_N$ in the first term and replace $\boldsymbol{T}_{N,\ell+1}$ by $S\boldsymbol{T}_{N,\ell+1}$ in the other two terms; the radius recursion remains unchanged.

This extension requires controlling both the VLM curvature and flow curvature over a tube with uncertain context. It introduces an additional context-approximation term and requires controlling flow curvature over candidate context displacements as well as state displacements. These additional uncertainties can make the bound more conservative than when the context change is measured directly, while the size of the context tensor (> 5 million for $\pi_{0.5}$) makes the curvature calculation computationally demanding.

A.6 TABULAR DATA FOR MAIN PAPER PLOTS

Table 2: Empirical Wasserstein-2 distances and numerical Theorem 1 estimates at all 10 evaluated context magnitudes for both observations.

		$\|\Delta c\|_F$									
Action Shift	**Metric**	1	5	10	25	50	100	150	250	375	500
LIBERO Spatial: Black Bowl											
Brightness	W_2	0.00036	0.0018	0.0036	0.0091	0.018	0.038	0.062	0.13	0.21	0.41
	Thm. 1	**0.00036**	**0.0018**	**0.0037**	**0.0095**	**0.020**	**0.045**	**0.083**	**0.21**	**0.48**	**0.57**
Contrast	W_2	0.00047	0.0024	0.0047	0.012	0.024	0.051	0.085	0.13	0.16	0.20
	Thm. 1	**0.00047**	**0.0024**	**0.0048**	**0.012**	**0.026**	**0.058**	**0.10**	**0.20**	**0.37**	**0.69**
Patch noise	W_2	0.0011	0.0049	0.0084	0.013	0.015	0.041	0.093	0.25	0.71	1.2
	Thm. 1	**0.0011**	**0.0049**	**0.0085**	**0.013**	**0.016**	**0.044**	**0.10**	**0.27**	**0.82**	**1.6**
Patch blur	W_2	0.00068	0.0034	0.0067	0.016	0.032	0.092	0.15	0.29	0.53	0.84
	Thm. 1	**0.00068**	**0.0034**	**0.0068**	**0.017**	**0.034**	**0.098**	**0.16**	**0.35**	**0.78**	**1.4**
Patch saturation	W_2	0.00081	0.0038	0.0069	0.013	0.018	0.042	0.083	0.18	0.33	0.55
	Thm. 1	**0.00081**	**0.0038**	**0.0071**	**0.014**	**0.024**	**0.064**	**0.14**	**0.42**	**1.1**	**2.9**
Patch warp	W_2	0.00083	0.0043	0.0088	0.023	0.042	0.071	0.11	0.19	0.33	0.50
	Thm. 1	**0.00083**	**0.0043**	**0.0089**	**0.023**	**0.045**	**0.080**	**0.12**	**0.26**	**0.67**	**2.2**
LIBERO Object: Cream Cheese											
Brightness	W_2	0.00088	0.0044	0.0088	0.022	0.044	0.081	0.10	0.050	0.22	0.42
	Thm. 1	**0.00088**	**0.0044**	**0.0090**	**0.024**	**0.051**	**0.11**	**0.15**	**0.063**	**0.40**	**1.6**
Contrast	W_2	0.0011	0.0054	0.011	0.028	0.056	0.12	0.17	0.29	0.43	0.54
	Thm. 1	**0.0011**	**0.0055**	**0.011**	**0.030**	**0.066**	**0.17**	**0.32**	**0.88**	**2.4**	**6.0**
Patch noise	W_2	0.00064	0.0031	0.0058	0.010	0.013	0.020	0.033	0.076	0.17	0.25
	Thm. 1	**0.00064**	**0.0032**	**0.0060**	**0.012**	**0.016**	**0.029**	**0.053**	**0.16**	**0.51**	**1.1**
Patch blur	W_2	0.00021	0.0011	0.0021	0.0058	0.013	0.023	0.022	0.089	0.24	0.35
	Thm. 1	**0.00021**	**0.0011**	**0.0022**	**0.0061**	**0.014**	**0.029**	**0.035**	**0.14**	**0.56**	**0.94**
Patch saturation	W_2	0.00073	0.0037	0.0074	0.018	0.036	0.074	0.096	0.11	0.12	0.12
	Thm. 1	**0.00073**	**0.0037**	**0.0075**	**0.019**	**0.039**	**0.090**	**0.12**	**0.17**	**0.24**	**79**
Patch warp	W_2	0.00034	0.0017	0.0034	0.0082	0.015	0.028	0.038	0.070	0.12	0.18
	Thm. 1	**0.00034**	**0.0017**	**0.0035**	**0.0089**	**0.019**	**0.050**	**0.21**	**9.5**	**260**	**120**

Table 3: Selected context magnitudes from the probabilistic bound comparison in A.3.

Perturbation	$\|\Delta c\|_F$	$\hat{Q}(A_\tau)$	Thm. 1 + Thm. 2	Thm. 1 + Cor. 1	Meas. W_2 + Cor. 1
Matched context magnitude ≈ 91.3					
Zoom	91.3	94.9%	51.3%	82.0%	85.6%
Contrast	91.3	95.5%	59.8%	83.8%	87.4%
Brightness	91.4	95.6%	67.8%	83.2%	88.1%
Saturation	91.3	95.8%	68.9%	86.9%	88.4%
Gaussian blur	91.2	95.9%	66.1%	79.9%	87.9%
Pixel noise	91.3	95.7%	66.8%	77.2%	88.6%
Matched context magnitude ≈ 180					
Zoom	180	93.1%	6.6%	60.6%	76.0%
Contrast	181	95.2%	22.7%	68.7%	82.4%
Brightness	180	94.6%	32.4%	62.9%	83.6%
Saturation	180	94.6%	37.1%	74.0%	83.5%
Gaussian blur	180	95.7%	11.8%	35.1%	84.4%
Pixel noise	180	95.2%	0.0%	0.0%	85.4%
Matched context magnitude ≈ 254					
Zoom	253	89.9%	0.0%	30.5%	65.6%
Contrast	253	95.0%	0.0%	51.4%	78.0%
Brightness	254	93.6%	0.0%	36.4%	78.6%
Saturation	254	93.4%	1.3%	48.7%	78.9%
Gaussian blur	254	95.1%	0.0%	0.0%	81.4%
Pixel noise	253	94.9%	0.0%	0.0%	82.6%