FLOW MATCHING IS NOT A CONTRACTIVE CLASS FOR VISION–LANGUAGE–ACTION POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies are increasingly employing flow matching for robot action generation. The process starts from random noise and produces an action over a few steps. It is currently believed that flow matching intrinsically contracts noise perturbations and this leads to robust VLA policies. Recent approaches often build upon this idea. We show that this belief is actually not correct. Contraction and expansion of flow in VLAs are mainly tied to the model architecture, not to the flow matching process. We establish this with a rigorous treatment of four diverse state-of-the-art VLAs. We achieve a clear split of two contractions and two expansions, which is not caused by flow matching, the direction of the flow, or the number of steps. Two policies sharing similar process design choices still land on the opposite sides, isolating the action-generation module as the primary cause. Our analysis further reveals two crucial insights. (a) Contraction in a deployed policy must be judged over the entire flow, not one step at a time, which is the current norm and can be misleading. We provide results supporting expansion at every step for a policy whose flow as a whole still contracts. (b) For flow-matching VLAs, robustness must not be inferred from the flow and needs to be measured independently because model robustness may not align with contraction. Our supporting evidence identifies a flow expanding policy as the most robust one. Our source code will be made public after acceptance.

1 INTRODUCTION

Modern Vision-Language-Action (VLA) policies increasingly generate robot actions using flow matching (Chi et al., 2025; Black et al., 2024; Bjorck et al., 2025; Kim et al., 2024; Octo Model Team et al., 2024). A flow-matching policy starts from a random noise and turns it into an action over a few steps. A small perturbation in the starting noise can contract as the action forms, or it can expand. Flow matching is widely assumed to contract it, and this contraction is the property from which these policies are argued to be robust (Davydov & Bullo, 2024). A growing line of work supports contraction and even trains for it (Lipman et al., 2022; Gao & Zhu, 2024; Tang & Zhao, 2026). The usual criterion employed is per-step contraction, forcing each step on its own not to expand (Lohmiller & Slotine, 1998). For instance, Contractive Diffusion Policies (CDP) build exactly this check into its training, penalizing any step that would expand to make the policy more robust (Abyaneh et al., 2026).

This per-step reasoning implicitly assumes that the whole flow behaves like its individual steps. We are the first to show that this assumption fails on deployed, state-of-the-art (SOTA) VLA policies. The result is counterintuitive. If every step shrinks a perturbation, the whole flow should shrink it too. It need not. In fact, a step contracts along its own directions, and those directions differ from one step to another. A perturbation that one step shrinks can be redirected into a direction the next step amplifies. Across multiple steps this phenomenon compounds, and a flow assembled from apparently contracting steps can still expand end-to-end. A per-step check is therefore not a highly useful criterion in the first place.

Given that VLA policies are now intended to drive general-purpose robots (Black et al., 2024; Bjorck et al., 2025; Kim et al., 2024), clearly understanding the contraction and expansion of the flow is crucial. For these policies, a contracting flow is taken as evidence of robustness. If that assumption is wrong, the robustness claims behind robots being deployed today rest on a false premise. This drives

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

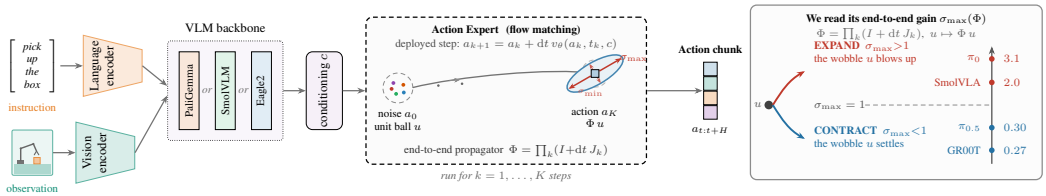


Figure 1: **Our analysis in one picture.** A flow-matching VLA turns noise a_0 into an action through its own sampler, and from that sampler we read a single measurement. Linearizing the flow gives a propagator Φ whose end-to-end gain $\sigma_{\max}(\Phi)$ is above 1 when the flow expands a perturbation and below 1 when it contracts one. Applied identically to four SOTA VLAs, this witness splits them cleanly into two expanders and two contractors (Sec. 5).

our rigorous analysis of SOTA VLAs. The question is whether a deployed policy's flow contracts a small perturbation, or amplifies it.

We measure it. We read a single number off each policy's own sampler, the worst-case factor by which it grows or shrinks a perturbation from noise to action (Fig. 1), and check it against a direct numerical estimate of the same quantity. We find a clean split across four SOTA VLAs on LIBERO (Liu et al., 2023) (Fig. 2a). SmolVLA (Shukor et al., 2025) and π_0 (Black et al., 2024) expand, while GR00T-N1.5 (Bjorck et al., 2025) and $\pi_{0.5}$ (Physical Intelligence et al., 2025) contract. In contrast to the common belief, flow matching does not define a contractive class.

The split survives the obvious framework-level explanations, denoising direction, training regime, and step count (Tab. 5). Most sharply, π_0 and $\pi_{0.5}$ share the same backbone, denoising direction, and step count yet land on opposite sides, which localizes the cause to the action-generation head, though not to a single design choice since the two differ in more than one.

We see this concretely in the per-step check itself. On $\pi_{0.5}$ it reports expansion at every step, yet the flow contracts end to end, a disagreement on a real deployed policy that is exactly why the whole flow must be measured rather than judged step by step (Sec. 3, App. C).

We also separate two easily conflated questions. Whether the flow contracts the noise it starts from, which is what we measure, differs from whether the policy's action is robust to changes in what its camera sees, since these perturb different inputs. We find they do not track each other, as the most observation-robust policy in our panel is one whose flow expands (Tab. 2). For a deployed VLA, robustness cannot be read off the flow.

Contributions. We establish that contraction in a deployed flow-matching VLA is a property of its architecture rather than of flow matching, and that it must therefore be measured rather than assumed.

- **A cross-architecture falsification.** One validated witness, applied identically to four SOTA VLAs on LIBERO, splits them into two expanders and two contractors. The split holds under every framework-level control, and a within-family pair localizes the cause to the action-generation head (Sec. 5; Fig. 2a, Tab. 5).
- **A methodological correction.** We demonstrate that the one-sided, per-step criterion the field certifies and trains against is neither necessary nor sufficient for end-to-end contraction, and we exhibit a deployed policy, $\pi_{0.5}$, on which the per-step and end-to-end verdicts disagree (Tab. 1).
- **A decoupling of geometry from robustness.** We find that contraction of the generation flow does not predict a policy's robustness to its observations. The most observation-robust model in our panel is itself an expander (Tab. 2), a result we report as an open-loop floor rather than closed-loop success.
- **A constructive tradeoff-null.** On both architectures, training against the two-sided criterion is the only way to contract the deployed flow, while the one-sided penalty fails. Forcing contraction, however, reduces how much the policy responds to its observation, so it is not a free robustness win (App. F).

2 Related Work

Prior work on VLA action generation falls into three lines, namely generative action policies based on diffusion and flow matching, the VLAs that instantiate these methods at scale, and contraction-based approaches that treat a bounded generation flow as a robustness objective. We position our study against all three.

Generative action policies. Modern robot policies increasingly model actions with an iterative generative sampler rather than a single feed-forward map. Diffusion policies (Chi et al., 2025; Song et al., 2020; Ho et al., 2020; Song et al., 2021) and flow-matching approaches (Lipman et al., 2022; Liu et al., 2022) transform noise into an action chunk through a learned score or velocity field. This formulation now underlies many of the strongest robot policies (Black et al., 2024; Physical Intelligence et al., 2025; Bjorck et al., 2025; Shukor et al., 2025).

Existing analyses, however, primarily characterise the *continuous-time* field or its training objective, not the gain of the *discrete, few-step sampler that is actually deployed*. That sampler is a composition of refinement steps whose geometry need not follow from the framework used to train it, so whether the resulting noise-to-action map expands or contracts remains an empirical question.

Vision-language-action models. Modern VLAs combine a pretrained vision-language backbone with an increasingly diverse set of action-generation mechanisms. OpenVLA uses autoregressive action-token prediction (Kim et al., 2024). π_0 integrates a flow field through a gemma action expert (Black et al., 2024), while $\pi_{0.5}$ modifies this design through adaptive layer-norm (adaLN) conditioning and state-in-prompt (Physical Intelligence et al., 2025). SmolVLA uses a lightweight flow-matching expert (Shukor et al., 2025), and GR00T-N1.5 uses a diffusion-transformer (DiT) action sampler (Bjorck et al., 2025).

Flow-matching VLAs share a generative framework while differing substantially in action-head architecture, denoising direction, and sampling depth. Any property attributed to "flow matching" must therefore survive substituting one head for another. Yet the end-to-end generation geometry of these deployed policies has not been measured on a common footing.

Contraction as an engineered robustness property. A separate line of work treats contraction not as an emergent property but as something to certify or enforce, on the premise that a contracting sampler is a more robust one. This premise is inherited from two mature literatures, classical contraction analysis in control (Lohmiller & Slotine, 1998; Davydov & Bullo, 2024) and convergence analysis of diffusion and flow samplers (Gao & Zhu, 2024; Tang & Zhao, 2026), both of which reason through a one-sided, per-step quantity derived from the symmetric part of the step Jacobian. Contractive Diffusion Policies (Abyaneh et al., 2026) carry this premise into action generation, training a policy to contract at every step precisely to make it more robust. Whether the premise holds for deployed VLA policies, however, has not been tested, and it is the question we take up.

Such a criterion characterises individual steps, whereas the deployed sampler is their finite-time composition, and for non-normal dynamics the two need not agree. A per-step criterion therefore does not by itself determine whether the complete noise-to-action map contracts end to end.

Contraction of the generation flow should also be distinguished from robustness of the policy to its observations. Existing VLA robustness evaluations perturb visual or multimodal inputs and measure the resulting policy behaviour (Guo et al., 2026). That probes observation sensitivity, which we denote J_c. Generation-flow contraction instead concerns how perturbations in the action noise propagate through the sampler, which we denote J_a. These are different Jacobians and different robustness questions, and we define both concretely in §3.1.

Our position. We measure whether the deployed flow actually contracts, rather than assuming it. We study the two-sided, end-to-end gain of each policy's deployed noise-to-action map, validated independently with finite differences, on a common footing across four flow-matching VLAs (§5.1, §5.2). To our knowledge, no prior work replaces the field's one-sided per-step contraction object with an end-to-end two-sided audit of deployed VLA samplers and exhibits a policy on which the two disagree.

3 THE CONTRACTION-GEOMETRY WITNESS

A flow-matching (FM) vision-language-action policy generates an action chunk by integrating a learned velocity field along a denoising schedule. It is this generation flow whose contraction geometry we measure. Let c be the frozen conditioning the backbone produces from the observation, and let $v_\theta(a, t, c)$ be the velocity field over the action latent a at flow time t. The latent stacks the emitted action chunk, with d_a=32 padded dimensions per step of which 7 are the real robot degrees of freedom (DoF). The policy deploys a fixed velocity-Euler sampler that runs K steps over knots $t_0, \dots, t_K$ with per-step size $\mathrm{d}t$,

$$a_{k+1} = a_k + \mathrm{d}t\, v_\theta(a_k, t_k, c), \qquad k = 0, \dots, K-1, \tag{1}$$

where a_0 is the sampled noise latent and a_K the emitted action chunk. The sign of $\mathrm{d}t$ carries the policy's native denoising orientation. It is negative for noise-to-data sampling and positive otherwise (per-policy roster in Sec. 4). We measure each policy on *its own* deployed sampler, step count, orientation, and in-distribution conditioning. The witness varies only the 7 active DoF per step and holds the padding fixed. Throughout we distinguish two routinely conflated Jacobians. J_a, the *action self-Jacobian*, governs how the flow propagates a noise-latent perturbation to the emitted action. J_c is the sensitivity of the emitted action to a perturbation of the conditioning c.

