CoT-Faith: Magnitude Scores Do Not Measure Faithfulness in Manipulation VLAs

Anonymous authors
Paper under double-blind review

Abstract

Vision-language-action models (VLAs) that emit chain-of-thought (CoT) are proposed as a route to interpretable robotics, on the premise that the emitted reasoning drives the action. The standard test scores an edit to the reasoning by how often it moves the action. We introduce CoT-Faith, a paired-intervention benchmark with 13 edit families, size-matched nulls and closed-loop evaluation, and apply it to 11 token-decoded VLA configurations from two model lineages. We show that the test's verdict depends on which meaning-preserving edit serves as its floor: under teacher-forced scoring, two judge-validated floors give opposite verdicts on 10 of 11 configurations, including the two competent policies. The score mostly records whether the action changes at all, which depends on what an edit touches rather than what it means. Nulls matched to each edit in size and tag restore a consistent sign, with direction flips beating their null on 11 of 11, but a control fine-tuned without a CoT target passes them too. Requiring the action to point the implied direction moves the top-ranked model from first of eight to seventh. On two competent LIBERO policies the action is sensitive to the reasoning at each step, yet editing or even emptying it leaves closed-loop success unchanged: sensitivity is not reliance. Edit-based scores should be read against matched nulls, a direction-aware score and closed-loop success before they license a claim.

1 Introduction

VLAs that emit chain-of-thought reasoning before their action tokens are proposed as a route to interpretable robots: the reasoning is expected to expose the model's plan for review, correction, or safety monitoring (Zawalski et al., 2024). That premise requires the reasoning to *drive* the action. The natural test edits the reasoning and asks whether the action changes. For driving VLAs, Nguyen et al. (2026) score such interventions by the magnitude of the induced trajectory change; for manipulation, Huang et al. (2026) read the action change caused by a hijacked CoT as evidence that the reasoning governs the decode. The standard test edits the reasoning and scores how often the action moves by more than a threshold. We ask whether it measures reliance on the reasoning, on manipulation VLAs whose actions are decoded as tokens.

It does not. Two reasonable meaning-preserving controls give opposite verdicts on the same models, because the score mostly records whether the action changes at all; matched controls fix the sign but not the question; and on competent policies in closed loop, the reasoning's content, and even its presence, does not affect success.

Contributions.

1. *A benchmark protocol with matched controls* (§2): 13 edit families, two judge-validated size-matched nulls, a size-matched primary statistic, and paired open- and closed-loop metrics, applied to 11 configurations from two model lineages.
2. *The magnitude verdict depends on the control* (§3.1–3.2): two judge-validated floors give opposite signs on 10 of 11 configurations, because the score is close to the rate at which the action changes at all.
3. *Moving is not following* (§3.3): size-matched controls fix the sign but a control without a CoT target passes them, and a direction-aware diagnostic shows the flagship public reasoning VLA moves without reversing, in distribution too.

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.

4. *Closed-loop evaluation on competent policies* (§3.4): editing or emptying the reasoning leaves success unchanged, so open-loop scores track how far the action moves, not reliance on the reasoning.

Related work. *CoT faithfulness in language models.* Stated reasoning often does not determine the output (Turpin et al., 2023; Lanham et al., 2023); paraphrasing the CoT is a standard faithfulness control (Lanham et al., 2023), and outputs are sensitive to meaning-preserving prompt changes (Sclar et al., 2024; Mizrahi et al., 2024). Our control result is that sensitivity read out in action space. *Reasoning VLAs.* ECoT (Zawalski et al., 2024), CoT-VLA (Zhao et al., 2025) and DeepThinkVLA (Yin et al., 2025) emit reasoning before acting; Chen et al. (2025a) find that ECoT-style reasoning helps largely through the representations it trains and that such policies can act without generating it, which our empty-CoT result agrees with. *Editing and attacking the reasoning.* On a LIBERO reasoning policy, entity swaps cost 8.3 points of success while reordering or reversing directions barely matter (Trinh et al., 2026a); our object swaps leave DeepThinkVLA's success unchanged, which may reflect the policy, its training or edit coverage. Reasoning-based safety monitors fail under adaptive attack (Trinh et al., 2026b); both works score task success or detection, not action magnitude. Huang et al. (2026) hijack the CoT with adversarial patches and read the action change as control by the reasoning, Wu et al. (2025) correct reasoning–action mismatch at runtime, Foutter et al. (2026) train a faithfulness critic as a reward for driving policies where we audit edit scores, and Nguyen et al. (2026)'s magnitude protocol separates a causal from an epiphenomenal CoT in driving VLAs with continuous heads; the choice of control matters for any such protocol. *Construct validity.* Whether a reported number measures what its name says has been questioned for emergent abilities (Schaeffer et al., 2023) and interpretability metrics, which pair a perturbation score with an explicit null (DeYoung et al., 2020; Hooker et al., 2019); our floors extend that design to action space. More is in Appendix A.

2 The CoT-Faith Benchmark

We distinguish two levels. *Step-level sensitivity* is whether the decoded action at a given state depends on the CoT's content; open-loop edit scores measure this. *Outcome-level reliance* is whether task success depends on the reasoning; this is what CoT-based monitoring and interpretability claims need. At the step level, we take a VLA's emitted CoT to be faithful to the extent that the decoded action depends counterfactually on the CoT's semantic content, and only on that content. Three conditions follow: a meaning-changing edit must move the action more reliably than a meaning-preserving one, an edit that reverses an implied direction must reverse the action, and every score must be read against a per-model null. A negative result below means an instrument is invalid, not that the CoT is unused.

2.1 Measurement Protocol

For each sample s and edit family f we decode greedily twice, once on the unedited CoT and once on the edited CoT, holding the observation, instruction and decode fixed, and compare the two 7-DoF actions:

$$\Delta_{\infty}^{s,f} = \max_{i \in [7]} \big| \, a_i(s, f(\text{cot}_s)) - a_i(s, \text{cot}_s) \, \big|, \tag{1}$$

with a_i normalized to $[-1, 1]$. The magnitude score is the response rate at $\tau = 0.05$,

$$\mathcal{F}(m, f) = \tfrac{1}{N} \textstyle\sum_s \mathbf{1}[\Delta_{\infty}^{s,f}(m) > \tau], \tag{2}$$

which we also write $\mathcal{F}_{\text{mag}}$. It is a paired, within-policy, within-observation intervention; it does not presuppose that the policy solves the task, and licenses no claim about one that does. Eq. 1 credits movement in any direction. For *direction_flip*, which reverses left↔right, up↔down and in↔out, we also require the sign:

$$\mathcal{F}_{\text{dir}} = \tfrac{1}{N} \textstyle\sum_s \mathbf{1}\big[c_s^f < -0.5\big], \tag{3}$$
$$c_s^f = \cos\big(a_{1:3}(s, \text{cot}_s),\, a_{1:3}(s, f(\text{cot}_s))\big),$$

with analogous criteria for *gripper_flip* and *negation* (Appendix B.6.2). The primary sweep teacher-forces the CoT annotation, so the reasoning text is fixed except for the edit, and scores the first step of 100 LIBERO-90 observations per family and sampling seed; self-generated CoT and random steps are tested separately.