3.1 THE PROPAGATOR AND ITS DECISION RULE

The generation geometry of equation 1 is a formal problem about a single linear map, the end-to-end noise-to-action Jacobian $\partial a_K / \partial a_0$. Differentiating the recursion equation 1 yields a product of per-knot factors, the discrete *propagator*

$$\Phi \;=\; \prod_{k=K-1}^{0} \big(I + \mathrm{d}t\, J_k\big), \qquad J_k \;=\; \left.\frac{\partial v_\theta(a, t_k, c)}{\partial a}\right|_{a=a_k}, \tag{2}$$

where I is the identity on the active block and J_k is the action self-Jacobian at the k-th knot along the realized trajectory. The ordered product composes the per-step linearizations in sampling order. Thus Φ maps a perturbation of the initial noise latent onto the emitted action, to first order. It is in fact the *exact* Jacobian, not an approximation, which makes it a faithful witness.

Proposition 1 (Propagator exactness)**.** *The end-to-end Jacobian of the deployed sampler equals the propagator,* $\partial a_K / \partial a_0 = \prod_{k=K-1}^{0} (I + \mathrm{d}t\, J_k) = \Phi$.

The chain rule makes Φ the whole flow's Jacobian exactly, not an approximation; full proof in App. A.

The two-sided decision rule. We report the operator norm of the propagator,

$$\sigma_{\max}(\Phi) \;=\; \|\Phi\|_2 \;=\; \max_{\|u\|=1} \|\Phi u\|, \qquad \begin{cases} \sigma_{\max}(\Phi) > 1 & \Rightarrow \text{ the flow } \textit{expands} \text{ some direction,} \\ \sigma_{\max}(\Phi) < 1 & \Rightarrow \text{ the flow } \textit{contracts} \text{ every direction,} \end{cases} \tag{3}$$

where $\sigma_{\max}(\Phi)$ is the largest singular value of Φ, u ranges over unit perturbations of the active block, and the contraction branch is scoped to the measured active 7-DoF subspace (§3.3). This is the true worst-case gain. It is a two-sided quantity that certifies expansion *and* its absence, so one scalar decides the sign without a one-sided surrogate.

Proposition 2 (Two-sided decision rule)**.** $\sigma_{\max}(\Phi) > 1$ *iff the linearized flow strictly expands some unit perturbation, and* $\sigma_{\max}(\Phi) < 1$ *iff it strictly contracts every unit perturbation.*

Because $\sigma_{\max}$ is the largest gain any unit direction receives, testing it against 1 decides the sign; full proof in App. A.

Scope: J_a**, not** J_c**.** The witness equation 3 is a statement about J_a, the very property robustness arguments for flow and diffusion policies implicitly assume (Tang & Zhao, 2026; Gao & Zhu, 2024; Davydov & Bullo, 2024). It is not about J_c, the robustness of the emitted action to *observation* perturbations, which this instrument does not measure. The $J_a \to J_c$ gap is the honest floor of what equation 3 earns.

Algorithm 1 Contraction-geometry witness $\sigma_{\max}(\Phi)$

Require: deployed velocity field v_θ, step size $\mathrm{d}t$, step count K, frozen conditioning c, noise latent a_0

$\Phi \leftarrow I$ ▷ active-block propagator
for $k = 0, \ldots, K-1$ **do**
  $J_k \leftarrow \partial v_\theta(a, t_k, c)/\partial a|_{a=a_k}$ ▷ per-knot velocity Jacobian, Eq. 2
  $\Phi \leftarrow (I + \mathrm{d}t\, J_k)\, \Phi$ ▷ accumulate in sampling order, Eq. 2
  $a_{k+1} \leftarrow a_k + \mathrm{d}t\, v_\theta(a_k, t_k, c)$ ▷ advance the deployed sampler, Eq. 1
end for
if $\varepsilon_{\mathrm{fd}} > 10^{-2}$ **then** ▷ finite-difference gate, Eq. 4
  return $\perp$ ▷ REFUSE: untrusted or degenerate map
end if
$\sigma \leftarrow \sigma_{\max}(\Phi)$ ▷ operator norm, Eq. 3
verdict $\leftarrow$ EXPAND **if** $\sigma > 1$ **else** CONTRACT

Ensure: gain σ and EXPAND/CONTRACT verdict, after the self-test (Eq. 5), predictiveness (Eq. 6), and non-degeneracy gates

3.2 THE INCUMBENT ONE-SIDED PER-STEP OBJECT, AND WHY TWO-SIDED IS CORRECT

Prior contraction analyses of diffusion and flow samplers do not measure $\sigma_{\max}(\Phi)$ directly. They certify contraction from a *per-step, one-sided* object. It requires the largest symmetric-part eigenvalue of each J_k to be nonpositive at every knot, $\lambda_{\max}(\mathrm{sym}\, J_k) \leq 0$. Equivalently, the per-step log-contraction abscissa $\eta_{\mathrm{loc}} = -\lambda_{\max}(\mathrm{sym}\, J_k)$ is nonnegative, and its integral $\int \eta_{\mathrm{loc}}\, \mathrm{d}t \geq 0$ is read as guaranteeing a contracting flow. A *negative* abscissa flags an expanding one. The Contractive Diffusion Policies class (Abyaneh et al., 2026) *engineers* a criterion of this type into training. But this is a one-sided bound on a non-normal, finite-time composition. It certifies the wrong-sided quantity, in general neither necessary nor sufficient for end-to-end contraction. The insufficiency is provable.

Proposition 3 (The one-sided criterion is not sufficient)**.** *There exist step factors, each with nonpositive symmetric-part spectral abscissa (per-step "contracting" under the one-sided criterion, $\eta_{\mathrm{loc}} \geq 0$ at every knot), whose product Φ nonetheless has $\sigma_{\max}(\Phi) > 1$.*

Two non-normal factors that each contract along their own axes can, when composed, hand a perturbation between axes and amplify it. Passing the one-sided test at every step therefore does not certify the flow. The full proof is an explicit 2×2 construction (App. A). The propagator gain equation 3 instead measures the composed map directly, and gate (ii) below exhibits the same phenomenon on a known-answer 2×2 system.

3.3 THE WITNESS AND ITS VALIDATION PROTOCOL

We adopt $\sigma_{\max}(\Phi)$ equation 3 as the contraction-geometry witness. No per-model verdict is trustworthy until the instrument passes a fixed validation protocol. A witness that decides a sign must first recover a *known* sign. The protocol comprises four gates applied uniformly to every policy: (i) finite-difference agreement, (ii) a known-answer self-test with an engineered one-sided disagreement, (iii) nonlinear-predictiveness, and (iv) a non-degeneracy gate for the surprising contractor. Algorithm 1 outlines the end-to-end measurement, from rolling out a policy's own sampler to reading the verdict and refusing on a failed gate. Measured pass values are in App. C, Tab. 3.

Applicability. The witness is well-posed only for velocity-Euler samplers of the form equation 1, whose per-knot J_k and hence Φ are defined. It is *not* applied to predict-action or one-shot rectified-flow samplers (Liu et al., 2022). Their emitted action is near-deterministic in the sampling noise, so $\partial a_K/\partial a_0$ collapses, and gate (i) below *refuses* such degenerate maps rather than returning a spurious number.

Gate (i): finite-difference agreement. We trust the autograd velocity Jacobian only once an independent differentiator reproduces it. At a knot we compare the autograd Jacobian-vector product

$J_k d$ against a central finite difference of the velocity along random unit directions d,

$$\varepsilon_{\mathrm{fd}} \;=\; \min_{h}\,\max_{d}\,\frac{\left\|\,\widehat{J_k d}(h) - J_k d\,\right\|}{\|J_k d\|} \;\ll\; 10^{-2}, \qquad \widehat{J_k d}(h) \;=\; \frac{v_\theta(a_k{+}hd, t_k, c) - v_\theta(a_k{-}hd, t_k, c)}{2h}, \tag{4}$$

where $\widehat{J_k d}(h)$ is the central finite-difference estimate of the Jacobian-vector product and h is the step. We report the minimum over h, the truncation-versus-cancellation U-curve. Every velocity path runs in float32, since the check is meaningless in bf16 (App. C). Since Φ is the exact product of the J_k (Prop. 1), validating them validates Φ.

Gate (ii): known-answer self-test. On a 2-D system whose closed-form gain is known we *engineer* the one-sided verdict to *disagree* with the true one and require $\sigma_{\max}(\Phi)$ to recover the known gain in both cases,

$$\underbrace{\eta_{\mathrm{loc}} < 0 \;\text{ yet }\; \sigma_{\max}(\Phi) < 1}_{\text{contract case: one-sided flags expansion}} \qquad \text{and} \qquad \underbrace{\eta_{\mathrm{loc}} > 0 \;\text{ yet }\; \sigma_{\max}(\Phi) > 1}_{\text{expand case: one-sided certifies contraction}}\,, \tag{5}$$

so the one-sided abscissa points the wrong way while $\sigma_{\max}(\Phi)$ reads the true sign. This is a known-answer instance of the same non-normal mechanism, certifying the witness reads the correct two-sided object.

Gate (iii): nonlinear-predictiveness. Because Φ only linearizes the sampler equation 1, we check that it predicts the true gain. Perturbing the noise along Φ's top singular direction u_1 and running the full sampler must reproduce the realized gain ρ,

$$\rho \;=\; \frac{\|a_K(a_0 + \epsilon u_1) - a_K(a_0)\|}{\epsilon}, \qquad u_1 = \arg\max_{\|u\|=1} \|\Phi u\|, \tag{6}$$

where ρ must match $\sigma_{\max}(\Phi)$ and stay on the same side of 1 out to $\epsilon = 0.1$. **Gate (iv)** confirms, for the least-expected contractor, that the sampler is a functioning, observation-dependent flow whose emitted action varies with the observation, ruling out a near-constant output that contracts by construction.

Scope of the contraction branch. The verdict $\sigma_{\max}(\Phi) < 1$ is scoped to the active 7-DoF subspace. Φ in equation 2 is a *product of restrictions*, active-to-active step-Jacobians with padding held fixed, so it is not by itself a two-sided bound on the full deployed flow. We corroborate every verdict against a directly measured full-dimensional deployment Jacobian (Tab. 7, App. D). The instrument is precise about what it measures, J_a on the active block and its sign relative to 1, and explicit about what it does not, a magnitude, a full-dimensional bound, or J_c.

4 Experimental Setup

We assess whether flow matching yields a contractive generation flow with one witness, the propagator gain $\sigma_{\max}(\Phi)$ of §3. We apply it identically across four deployed flow-matching (FM) vision-language-action (VLA) policies on a shared robot-manipulation benchmark. This section fixes the four objects a reader needs before any number, namely the benchmark, the policies and their *as-deployed* samplers, the protocol, and the instrument validation. Full configuration, per-suite tables, and the complete validation matrix are in the appendix.

Benchmark. All experiments are on LIBERO (Liu et al., 2023). Its four task families stress complementary axes of generalization: spatial arrangement (LIBERO-SPATIAL), object identity (LIBERO-OBJECT), task goal (LIBERO-GOAL), and long-horizon composition (LIBERO-LONG). Every model is measured on the same 4 suites $\times$ 5 tasks $\times$ 4 episodes ($n = 80$ per model). LIBERO is the only dataset with checkpoints of all four policies, so it is the common footing for a cross-architecture comparison. Each policy is measured on its own in-distribution observations and normalization, its ecologically valid operating point (App. D).

Models. We test four deployed FM-VLA policies. They are chosen to vary the architectural axes a framework-level robustness claim would have to be invariant to. **(1) SmolVLA** (Shukor et al., 2025) is a compact generalist deploying a 10-step velocity-Euler sampler at negative time step ($\mathrm{d}t < 0$).

(2) π_0 (Black et al., 2024) is a PaliGemma-plus-gemma-expert generalist, also 10 steps at $\mathrm{d}t < 0$. **(3) GR00T-N1.5-3B** is the N1.5 release of the GR00T N1 model family (Bjorck et al., 2025),[1] a per-suite specialist whose DiT action head deploys a 4-step sampler at positive time step ($\mathrm{d}t > 0$). **(4)** $\pi_{0.5}$ (Physical Intelligence et al., 2025) is a generalist sharing π_0's PaliGemma-plus-gemma-expert backbone but differing in its action head. It uses adaLN time-conditioning and folds state into the prompt rather than the velocity path, at 10 steps and $\mathrm{d}t < 0$. The selection is deliberate. $\pi_{0.5}$ and π_0 share backbone family, denoising direction, and step count, while $\pi_{0.5}$ and GR00T span the generalist–specialist and $\mathrm{d}t$-sign axes. Any split therefore reads against each candidate confound rather than confounded with it.

Compared object. The witness is contractive-analysis machinery, so the comparison is against a decision boundary, not a baseline family. We test each policy's measured $\sigma_{\max}(\Phi)$ against the threshold $\sigma_{\max} = 1$: above it the flow expands some direction, below it the flow contracts every direction of the active 7-DoF block. We report two contrasts against this boundary. First, the four models against each other by *sign* relative to 1, never raw magnitude, since estimators and grids differ. Second, within each model, the two-sided witness against the one-sided per-step object of the contractive-diffusion literature. That object is the integral of $\lambda_{\max}$ of the symmetric part, a CDP-class criterion (Abyaneh et al., 2026), and it is the methodological foil our witness corrects (§5.2). Throughout we keep the generation-flow gain (J_a) strictly distinct from output robustness to observation perturbations (J_c); the setup measures the former.

Protocol. We linearize each policy's own velocity-Euler sampler about its in-distribution operating point. We report $\sigma_{\max}(\Phi) = \|\Phi\|_2$ per episode, averaged over sampling draws. All downstream statistics aggregate these per-episode values, using the median and a cluster bootstrap over the 5 tasks per suite, pre-registered before running. The witness is restricted to the active 7-DoF LIBERO action block with padding held fixed. A contraction verdict is therefore a statement about that active subspace, corroborated separately against the full 32-dimensional deployed flow (App. D). All velocity paths run in float32, since the finite-difference check is meaningless in bf16. The verdict sign is stable across 10–40 denoising steps for the family we vary it on. Per-episode averaging, normalization handling, and estimator settings are in App. D–C.

Instrument validation. We validate the witness on the four gates of §3.3 before trusting any per-model number. The central claim is only as credible as the instrument producing the sign. All four models pass. Finite-difference agreement holds at the U-curve minimum, with per-model relative errors spanning 2.6×10^{-5} to 3.2×10^{-3}, all $\ll 10^{-2}$. The 2-D known-answer self-test recovers the true sign with the one-sided abscissa *engineered to disagree*. Nonlinear-predictiveness of the full sampler holds out to $\varepsilon = 0.1$. Non-degeneracy holds for the surprising contractor $\pi_{0.5}$. The same finite-difference gate *correctly refuses* near-deterministic predict-action samplers such as X-VLA, for which the witness is ill-posed. That refusal is the gate working as designed. The complete matrix, U-curve sweeps, and dual-estimator cross-check are in App. C, Tab. 3.

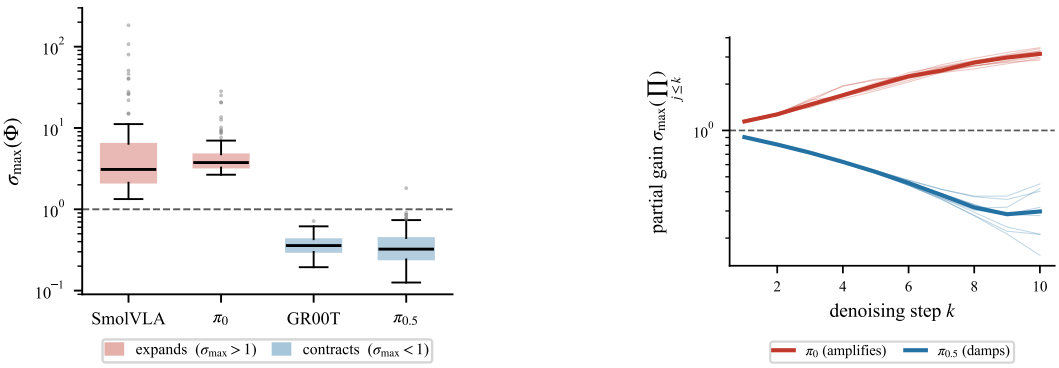
5 RESULTS

We measure the end-to-end propagator gain $\sigma_{\max}(\Phi)$ for the four deployed flow-matching VLAs of Section 4, following that section's protocol. We compare only the *sign* relative to 1, never raw magnitudes across models. A value above 1 means the generation flow expands some perturbation direction, and a value below 1 means it contracts on the measured active subspace. We read this two-sided verdict against the field's one-sided per-step criterion (the CDP-class $\lambda_{\max}(\mathrm{sym}\, J_k)$, equivalently the integral-η rule; Abyaneh et al., 2026). Throughout, $\sigma_{\max}(\Phi)$ is a generation-flow / action self-Jacobian (J_a) quantity, not output robustness (J_c).

5.1 EXPANSION VERSUS CONTRACTION IS A DESIGN PROPERTY

The four policies split cleanly around the contraction boundary (Fig. 2a). SmolVLA and π_0 expand while GR00T and $\pi_{0.5}$ contract. Policies built from the same flow-matching framework therefore

[1] We run the released GR00T-N1.5 checkpoints; Bjorck et al. (2025) describes the N1 model of the same family (there is no separate N1.5 arXiv paper).



(a) The four-model split.

(b) Its per-step mechanism.

Figure 2: **(a)** Per-episode end-to-end gain $\sigma_{\max}(\Phi)$ for four deployed VLAs under one finite-difference-validated witness. SmolVLA and π_0 expand $(80/80,\ 80/80)$; GR00T and $\pi_{0.5}$ contract $(80/80,\ 79/80)$. **(b)** Running per-step gain of a paired $\pi_0/\pi_{0.5}$ flow on identical inputs. π_0 climbs above 1 at every knot while $\pi_{0.5}$ stays below, so the mechanism is distributed composition, not one dominant step.

reach opposite end-to-end geometries. Flow matching is not a contractive class. One instrument applied identically to policies that all instantiate that framework cannot yield a framework property, so whether a deployed flow expands or contracts is a *model-design* property, and this falsifies the implicit "flow matching contracts by construction" premise.

The split does not follow the obvious framework-level factors (App. D, Tab. 5, 6). It is not the denoising direction, since $\pi_{0.5}$ contracts under the same $\mathrm{d}t < 0$ orientation for which SmolVLA and π_0 expand; not the training regime, since a specialist and a generalist both contract while two generalists expand; and not the step count, since the split persists at every matched count tested. The sharpest comparison is within the π family. π_0 and $\pi_{0.5}$ share the same PaliGemma-plus-gemma-expert backbone, denoising direction, and step count, yet fall on opposite sides of 1, localising the difference to the action-generation head. It does not isolate a single architectural cause, since the pair differs in more than one action-head choice and in its training recipe.

The result is consistent across all sixteen (model, suite) cells, each on its model's side of 1 with pre-registered conservative confidence intervals (CIs) excluding 1 (App. D, Tab. 4); a few heavy-tailed expander suites carry INCONCLUSIVE flags that we retain and detail there. The verdict is also not a frozen-padding artifact. Measured on the full 32-dimensional deployment Jacobian, $\sigma_{\max}^{\mathrm{deploy}}$ lands on the same side of 1 as the active-block witness for all four models (App. D, Tab. 7).

Figure 2b shows how the two backbone-siblings reach opposite verdicts. The difference is distributed across the composition rather than explained by a single dominant step, exactly why a per-step one-sided penalty misfires, firing on every one of the deepest contractor's knots (App. E).

5.2 THE ONE-SIDED PER-STEP CRITERION GIVES THE WRONG VERDICT

The standard one-sided per-step criterion does not reliably determine whether the deployed sampler contracts end to end (Tab. 1). It agrees for three models but labels $\pi_{0.5}$ an expander on every episode while its end-to-end flow contracts, a live counterexample on a real generalist policy rather than a constructed toy, so the one-sided abscissa cannot be *necessary* for contraction. With a constructed self-test showing it need not be *sufficient* either, the one-sided per-step object does not determine the contraction geometry of the composed sampler.

Across the measured panel its errors all run one way, in the safe direction of over-constraint (Tab. 1). That makes it the wrong *training* target as much as the wrong diagnostic. This over-warning with zero *observed* false-safes is an empirical observation on the evaluated policies, not a guarantee that false-safes cannot occur.

Table 1: **The one-sided per-step criterion certifies the wrong-sided quantity.** One-sided sign ($\int \eta$) versus the two-sided verdict ($\sigma_{\max}$), with per-step false-alarm and false-safe counts (episodes). The highlighted $\pi_{0.5}$ is called an expander while its flow contracts, so the criterion is *not necessary* for end-to-end contraction. Aggregating over 240 episodes (SmolVLA, π_0, $\pi_{0.5}$) gives 161 agree, 79 false-alarm, 0 false-safe.

Model	One-sided ($\int \eta$)	Two-sided ($\sigma_{\max}$)	Correct?	False-alarm	False-safe
SmolVLA	expander	expands	yes	0	0
π_0	expander	expands	yes	0	0
GR00T	not-expand	contracts	yes	$0/2/8/1^{\dagger}$	0
$\pi_{0.5}$	**expander**	**contracts**	**no, wrong side**	**79**	**0**

†GR00T false-alarms per LIBERO suite (n=10 each); its false-safe is 0 on all four, and GR00T is not in the 240-episode aggregate.

Table 2: Anisotropy (J_a, left) and its decoupling from an observation-sensitivity proxy (J_c, right; R_{obs} over 64 image-noise directions at σ=0.03) on `libero_long`. The J_a sign does not order R_{obs}; Rank* is 1 for the most observation-robust model.

	Anisotropy (J_a)				Decoupling (J_c proxy)	
Model	$\sigma_{\max}$ med.	g_{rnd} med.	Ratio med. [min, max]	Class	R_{obs} med. [min, max]	Rank*
π_0	3.14	1.01	3.04 [2.88, 3.40]	EXPAND	**0.0037** [0.0028, 0.0043]	**1**
$\pi_{0.5}$	0.30	0.035	9.28 [7.35, 14.89]	CONTRACT	0.022 [0.015, 0.055]	2
GR00T	0.27	0.049	5.85 [5.17, 6.20]	CONTRACT	0.029 [0.017, 0.057]	3
SmolVLA	2.05	0.276	7.86 [5.59, 18.88]	EXPAND	0.481 [0.27, 0.54]	4

5.3 Generation-flow geometry does not predict observation robustness

Contraction of the generation flow does not imply robustness to observation perturbations (Tab. 2). The J_a expand/contract classification does not order the observation-sensitivity proxy R_{obs}; most notably, π_0 is an expander yet is the *most* observation-robust of the four.

The distinction is structural. $\sigma_{\max}(\Phi)$ perturbs the initial action noise with the observation held fixed, while R_{obs} perturbs the observation with the generation noise held fixed. These are different Jacobians, J_a versus J_c.

All four flows are also strongly anisotropic, so the $\sigma_{\max}$ verdict is a worst-case direction rather than typical behaviour (strong for SmolVLA, near-neutral for π_0). We therefore make no claim that a contractive flow is by itself robust, and R_{obs} is an open-loop, single-step proxy rather than closed-loop task success. Observation robustness must be measured directly, not inferred from generation-flow contraction.

5.4 Contraction can be engineered, but not for free

Finally, we ask whether the corrected end-to-end quantity can serve as a training objective. In a controlled A/B/C probe on SmolVLA and π_0 (App. F, Tab. 9), only the two-sided end-to-end penalty drives $\sigma_{\max}(\Phi)$ below 1; the one-sided CDP-class penalty fails to contract the deployed flow on both architectures ($2/2$), despite training driving its per-step abscissa down.

This does not buy a clean robustness win. Under a same-recipe non-degeneracy probe, inducing contraction collapses observation-coupling (Tab. 10), so the pre-registered outcome is a TRADEOFF NULL. Contraction can be engineered, but not while preserving the measured observation-coupling under the tested setup.

6 Conclusion

We applied one finite-difference-validated witness, the operator norm $\sigma_{\max}(\Phi)$ of each policy's own noise→action propagator, to four deployed flow-matching VLAs on LIBERO. The four split by sign, and the split survives controls on direction, regime, and step count while a within-family pair still diverges. Expansion versus contraction is therefore a property of action-generation design, not of the framework. The one-sided per-step criterion the literature trains against is neither necessary nor, by a constructed self-test, sufficient for end-to-end contraction, and J_a does not order observation robustness J_c, which must be measured rather than inferred. These findings are scoped to a worst-case, open-loop reading on four architectures.

Ethics Statement

This work analyzes the robustness properties of publicly available vision-language-action policies on a public benchmark. It collects no new data involving human subjects and introduces no new capability for attacking deployed systems. Its purpose is to inform safer evaluation and deployment of flow-matching policies by showing that contraction of the generation flow must not be read as a robustness guarantee. We identify no additional ethical concerns.

Reproducibility Statement

Every quantity we report is produced by a single estimator applied identically to all four policies, described in the method section and detailed in the appendix. The policies and the benchmark are publicly available, and the evaluation protocol, hyperparameters, and seeds are specified in the appendix. Our source code will be released to reproduce all reported numbers.

AI Use Statement

We used generative AI tools to assist with prose drafting and editing and with LaTeX integration. We did not use generative AI to produce the experimental results, the figures, or any of the measured quantities; all reported numbers are computed directly from the policies. We reviewed all AI-assisted content, and in particular we verified every reference in the bibliography against its authoritative primary source. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for details.

References

Amin Abyaneh, Charlotte Morissette, Mohamad H Danesh, Anas El Houssaini, David Meger, Gregory Dudek, and Hsiu-Chin Lin. Contractive diffusion policies: Robust action diffusion via contractive score-based sampling with differential equations. *arXiv preprint arXiv:2601.01003*, 2026.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision-Language-Action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Alexander Davydov and Francesco Bullo. Perspectives on contractivity in control, optimization, and learning. *IEEE Control Systems Letters*, 8:2087–2098, 2024.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Jianing Guo, Zhenhong Wu, Chang Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, et al. On robustness of Vision-Language-Action model against multi-modal perturbations. In *International Conference on Learning Representations*, volume 2026, pp. 70248–70272, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Winfried Lohmiller and Jean-Jacques E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, et al. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A Vision-Language-Action model with Open-World generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. SmolVLA: A Vision-Language-Action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Wenpin Tang and Hanyang Zhao. Contractive diffusion probabilistic models. *SIAM Journal on Imaging Sciences*, 19(2):1364–1389, 2026.

A PROOFS

This appendix gives the full proofs of the three propositions stated in the Method (§3); each statement is restated verbatim before its proof. Together they justify the three load-bearing steps of the witness: that the propagator *is* the sampler's Jacobian (Proposition 1, formalizing Eq. 2), that its top singular value decides the sign two-sidedly (Proposition 2, Eq. 3), and that the incumbent one-sided per-step object cannot make this decision (Proposition 3, the mechanism behind the self-test Eq. 5). Throughout we keep the corrected sign convention $\eta_{\mathrm{loc}} = -\lambda_{\max}(\mathrm{sym}\, J)$, so $\eta_{\mathrm{loc}} \geq 0$ is per-step contraction and a *negative* abscissa flags expansion.

Proposition 4 (Propagator exactness, cf. Proposition 1)**.** *The end-to-end Jacobian of the deployed sampler equals the propagator,* $\partial a_K / \partial a_0 = \prod_{k=K-1}^{0} (I + \mathrm{d}t\, J_k) = \Phi$.

Proof. We argue by induction on the number of executed steps, differentiating the deployed recursion $a_{k+1} = a_k + \mathrm{d}t\, v_\theta(a_k, t_k, c)$ of Eq. 1. *Base case.* Before any step, $\partial a_0/\partial a_0 = I$, the empty product. *Inductive step.* Assume $\partial a_k/\partial a_0 = \prod_{j=k-1}^{0}(I + \mathrm{d}t\, J_j)$. The conditioning c is frozen and t_k is a fixed schedule value, so the only dependence of a_{k+1} on a_0 is through a_k. By the chain rule applied to Eq. 1,

$$\frac{\partial a_{k+1}}{\partial a_0} = \frac{\partial a_{k+1}}{\partial a_k}\,\frac{\partial a_k}{\partial a_0} = \Big(I + \mathrm{d}t\,\tfrac{\partial v_\theta(a,t_k,c)}{\partial a}\big|_{a=a_k}\Big)\frac{\partial a_k}{\partial a_0} = (I + \mathrm{d}t\, J_k)\,\frac{\partial a_k}{\partial a_0},$$

where the middle equality differentiates the update and the last uses the definition $J_k = \partial v_\theta(a, t_k, c)/\partial a|_{a=a_k}$ of Eq. 2. Substituting the inductive hypothesis left-multiplies the running product by the new factor, $\partial a_{k+1}/\partial a_0 = (I+\mathrm{d}t\, J_k)\prod_{j=k-1}^{0}(I+\mathrm{d}t\, J_j) = \prod_{j=k}^{0}(I+\mathrm{d}t\, J_j)$, which is the claim with $k+1$ in place of k. Taking $k = K$ gives $\partial a_K/\partial a_0 = \prod_{k=K-1}^{0}(I + \mathrm{d}t\, J_k) = \Phi$, so the deployed sampler's end-to-end Jacobian is exactly the propagator of Eq. 2. □

Proposition 5 (Two-sided decision rule, cf. Proposition 2)**.** *$\sigma_{\max}(\Phi) > 1$ iff the linearized flow strictly expands some unit perturbation, and $\sigma_{\max}(\Phi) < 1$ iff it strictly contracts every unit perturbation.*

Proof. By the variational definition of the operator norm in Eq. 3, $\sigma_{\max}(\Phi) = \|\Phi\|_2 = \max_{\|u\|=1}\|\Phi u\|$, and the maximum is attained at the top right-singular vector u_1 (the unit sphere is compact and $u \mapsto \|\Phi u\|$ continuous). By Proposition 1, Φu is exactly the first-order image of the unit perturbation u under the deployed flow, so $\|\Phi u\|$ *is* the gain along u. *(⇒, expansion).* If $\sigma_{\max}(\Phi) > 1$, then the maximizer u_1 satisfies $\|\Phi u_1\| = \sigma_{\max}(\Phi) > 1$, so u_1 is a unit perturbation the flow strictly expands. Conversely, any strictly expanded unit u has $\|\Phi u\| > 1$, and since $\sigma_{\max}(\Phi) \geq \|\Phi u\|$ by maximality, $\sigma_{\max}(\Phi) > 1$. *(⇔, contraction).* $\sigma_{\max}(\Phi) < 1$ means $\max_{\|u\|=1}\|\Phi u\| < 1$, i.e. *every* unit u has $\|\Phi u\| < 1$ and is strictly contracted; conversely, if every unit direction is strictly contracted then the maximum over the (compact) sphere is < 1. Therefore testing the single scalar $\sigma_{\max}(\Phi)$ against 1 certifies expansion of some direction and its absence, the two-sided reading of Eq. 3. □

Proposition 6 (The one-sided criterion is not sufficient, cf. Proposition 3)**.** *There exist step factors, each with non-positive symmetric-part spectral abscissa (per-step "contracting" under the one-sided criterion, $\eta_{\mathrm{loc}} \geq 0$ at every knot), whose product Φ nonetheless has $\sigma_{\max}(\Phi) > 1$.*

Proof. We prove existence by construction, exhibiting the non-normal 2×2 mechanism that underlies the known-answer self-test of Eq. 5. *The factors.* Take $\mathrm{d}t = 1$ and the two step-Jacobians

$$J_0 = \begin{pmatrix} -0.8 & 1.5 \\ 0 & -0.8 \end{pmatrix}, \qquad J_1 = \begin{pmatrix} -0.8 & 0 \\ 1.5 & -0.8 \end{pmatrix},$$

so the step factors are $M_0 = I + J_0 = \left(\begin{smallmatrix} 0.2 & 1.5 \\ 0 & 0.2 \end{smallmatrix}\right)$ and $M_1 = I + J_1 = \left(\begin{smallmatrix} 0.2 & 0 \\ 1.5 & 0.2 \end{smallmatrix}\right)$, a pair of opposite-oriented shears. *Each factor is per-step contracting.* The symmetric part of each step-Jacobian is $\mathrm{sym}\, J_k = \left(\begin{smallmatrix} -0.8 & 0.75 \\ 0.75 & -0.8 \end{smallmatrix}\right)$, with eigenvalues $-0.8 \pm 0.75 = \{-1.55, -0.05\}$. Hence $\lambda_{\max}(\mathrm{sym}\, J_k) = -0.05 < 0$, i.e. the corrected one-sided abscissa $\eta_{\mathrm{loc}} = -\lambda_{\max}(\mathrm{sym}\, J_k) = 0.05 \geq 0$ at both knots, each factor is certified *contracting* by the per-step criterion (indeed each has spectral radius $0.2 < 1$). *Yet the product expands.* The propagator of Eq. 2 is $\Phi = M_1 M_0 = \left(\begin{smallmatrix} 0.04 & 0.30 \\ 0.30 & 2.29 \end{smallmatrix}\right)$. Its Gram matrix $\Phi^\top\Phi = \left(\begin{smallmatrix} 0.0916 & 0.699 \\ 0.699 & 5.3341 \end{smallmatrix}\right)$ has $\lambda_{\max}(\Phi^\top\Phi) = 5.4257$, so $\sigma_{\max}(\Phi) = \sqrt{5.4257} = 2.329 > 1$: the composed map strictly expands its top singular direction while every factor was per-step contracting. The amplification is the non-normal cross-axis coupling the symmetric part never sees, M_0 shears the first axis into the second and M_1 shears it back, so a perturbation is handed between axes and grows even though each factor damps along its own axes. Repeating the same alternating pair over more steps drives $\sigma_{\max}(\Phi)$ up without bound ($K = 10$ gives $\sigma_{\max}(\Phi) \approx 68.6$). The known-answer self-test of Eq. 5 is a *separate* engineered 2-D instance in this same regime, per-step contraction, end-to-end expansion, whose expand case reads $\sigma_{\max}(\Phi) = 58.96$ against a per-step verdict of contraction. Therefore per-step contraction at every knot ($\eta_{\mathrm{loc}} \geq 0$) does not imply end-to-end contraction: the one-sided criterion is not sufficient, and the two-sided witness $\sigma_{\max}(\Phi)$ of Eq. 3 is required to read the correct sign. This 2×2 is an analytic, hand-verified counterexample (its arithmetic is reproduced in full above, not drawn from any measured artifact); it is logically distinct from the multimodal-fidelity impossibility of App. F, which concerns a different obstruction, global contraction versus marginal fidelity, rather than per-step versus composed contraction. □

B Relation to Classical Contraction Theory and CDP

We claim no new inequality. The gap between *per-step* symmetric-part contraction (the one-sided log-norm criterion $\eta_{\mathrm{loc}} = -\lambda_{\max}(\mathrm{sym}\, J) \geq 0$) and *finite-time end-to-end* contraction ($\sigma_{\max}(\Phi) < 1$) is the classical non-normal transient-growth phenomenon. A product of factors each contracting in its own symmetric part can still amplify, because the symmetric part is blind to the cross-axis coupling that non-normal factors hand between directions (Proposition 3; Lohmiller & Slotine, 1998; Davydov & Bullo, 2024); related global operator-norm and Lipschitz-constant analyses bound sensitivity without composing the per-step factors (Miyato et al., 2018; Fazlyab et al., 2019). Our contribution is not this mechanism but its *instantiation and consequence* on deployed policies, in three parts. *(i)* The object the flow- and diffusion-policy literature certifies and, in the CDP class, trains against, the per-step largest symmetric-part eigenvalue $\lambda_{\max}(\mathrm{sym}\, J_k)$ (Abyaneh et al., 2026), is the *wrong-sided* quantity for a deployed sampler. It is neither necessary (§5.2: $\pi_{0.5}$ contracts on $79/80$ episodes yet is flagged an expander on $80/80$) nor sufficient (the 2×2 construction of Proposition 3). *(ii)* We do not stop at the per-step object. We form and diagonalize the *full composed propagator* $\Phi = \prod_k (I + \mathrm{d}t\, J_k)$ that the per-step criterion never composes, and report its exact operator norm. *(iii)* We exhibit a cross-architecture sign counterexample, two expanders and two contractors, each verdict on $\geq 79/80$ episodes, which no framework-level or per-step-certified account predicts. To our knowledge these three together, a deployed-VLA operator-norm audit, the composed propagator, and the 2-vs-2 dissociation, are absent from prior contraction analyses of generative policies, whose guarantees characterise the continuous-time target field or its training objective rather than the discrete, few-step map that is actually deployed (Lipman et al., 2022; Gao & Zhu, 2024; Tang & Zhao, 2026; Albergo et al., 2023; Karras et al., 2022).

The CDP object, precisely. Contractive Diffusion Policies (Abyaneh et al., 2026) enforce contraction by a *per-step* constraint on the largest eigenvalue of the symmetric part of the score-network Jacobian. Their contraction condition (Theorem 3.1) is $\lambda_{\max}(J^{\mathrm{sym}}_{\varepsilon_\theta}) < -f(t)h(t)^{-1}$ for all $t \in [0,1]$, imposed through a contraction loss $\mathcal{L}_c(\theta) = \max\big(-\beta,\ \lambda_{\max}(J^{\mathrm{sym}}_{\varepsilon_\theta}) + f(t)h(t)^{-1}\big)$ added to the score-matching objective as $\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{score}}(\theta) + \gamma\, \mathcal{L}_c(\theta)$ (their Eq. 9; f, h schedule functions, $\beta, \gamma > 0$ a margin and weight). This is exactly the one-sided, per-step $\lambda_{\max}(\mathrm{sym}\, J)$ object our witness corrects: it constrains the symmetric part at *each* step and never the composed propagator Φ, so it is blind to the non-normal cross-axis amplification that Proposition 3 exhibits and that $\sigma_{\max}(\Phi)$ measures. Our Arm C (App. F) penalizes precisely $\sum_k \mathrm{relu}(\lambda_{\max}(\mathrm{sym}\, J_k))$, the training analogue of this object, and reproduces its failure to contract the deployed flow end-to-end.

C Instrument Validation

Every per-model sign in this paper is produced by a single instrument: the operator norm $\sigma_{\max}(\Phi) = \|\Phi\|_2$ of the linearized discrete noise→action propagator $\Phi = \prod_k (I + \mathrm{d}t\, J_k)$ of each policy's *own* deployed velocity-Euler sampler, restricted to the active 7-DoF LIBERO action block. Because the entire cross-architecture claim rests on the trustworthiness of this sign, we validate the witness on four gates before reporting any per-model number, applying the identical protocol to all four policies. Table 3 summarizes the outcome; the prose below documents the rationale of each gate.

Gate 1, finite differences at the U-curve minimum. For each policy we check that the autograd action self-Jacobian agrees with a central finite-difference estimate. The subtlety the gate must respect is that FD error is a U-curve in the step size h: truncation error dominates at large h and float32 subtractive cancellation dominates at small h, so the true agreement is the *sweep minimum*, not the value at any fixed h. We therefore sweep h and report the minimum. This matters for π_0, whose deeper action expert has a larger local Jacobian norm ($\|J\| \approx 30$) and hence a higher cancellation floor that pushes the default $h = 10^{-3}$ off the true minimum; sweeping recovers the correct agreement (3.4×10^{-4} at $h = 0.1$). Every velocity path is computed in float32, because the FD check is meaningless in bf16. The four measured minima span 2.6×10^{-5} ($\pi_{0.5}$) to 3.2×10^{-3} (SmolVLA), all $\ll 10^{-2}$ (Table 3, column 1). The GR00T environment workarounds are measurement-neutral: the Eagle2 backbone's attention was switched flash→sdpa (mathematically identical) and runs in bf16, but the backbone produces only the *frozen* conditioning; the Jacobian flows solely through the float32 DiT action head, and the clean 8.1×10^{-4} FD confirms these environment choices cannot affect $\sigma_{\max}(\Phi)$.

Table 3: Instrument-validation summary for the four FM-VLA policies: FD relative error at the U-curve minimum (all $\ll 10^{-2}$); 2-D known-answer self-test (all PASS); nonlinear-predictiveness error; and non-degeneracy of the surprising contractor $\pi_{0.5}$. J_a sign relative to 1 only.

Model	FD rel. err. (autograd vs. FD)	Self-test (2-D known)	Nonlin.-predict. rel. err. (small ϵ)	Non-degen.
SmolVLA	3.2×10^{-3}	PASS	–†	PASS
π_0	3.4×10^{-4}	PASS	0.4%	PASS
GR00T-N1.5-3B	8.1×10^{-4}	PASS	–†	PASS
$\pi_{0.5}$	$\mathbf{2.6 \times 10^{-5}}$	PASS	$\mathbf{0.0\%}$	PASS
2-D known-answer self-test (identical for all four models): contract case $\sigma_{\max}(\Phi) = 1.19 \times 10^{-3}$; expand case $\sigma_{\max}(\Phi) = 58.96$; IMPL_OK/PASS=true.				

† Not-yet-witnessed: the common-footing nonlinear-predictiveness JSON witness for SmolVLA and GR00T is not present in the exported artifact set (their earlier single-direction figures come from a different estimator and are not carried here). The π_0 and $\pi_{0.5}$ entries (0.4%, 0.0%) are witnessed from `item3_nonlinear` in the Tier-2 JSONs.

Gate 2, the 2-D known-answer self-test with an engineered one-sided disagreement. The witness must recover a *known* sign before it is trusted on an unknown one. We construct a 2-D system in which the sign of the naive one-sided abscissa η (the per-step symmetric log-norm / spectral-abscissa surrogate that the field uses) is *engineered to disagree* with the true map gain. On this system our $\sigma_{\max}(\Phi)$ code reproduces the closed-form gain and the brute-force direction-swept growth: $\sigma_{\max}(\Phi) = 1.19 \times 10^{-3}$ in the "contract" case and $\sigma_{\max}(\Phi) = 58.96$ in the "expand" case, with IMPL_OK=true (Table 3, lower block). This is the direct witness that motivates the paper's methodological correction: it demonstrates a concrete case where the one-sided abscissa points the wrong way while the two-sided operator norm is correct, mirroring the real-world $\pi_{0.5}$ counterexample (one-sided rule labels it an expander, the two-sided witness shows it contracts). The self-test also passes as a propagator gate (`selftest_propagator`=true) on all four Tier-2 runs.

Gate 3, nonlinear predictiveness of the deployed sampler. The linearization is only useful if it predicts the behavior of the actual nonlinear sampler. We perturb the initial noise along Φ's top singular direction, run the *full nonlinear* sampler, and check that the measured gain reproduces $\sigma_{\max}(\Phi)$ and stays on the same side of 1 out to $\epsilon = 0.1$. The witnessed pi-model errors on the `libero_long` Item-3 subset (5 episodes $\times$ 2 seeds) are 0.4% (π_0) and 0.0% ($\pi_{0.5}$), with `all_same_side`=true, confirming the linearization is predictive of the deployed sampler where the common-footing witness is in hand. The corresponding SmolVLA and GR00T entries are not-yet-witnessed on this common footing (their earlier single-direction figures come from a different estimator and are not mixed in); we mark them explicitly (Table 3, †).

Gate 4, non-degeneracy of the surprising contractor. Because $\pi_{0.5}$ is the counterintuitive result (a generalist that contracts), we verify its contraction is a property of a functioning flow rather than a collapsed or constant output. Its `predict_action_chunk` emits finite, observation-dependent actions whose spread is non-trivial: the emitted chunk varies with the observation and carries non-trivial within-chunk per-dimension spread. This qualitative non-degeneracy is corroborated by the smooth variation of per-suite $\sigma_{\max}$ across the 80 episodes.

Dual-estimator cross-checks and the anisotropy self-test. Two independent estimators of the per-knot symmetric-eigenvalue agree on SmolVLA (`eigvalsh` vs. shifted-power: knot-0 relative 0.0164, knot-9 relative 5.1×10^{-5}). The anisotropy estimator used in App. D passes a three-leg known-answer gate (`selftest_grnd`: dense / rank-1 / float32-N=256-production) on all four models, and the batched per-row Jacobian reproduces the single-flow path to $\sim 10^{-6}$ (the batched runs inherit the FD validation). The verifier scripts of record (`verify_results.py`, `verify_tier2.py`) re-derive every headline number directly from the per-episode JSONs and print PASS next to each; on re-run they report ALL CHECKED CLAIMS PASS: TRUE with full four-model coverage, and all Tier-2 instrument gates PASS.

Corroborating quantities on a narrower footing. Three corroborating quantities are reported on a narrower footing than the headline numbers, and we flag them explicitly rather than fold them in silently. The SmolVLA (0.9%) and GR00T (9%) nonlinear-predictiveness figures come from an earlier single-direction estimator and lack a common-footing JSON witness in the exported set (marked † in Table 3). The $\pi_{0.5}$ non-degeneracy scalars (mean action shift between observations 0.29–0.44, within-chunk per-dimension std 0.02–0.17) and the batched-versus-single Jacobian agreement ($\sim 10^{-6}$) are corroborating as well. None is load-bearing for a *sign* verdict.

Scope of the instrument (honest floor). The witness is well-posed only for velocity-Euler FM samplers ($a_{k+1} = a_k + \mathrm{d}t\, v_\theta$), which is exactly what all four policies deploy; it is *not* applied to predict-action / rectified-flow samplers (e.g. X-VLA), whose output is near-deterministic in the sampling noise, and the FD gate *correctly refuses* that degenerate map. Two further scope statements are carried into the main text and repeated here so they are not softened. First, $\sigma_{\max}(\Phi)$ is a generation-flow / action self-Jacobian (J_a) instrument; it is *not* an output-robustness-to-observation (J_c) instrument, and the two are kept strictly distinct throughout (App. D and the main text quantify their decoupling). Second, the earlier one-sided rule "$\int \eta_{\mathrm{loc}}\, \mathrm{d}t < 0 \Rightarrow$ expansion" is a one-sided bound that does *not* imply expansion and is **refuted**: $\pi_{0.5}$ has $\int \eta_{\mathrm{loc}} < 0$ on $80/80$ episodes yet contracts on $79/80$. At panel scale this one-sided per-step rule *over-warns*: it false-alarms on genuine contractors ($\pi_{0.5}$ $79/80$; GR00T per-suite $0/2/8/1$ across spatial/object/goal/`libero_10`) with 0 false-safe on all four models (240-episode aggregate: 161 agree, 79 false-alarm, 0 false-safe). The earned claim is that the one-sided object is the *wrong diagnostic and training target*, neither necessary nor sufficient, *not* that any policy is provably unsafe; "never false-safe" is an empirical observation on this panel, not a guarantee. All cross-model comparisons are made on sign relative to 1, never on raw magnitudes across estimators, grids, or domains.

D Extended Experimental Setup

This appendix records the per-model deployment details, the per-episode averaging scheme, the active-block and normalization handling, and the pre-registered estimator, so that every number in the main text can be reproduced from its ecologically valid "as-deployed" operating point. All experiments are on LIBERO (Liu et al., 2023); each policy is measured on its own in-distribution observations and normalization.

Per-model deployed samplers. Each policy is measured on *its own* deployed sampler, not a re-imposed common one. **SmolVLA** (Shukor et al., 2025) deploys a 10-step velocity-Euler sampler at negative time step ($\mathrm{d}t < 0$), a compact generalist checkpoint. π_0 (Black et al., 2024) deploys 10 steps at $\mathrm{d}t < 0$ with a PaliGemma-plus-gemma-expert backbone and a deeper action expert. **GR00T-N1.5-3B** (Bjorck et al., 2025) is a per-suite specialist that deploys a 4-step sampler at positive time step ($\mathrm{d}t > 0$) through a diffusion-transformer (DiT) action head; its Eagle2 backbone runs in bf16 and produces only the frozen conditioning, while the Jacobian flows through the float32 DiT head (see App. C, Gate 1). $\pi_{0.5}$ (Physical Intelligence et al., 2025) is a generalist that shares the PaliGemma-plus-gemma-expert backbone *family* with π_0 but differs in its action head (adaLN time-conditioning and state folded into the prompt rather than the velocity path), deploying 10 steps at $\mathrm{d}t < 0$. This is the pair that isolates the split to action-head design: π_0 and $\pi_{0.5}$ agree on backbone family, denoising direction, and step count yet land on opposite sides of 1.

Per-episode averaging. Per-episode $\sigma_{\max}(\Phi)$ is averaged over sampling draws within the episode. For SmolVLA and GR00T the average is over frame 0 with two noise seeds (frame$_0 \times 2$ seeds); for π_0 and $\pi_{0.5}$ it is over frames $\{0, \mathrm{mid}\}$ with two seeds (frames$\{0, \mathrm{mid}\} \times 2$ seeds). All downstream statistics aggregate these per-episode values. The panel is 4 suites $\times 5$ tasks $\times 4$ episodes $= 80$ episodes per model.

Active block, padding, and normalization. The witness is formed from active-to-active step-Jacobians on the 7-DoF LIBERO action block with the padding latent dimensions held fixed. Φ is therefore a *product of restrictions*, not a principal submatrix of the full 32-dimensional propagator, so a contraction verdict is scoped to the active subspace and is not by itself a guaranteed two-sided bound on the full deployed flow. Because Φ is a product of restrictions and not a principal submatrix, Cauchy eigenvalue interlacing does *not* apply, so the active-block witness bounds the full

32-dimensional propagator in *neither* direction; an earlier interlacing "conservative bound" justification was invalid here and has been removed. Precisely because the active witness carries no such guarantee, we corroborate its sign against a directly measured full 32-dimensional deployment Jacobian (Tab. 7). For π_0 the shipped processor's normalizer carried the recipe but empty statistics; the real mean/std live in the checkpoint's model buffers and were applied manually to the state. $\sigma_{\max}(\Phi)$ is insensitive to the exact state operating point (raw vs. normalized state: 3.28 vs. 3.28 at the smoke observation), because the witness lives entirely in normalized-action latent space (action normalization statistics are irrelevant to it); the in-distribution point is used throughout.

Estimator, pre-registration, and the honest inconclusive note. Per-suite statistics use a cluster bootstrap (resample the 5 tasks per suite, then episodes) with a small-cluster cluster-t cross-check; the thresholds, sample sizes, estimators, and verdict rule were locked in `CLOSEOUT_PREREG.md` before running, and reported symmetrically for expanders and contractors, with no estimator changed to alter a verdict. Heavy-tailed expander magnitudes are summarized by proportion, median, and geometric mean; the arithmetic mean is retained only as a magnitude descriptor. We report the pre-registration faithfully: the conservative rule (widest of the bootstrap-mean and cluster-t intervals) returns INCONCLUSIVE on SmolVLA `libero_goal` and `libero_spatial`, and on π_0 `libero_long`, *solely* because the normal-theory cluster-t lower bound on a heavy-tailed non-negative quantity dips below 1 with only five task-clusters (the cluster-t lower bound is -3.68 on SmolVLA `goal` and 0.350 on π_0 `long`, a small-cluster normal-theory artifact on a heavy right tail), not because any episode in those suites contracts (the proportion remains $20/20$). The conclusion is unanimous under the appropriate robust estimators (proportion, median-CI, and geometric-mean-CI all exclude 1 on four of four suites). Two controls are reported on subsets rather than the full 80-episode panel, the matched-step-count dissociation table ($n = 4$ `libero_long` observations, frame 0, seed 0) and the R_{obs} bridge (5 `libero_long` scenes).

Per-suite $\sigma_{\max}(\Phi)$ (full panel). Table 4 records the full per-suite summary underlying the headline dichotomy: on every one of the sixteen (model, suite) cells the value lands on the same side of 1 as the per-model verdict, and the pre-registered conservative confidence intervals for the robust estimators exclude 1 on all sixteen cells. $\pi_{0.5}$ crosses 1 on exactly one episode (`libero_spatial`, episode 1275, mean $\sigma_{\max} = 1.816$, across-draw sd 1.606, driven by a single high draw), so the four-model supports are not strictly disjoint; the classes are separated cleanly *by sign*, not by support-disjointness.

Confound controls. Table 5 records the per-model summary underlying the main-text confound analysis: $\pi_{0.5}$ shares denoising direction ($\mathrm{d}t < 0$) and step count (10) with the expanders yet contracts, and class does not track training regime.

Matched step-count control. Table 6 dissociates the two classes at every matched step count tested, ruling out step count as the cause of the split.

Deployment-Jacobian corroboration. Table 7 answers the frozen-padding objection by measuring the full end-to-end deployment Jacobian directly, with all 32 latent dimensions evolving through each policy's own sampler: the directly measured $\sigma_{\max}^{\text{deploy}}$ lands on the same side of 1 as the active-block witness for all four models, and padding-to-active coupling redistributes without ever flipping the sign. The deploy-versus-active deviation is $\sim 1\%$ for π_0, $\pi_{0.5}$, and SmolVLA; for GR00T it reaches a true maximum of $\sim 8.5\%$ (episode-4 ratio 1.0193, deploy marginally *above* active yet still < 1), and GR00T's padding coupling is a non-trivial $\sim 45\%$ that only redistributes magnitude without flipping the sign.

Full-suite anisotropy ratios. We scope the dichotomy explicitly as a *worst-case* $\sigma_{\max}$ claim, not a typical or RMS-direction claim, and we pre-empt the "cherry-picked direction" objection with a pre-registered random-direction estimator. Table 8 extends the main-text `libero_long` anisotropy result to all four suites under one common pre-registered estimator: $N = 256$ random unit directions, linear matvec $\|\Phi d\|$ on the materialized active-block Φ, fixed seed 777, with g_{rnd} the RMS gain over the 256 directions. Because RMS $\geq$ median, g_{rnd} upper-bounds the typical direction's gain, so the ratio $\sigma_{\max}/g_{\text{rnd}}$ is a conservative anisotropy measure. The estimator is validated against the closed form $\mathbb{E}\|\Phi d\|^2 = \|\Phi\|_F^2/n$ (the 256-direction RMS must converge to $\|\Phi\|_F/\sqrt{n}$; maximum

Table 4: **Per-suite $\sigma_{\max}(\Phi)$ summary.** Median and $[\min, \max]$ of per-episode $\sigma_{\max}(\Phi)$, count on the model's side of 1, and the pre-registered conservative median-CI (lower bound relative to 1) with its status. Every cell lands on the same side of 1 as the per-model class; INC. marks the pre-registered cluster-t INCONCLUSIVE flag on heavy-tailed suites (proportion still $20/20$; robust estimators unanimous). $n = 20$ episodes per (model, suite) cell.

Model (class)	Suite	Median	$[\min, \max]$	# on-side	Median-CI vs. 1	Status
SmolVLA (EXPAND)	long	3.78	$[1.34, 46.33]$	$20/20$	$[2.62, \cdot] > 1$	EXPAND
	goal	2.88	$[1.94, 107.22]$	$20/20$	$[2.22, \cdot] > 1$	INC.†
	object	2.09	$[1.52, 21.92]$	$20/20$	$[1.89, \cdot] > 1$	EXPAND
	spatial	3.09	$[2.02, 183.37]$	$20/20$	$[2.47, \cdot] > 1$	INC.†
$\boldsymbol{\pi_0}$ (EXPAND)	long	3.24	$[2.66, 20.61]$	$20/20$	$[2.90, \cdot] > 1$	INC.†
	goal	3.59	$[2.91, 4.46]$	$20/20$	$[3.28, \cdot] > 1$	EXPAND
	object	5.99	$[3.92, 28.32]$	$20/20$	$[4.54, \cdot] > 1$	EXPAND
	spatial	3.54	$[2.75, 4.43]$	$20/20$	$[3.11, \cdot] > 1$	EXPAND
GR00T (CONTRACT)	long	0.344	$[0.261, 0.503]$	$0/20$	$[\cdot, 0.503] < 1$	CONTRACT
	goal	0.468	$[0.194, 0.617]$	$0/20$	$[\cdot, 0.617] < 1$	CONTRACT
	object	0.356	$[0.237, 0.396]$	$0/20$	$[\cdot, 0.396] < 1$	CONTRACT
	spatial	0.410	$[0.247, 0.716]$	$0/20$	$[\cdot, 0.716] < 1$	CONTRACT
$\boldsymbol{\pi_{0.5}}$ (CONTRACT)	long	0.296	$[0.167, 0.851]$	$0/20$	$[0.21, 0.36] < 1$	CONTRACT
	goal	0.412	$[0.213, 0.852]$	$0/20$	$[0.32, 0.59] < 1$	CONTRACT
	object	0.338	$[0.126, 0.905]$	$0/20$	$[0.28, 0.42] < 1$	CONTRACT
	spatial	0.257	$[0.189, 1.816]$	$1/20$	$[0.22, 0.44] < 1$	CONTRACT‡

†Pre-registered INCONCLUSIVE: the normal-theory cluster-t lower bound dips below 1 on a heavy right tail with five task-clusters; no episode contracts (proportion $20/20$) and the robust estimators exclude 1. ‡One `libero_spatial` outlier (episode 1275, mean $\sigma_{\max} = 1.816$, across-draw sd 1.606, driven by a single high draw) crosses 1; the conservative and median CIs still exclude 1 from below.

Table 5: Confound controls: per-model class, median gain, and the three candidate framework-level explanations (denoising direction, step count, training regime), compared by sign relative to 1 only. $\boldsymbol{\pi_{0.5}}$ shares denoising direction and step count with the expanders yet contracts, breaking all three at once (matched step-count control in Tab. 6).

Model	Class	Median $\sigma_{\max}$	On-side	dt sign	Regime
SmolVLA	EXPAND	3.09	$80/80 > 1$	< 0	generalist
π_0	EXPAND	3.75	$80/80 > 1$	< 0	generalist
GR00T	CONTRACT	0.358	$80/80 < 1$	> 0	specialist
$\boldsymbol{\pi_{0.5}}$	CONTRACT	$\mathbf{0.324}$	$\mathbf{79/80 < 1}$	$< \mathbf{0}$	generalist

relative error 3.4% across the 40 `libero_long` flows) and passes the three-leg `selftest_grnd` on all four models. Every one of the sixteen cells satisfies $g_{\mathrm{rnd}} \leq \sigma_{\max}$ (`bounds_ok`=true), and the g_{rnd}-vs-closed-form check is at most 9.05% at its worst cell (SmolVLA / `libero_spatial`). The reading is uniform: all four models are anisotropic on every suite (median ratio ~ 3 to ~ 12), so the expand/contract dichotomy is a statement about the *worst-case* direction $\sigma_{\max}$, not the typical (RMS) direction. The thin-subspace reading is strong for SmolVLA (its RMS gain sits well below its worst case) and weak for π_0 (its RMS gain is already ≈ 1.0, near-neutral rather than a contracting background).

Anisotropy and $J_a \neq J_c$ decoupling (`libero_long`). The main-text Table 2 pairs the `libero_long` anisotropy of the generation flow with an observation-sensitivity proxy R_{obs}, and the two axes come out *decoupled* (Fig. 3). All four models are strongly anisotropic, so the dichotomy is a worst-case statement. The J_a expand/contract sign does *not* order R_{obs} (median, seed 777): π_0 0.0037 $[0.0028, 0.0043]$, $\pi_{0.5}$ 0.022 $[0.015, 0.055]$, GR00T 0.029 $[0.017, 0.057]$, SmolVLA 0.481 $[0.27, 0.54]$. The two *expanders* bracket the entire range: π_0 is the *most* observation-robust of the four, also in absolute terms (mean action shift $\|\Delta a\| \approx 0.084$ versus SmolVLA's 4.79), while SmolVLA is the least. A generation-flow verdict therefore carries no order information about observation robustness. **Non-circularity.** R_{obs} and $\sigma_{\max}(\Phi)$ are not two views of one measurement. R_{obs} is the RMS relative action shift under $\sigma = 0.03$ image perturbations along 64 random directions with

Table 6: Matched step-count control (common-footing subset*): SmolVLA worst-case min $\sigma_{\max}$ (all > 1) vs. GR00T worst-case max $\sigma_{\max}$ (all < 1) at every tested step count. Sign relative to 1 only.

Steps	SmolVLA min $\sigma_{\max}$ (all > 1)	GR00T max $\sigma_{\max}$ (all < 1)	Split
4	4.31	0.31	✓
8	3.13	0.39	✓
10	**1.73**	**0.41**	✓
16	1.69	0.44	✓

*$n = 4$ `libero_long` observations (frame 0, seed 0); these per-step values are on a common footing with each other but are *not* directly comparable to the full-suite numbers of Table 4, and the control matches step count but not denoising orientation for GR00T (its $\mathrm{d}t > 0$ confound is broken separately by $\pi_{0.5}$). They are reported on this subset rather than the full 80-episode panel.

Table 7: Deployment-Jacobian corroboration: directly measured $\sigma_{\max}^{\text{deploy}}$ over all 32 latent dimensions (5 `libero_long` episodes). For every model the deploy Jacobian lands on the same side of 1 as the active-block witness; padding coupling redistributes without flipping the sign.

Model	$\sigma_{\max}^{\text{deploy}}$ range	Side of 1	Coupling (padded/active)	Deploy ≈ active?
$\pi_{0.5}$	**0.20–0.44**	< 1	7.3–8.6%	yes (∼1%)
GR00T	0.24–0.27	< 1	40.5–50.6%	yes† (max dev ∼8.5%)
π_0	3.01–3.39	> 1	9.1–10.3%	yes (∼1%)
SmolVLA	1.69–123.6	> 1	10.1–13.4%	yes (∼1%)

†GR00T's end-to-end finite-difference check is float32-curvature-limited on its short 4-step map; the contraction verdict rests on the fully finite-difference-clean per-step witness.

the generation noise *held fixed*; it perturbs the *observation* input. $\sigma_{\max}(\Phi)$ perturbs the *noise latent* with the observation held fixed. Different inputs, different Jacobians (J_c versus J_a). **Honest floor.** R_{obs} is an open-loop, single-step proxy, not closed-loop task success; its ordering agrees only *directionally* with an external closed-loop robustness report (Guo et al., 2026), and output robustness must be measured, not inferred from J_a. The full-suite anisotropy ratios above (Table 8) extend the anisotropy leg to all four suites.

Subset-scale controls. Four controls are reported on subsets rather than the full 80-episode panel, the matched-step-count dissociation ($n = 4$ `libero_long`, frame 0, seed 0; Table 6), the per-step Track-1 mechanism (10 flows per suite), the deployment Jacobian (5 episodes per model; Table 7), and the R_{obs} bridge (5 `libero_long` scenes). A direct denoising-direction flip control (reversing the time orientation) was *not* run, because $\pi_{0.5}$ already breaks the $\mathrm{d}t$-sign collinearity observationally (it contracts at $\mathrm{d}t < 0$), making that control confirmatory rather than load-bearing.

E Per-Step Mechanism

This appendix expands the distributed-composition mechanism summarized in the main text (§5.1, Fig. 2b). Tracking the running partial product $\sigma_{\max}(\prod_{j \le k})$ along the flow shows *how* two backbone-siblings reach opposite verdicts on the same observations and noise seeds, corroborating that the split is a genuine property of the flow, not a single pathological step. π_0's partial product crosses the boundary early and climbs on every knot, so its expansion is the compounded product of many mild per-step gains; $\pi_{0.5}$, on identical inputs, stays below 1 throughout, with only a lone knot ever exceeding unity. The mechanism is thus distributed composition, exactly why a per-step one-sided penalty is the wrong lever: it would fire on every one of the deepest contractor's knots (two independent instruments agree here to roughly two decimals). The signature is clean on three of four suites; on `libero_object` the median π_0 flow is still composition-driven but its heavy right tail includes a single dominant-knot event the pre-registered rule flagged rather than absorbed, the dichotomy is nonetheless unaffected, holding on every flow of every suite. The paired running partial products are the per-flow decomposition of the split reported in the main text; the two independent instruments agree to ∼two decimal places on every suite.

Table 8: Full-suite anisotropy ratios (median $\sigma_{\max}/g_{\text{rnd}}$ per (model, suite) cell), pre-registered N=256, seed-777 estimator. All sixteen cells satisfy $g_{\text{rnd}} \leq \sigma_{\max}$; worst g_{rnd}-vs-closed-form error 9.05%. Source: `item4_anisotropy.summary` in the Tier-2 JSONs.

Model	median anisotropy ratio $\sigma_{\max}/g_{\text{rnd}}$				range across
	`long`	`goal`	`object`	`spatial`	suites
SmolVLA (EXPAND)	7.86	7.01	6.81	7.67	6.8–7.9
π_0 (EXPAND)	3.04	4.08	**11.65**	8.30	3.0–11.6
GR00T (CONTRACT)	5.85	7.79	6.41	5.75	5.8–7.8
$\pi_{0.5}$ (CONTRACT)	9.28	8.86	9.84	8.79	8.8–9.8

All sixteen cells witnessed from `tier2_{model}_libero_{suite}.json` (`item4_anisotropy.summary.anisotropy_ratio_median`). The witnessed `libero_long` row (with $\sigma_{\max}$ median, g_{rnd}, and ratio range) is: SmolVLA 2.05, 0.276, [5.59, 18.88]; π_0 3.14, 1.01, [2.88, 3.40]; GR00T 0.27, 0.049, [5.17, 6.20]; $\pi_{0.5}$ 0.30, 0.035, [7.35, 14.89].

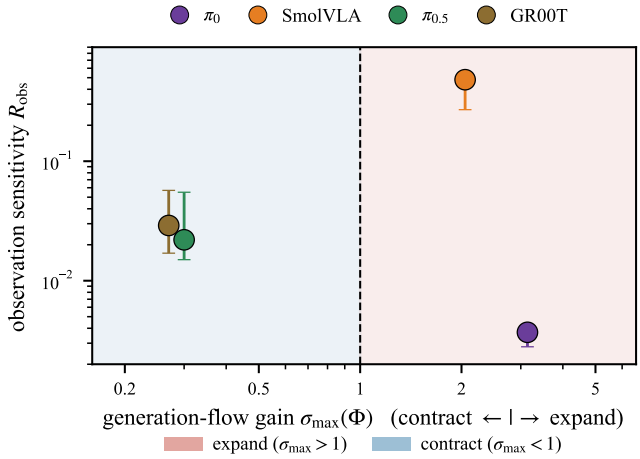


Figure 3: Generation-flow geometry (J_a) does not order observation robustness (J_c), on `libero_long`. Each policy is placed by its generation-flow gain $\sigma_{\max}(\Phi)$ (horizontal; left of 1 contracts, right expands) and its observation-sensitivity proxy R_{obs} (vertical; lower is more robust). The two *expanders* bracket the entire R_{obs} range, π_0 is the most observation-robust of the four and SmolVLA the least, so the expand/contract sign carries no order information about R_{obs}.

Per-flow decomposition and scope. On the paired `libero_long` flow (identical observations and seeds), π_0 exceeds unity at $10/10$ knots (median top-knot gain $G_k \approx 1.33$, $\sigma_{\max} = 3.14$), while $\pi_{0.5}$ exceeds unity at only $1/10$ knots (median top-knot $G_k \approx 1.12$, $\sigma_{\max} = 0.30$): opposite verdicts from distributed per-step profiles, not one dominant knot. Track-1 (per-knot) and Tier-2 (batched) $\sigma_{\max}$ medians agree to ∼two decimals on every suite, but they are *different aggregates* at different n and are never conflated. The one exception to the no-dominant-knot signature is `libero_object`, where $2/10$ π_0 flows carry a heavy tail ($\sigma_{\max} \approx$ 26–27) with a single knot ($G_k = 3.03$) exceeding the pre-registered 2.5 bound; the pre-registered rule *flags* rather than absorbs this, and the sign dichotomy ($10/10$ versus $10/10$) is unaffected.

Existential, not causal. The within-family π_0-versus-$\pi_{0.5}$ flip localizes the split to design *within the action-generation head* (shared backbone family, $\mathrm{d}t < 0$, 10 steps, opposite signs), but it does *not* isolate the single responsible choice: the pair differs in more than one way (adaLN versus additive time-conditioning; state folded into the prompt versus the velocity path; training recipe and checkpoint). Causal attribution to one design knob is pre-registered **forbidden** and left as future work. The confounds we *do* break are the framework-level ones (denoising direction, generalist/specialist, step count); the mechanism here is descriptive, not causal.

F CONSTRUCTIVE PROBE: THE TRADEOFF NULL

A constructive probe on SmolVLA (replicated on π_0) asks whether *penalizing* the correct two-sided end-to-end object, $\sigma_{\max}(\Phi)$, the operator norm of the noise→action propagator, a J_a quantity, can *induce* generation-flow contraction, and at what cost. This is a supplementary honest-negative that pre-empts the "just train for contraction" pushback. Everything here is a J_a statement, makes no J_c claim, and compares sign relative to 1.

Design. In a controlled A/B/C comparison with an identical recipe (all-suites 75 training entries, 150 steps, seed 42, expert-only, auto-λ init-balanced, three fixed penalty observations with a fixed generation-noise seed), arm **A** is an FM-only control; arm **B** adds the two-sided end-to-end penalty $\mathrm{relu}(\log \sigma_{\max}(\Phi))$; and arm **C** adds a one-sided per-step penalty $\mathrm{relu}(\lambda_{\max}(\mathrm{sym}\, J_k))$, a faithful CDP-class object (Abyaneh et al., 2026), the diffusion-policy analogue of the per-step contraction penalty. Held-out assessment is on `libero_10` (5 episodes $\times$ 2 seeds). Before committing GPU, a Gate-0 feasibility check confirmed the two-sided penalty back-propagates matrix-free without OOM and that $\sigma_{\max}$ is a well-posed monotone knob (a fixed-trajectory gain-scale sweep falls monotonically, $1.633 \to 1.192$ as the step-Jacobian scale α goes $1.00 \to 0.25$); the eigenvalue step reproduces the FD-validated per-row reference to $\leq 5.5 \times 10^{-6}$ (a matrix-free Lanczos cross-check is an optional additional validation we did not run).

The penalized objects, and Arm C faithfulness. Both penalties are computed matrix-free (no materialized Jacobian) by forward-over-reverse JVP power iteration, so both are differentiable in the policy parameters and fit in $\sim$2.9 GB. Arm B penalizes the *composed, end-to-end* operator norm, $\mathrm{relu}(\log \sigma_{\max}(\Phi) - \log \tau)$ with $\Phi = \prod_k (I + \mathrm{d}t\, J_k)$ and $\tau \leq 1$, targeting exactly the deployed gain the paper reports rather than a per-knot proxy. Arm C penalizes the *raw per-step* symmetric-part abscissa, $\sum_k \mathrm{relu}(\lambda_{\max}(\mathrm{sym}\, J_k))$, never $(I + \mathrm{d}t\, J_k)$ and never composed, which is the standard one-sided log-norm object and the exact quantity pre-registered as the CDP-class foil. Training verifiably drove Arm C's per-step abscissa down (SmolVLA batch mean $\to 0.42$; $\pi_0 \to 0.006$), yet the deployed flow still failed to contract (SmolVLA 1.06, π_0 2.75), mirroring its diagnostic false-alarm failure. Two scope notes follow. First, Arm B and Arm C differ on *two* axes at once, sidedness (two- versus one-sided) and composition (end-to-end versus per-step), so the earned claim is disjunctive: the two-sided end-to-end object contracts the deployed flow where the one-sided per-step object cannot; we do *not* claim sidedness alone is responsible. Second, the differentiable $\sigma_{\max}$ was validated on a known-answer toy (it matches a brute-force SVD to $< 2\%$, has non-zero parameter gradient, and gradient descent on $\log \sigma_{\max}$ provably reduces the true $\sigma_{\max}$), ruling out the inert-penalty failure mode of an earlier implementation.

Result: the two-sided object is the right target; the one-sided CDP-class object is not. Table 9 reports the arms. On SmolVLA the two-sided arm B is the *only* one that drives $\sigma_{\max}(\Phi)$ below 1 end-to-end (0.97 at step 150, from a baseline 1.867), while the one-sided CDP-class arm C *fails to contract* (1.06) despite training driving its per-step abscissa down. B and A are identical on every measured axis (e.g. `cross_obs_var` 0.70 vs. 0.70) except $\sigma_{\max}$ (B 0.97 contracts vs. A 1.43 expands), so B's contraction is attributable to the penalty, not to the shared recipe. B contracts harder earlier ($\sigma_{\max} = 0.71$ at step 50, 0.85 at 100, 0.97 at 150) as the relu deactivates once $\sigma < 1$ and the FM loss drifts it back toward 1; a $\tau = 0.8$ variant settles stably at 0.70 (step 150), showing the transient drift is a relu-deactivation artifact and a stable deeper contractor is reachable. The π_0 replication reproduces the ordering across architecture (two-sided B is the only arm to contract, $3.04 \to 0.99$; one-sided C fails at 2.75); the "CDP wrecks fidelity" effect is SmolVLA-specific (on π_0, arm C's FM-loss 1.62 is fine, below the baseline 2.60, it simply fails to contract), so only "two-sided engineers contraction where the one-sided CDP-class object cannot" is architecture-general ($2/2$ models).

The gentle-LR non-degeneracy probe: contraction causally costs observation-coupling. The fine-tuning recipe drops observation-coupling for *every* arm, the control included, so the binary gate (contract while preserving observation-coupling $\geq 0.7\times$ the raw baseline) is failed by every arm and the pre-registered verdict is TRADEOFF NULL. To isolate cause from recipe, a gentle-LR probe (learning rate 2×10^{-5}, 120 steps) uses a recipe gentle enough that FM-only *preserves* observation-coupling. Table 10 shows the decisive same-recipe, penalty-only-differs isolation: FM-only (arm

Table 9: Constructive A/B/C probe (TRADEOFF NULL). Arm A: FM-only control. Arm B: FM + two-sided end-to-end $\mathrm{relu}(\log \sigma_{\max}(\Phi))$. Arm C: FM + one-sided per-step $\mathrm{relu}(\lambda_{\max}(\mathrm{sym}\, J_k))$ (faithful CDP-class object (Abyaneh et al., 2026)). "contracts" = $\sigma_{\max}(\Phi) < 1$. Held-out `libero_10` (5 ep $\times$ 2 seeds). Only the two-sided arm B reaches $\sigma_{\max} < 1$; the one-sided arm C fails on both architectures. Source: `sigma_max_median` in the per-arm `eval_arm_*` JSONs.

Architecture	Arm	$\sigma_{\max}(\Phi)$	contracts	FM-loss
SmolVLA @ step 150	baseline	1.867	×	0.068
	A (FM-only)	1.43	×	0.094
	B (two-sided e2e)	**0.97**	✓	0.063
	C (one-sided CDP)	1.06	×	0.245
π_0 @ step 100	baseline	3.04	×	2.60
	A (FM-only)	3.03	×	1.57
	B (two-sided e2e)	**0.99**	✓	1.50
	C (one-sided CDP)	2.75	×	1.62

Table 10: Gentle-LR non-degeneracy probe (SmolVLA, lr 2×10^{-5}, 120 steps): FM-only preserves observation-coupling and expands; the two-sided penalty contracts *only* by collapsing observation-coupling. `cross_obs_var` ratio relative to untuned base; (✓/×) = $\geq 0.7\times$ preservation gate.

Arm	step	$\sigma_{\max}(\Phi)$	`cross_obs_var` (ratio)	reading
A (FM-only)	30	2.57	4.69 (0.86×, ✓)	non-degenerate, EXPANDS
A (FM-only)	120	7.42	4.07 (0.75×, ✓)	non-degenerate, EXPANDS
B ($\sigma_{\max}$ penalty)	30	0.57	1.05 (0.19×, ×)	CONTRACTS, degenerate
B ($\sigma_{\max}$ penalty)	**120**	**0.87**	**1.29 (0.24×, ×)**	**CONTRACTS, degenerate**

A) is non-degenerate and *expands* ($\sigma_{\max}$ 2.57 $\rightarrow$ 7.42; `cross_obs_var` 0.86× then 0.75× the base, both above the 0.7× threshold), whereas the two-sided penalty (arm B) achieves contraction ($\sigma_{\max}$ 0.57–0.87) *only* by collapsing observation-coupling (`cross_obs_var` $\rightarrow$ 0.19–0.24× the base). Same recipe, only the penalty differs: the penalty causally trades observation-coupling for contraction. The π_0 replication shows the same tradeoff (all arms collapse observation-coupling to $\sim$9–13% of base, with B at 0.087× the arm that contracts).

Contraction-by-collapse: detection and a retracted result. Degeneracy is the primary threat to any constructive claim: a policy that ignores its input trivially has low noise$\rightarrow$action gain, so a $\sigma_{\max}$ drop is meaningful only if observation-coupling survives. We therefore gate on anti-collapse with a relative gate (observation responsiveness $\geq 0.7\times$ the untuned baseline) and an *absolute* gate (both the action scale and `cross_obs_var` $\geq 0.7\times$ base, closing the scale-invariant-shrink loophole). This gate did real work: it *killed* an earlier generation of the experiment, a provisional constructive PASS whose penalty was inert on 298/300 steps and whose $\sigma_{\max}$ drop was pure policy collapse (`cross_obs_var` 5.46 $\rightarrow$ 0.33, i.e. 0.06× base). That result is retracted and is not part of this paper; the arms reported here are the post-fix run, and the gentle-LR probe (Table 10) is the isolation that survives this gate for the FM-only control while exposing the penalty's collapse.

Verdict and honest scope. The pre-registered verdict is TRADEOFF NULL: no arm reaches $\sigma_{\max}(\Phi) < 1$ while preserving observation-coupling, so a "clean constructive win" (contraction without degradation) is *not* achievable from this base, one obtains either a non-degenerate expander (FM-only) or a degenerate contractor (penalty), not both. What is claimed, and is architecture-general (2/2 models), is bounded: (i) contraction is an *engineerable* design property (the B-vs-A causal isolation); and (ii) the two-sided end-to-end object is the right training target while the one-sided CDP-class object (Abyaneh et al., 2026) is not (B-vs-C), the constructive echo of the main-text diagnostic that the one-sided per-step object false-alarms on contractors. The failure of arm C shows the one-sided object is *not necessary* for end-to-end contraction; it does not show the object is unsafe (no false-safe was observed). The scope is honest and narrow: open-loop, a single base model per architecture, a single held-out suite (`libero_10`), and mild contraction ($\sigma_{\max} \sim$ 0.71–0.97); `cross_obs_var` and R_{obs} are open-loop proxies, not closed-loop task success. Combined with the

Table 11: Multimodal-fidelity impossibility (1-D optimal-transport demonstration, within-mode $\mathrm{sd} = 0.3$). As the mode gap Δ grows, the fidelity-preserving map's inter-mode-valley slope $\sup|T'| = \sigma_{\max}$ grows explosively while the within-mode slope stays < 1; forcing $|T'| \leq 1$ (a globally contractive map) collapses the far-mode coverage away from the true ≈ 0.5. Overlapping modes ($\Delta = 0.5$) are the built-in control, where contraction is free.

Mode gap Δ	$\sup\|T'\|$ (valley)	within-mode $\|T'\|$	far-mode cover	cover, forced $\|T'\| \leq 1$
0.5 (control)	0.43	0.34	0.46	0.46
1.0	1.20	0.35	0.47	0.46
1.5	6.83	0.35	0.48	0.38
2.0	77.6	0.35	0.49	0.25
3.0	8.1×10^4	0.35	0.50	0.00
4.0	1.3×10^9	0.35	0.50	0.00

decoupling of J_a from J_c in the main text, the prescriptive message is sharp: making an FM-VLA contractive is neither free nor beneficial as an optimization target.

A likely-fundamental tension: contraction versus multimodal fidelity (1-D demonstration). The tradeoff-null is not obviously a training artifact. A 1-D optimal-transport calculation suggests it reflects a genuine tension between global contraction and fidelity for *multimodal* action distributions. In one dimension the fidelity-preserving (correct-marginal) noise$\to$action map of minimal transport is the monotone rearrangement $T(z) = F^{-1}_{\mathrm{target}}(\Phi(z))$, whose slope *is* the local gain, $T'(z) = \varphi(z)/p_{\mathrm{target}}(T(z)) = \sigma_{\max}$ in 1-D (with φ, Φ the standard normal pdf and cdf). For a bimodal target with mode gap Δ and within-mode $\mathrm{sd} = 0.3$, the inter-mode valley (where p_{target} is tiny) forces the slope up while the within-mode slope stays below 1. Forcing $|T'| \leq 1$ everywhere is exactly a globally contractive map, and it provably cannot reproduce the bimodal marginal: it collapses mode coverage (Table 11). The reading is that a fidelity-preserving multimodal generator *must* expand at the inter-mode valley, with the expansion growing explosively in the mode separation, while an overlapping (effectively unimodal) target contracts for free. This is a 1-D exact demonstration, not a theorem for the 32-dimensional action chunk; the n-dimensional statement is owed as future work, and the corollary that our empirical contractors are *less* multimodal than the expanders is a testable *prediction* we have not verified. Its role here is to give the tradeoff-null a principled, non-artifactual reading, and to answer the "just add a smarter contraction loss" objection: for a genuinely multimodal action distribution, no loss can buy global contraction without sacrificing a mode (shown here in 1-D).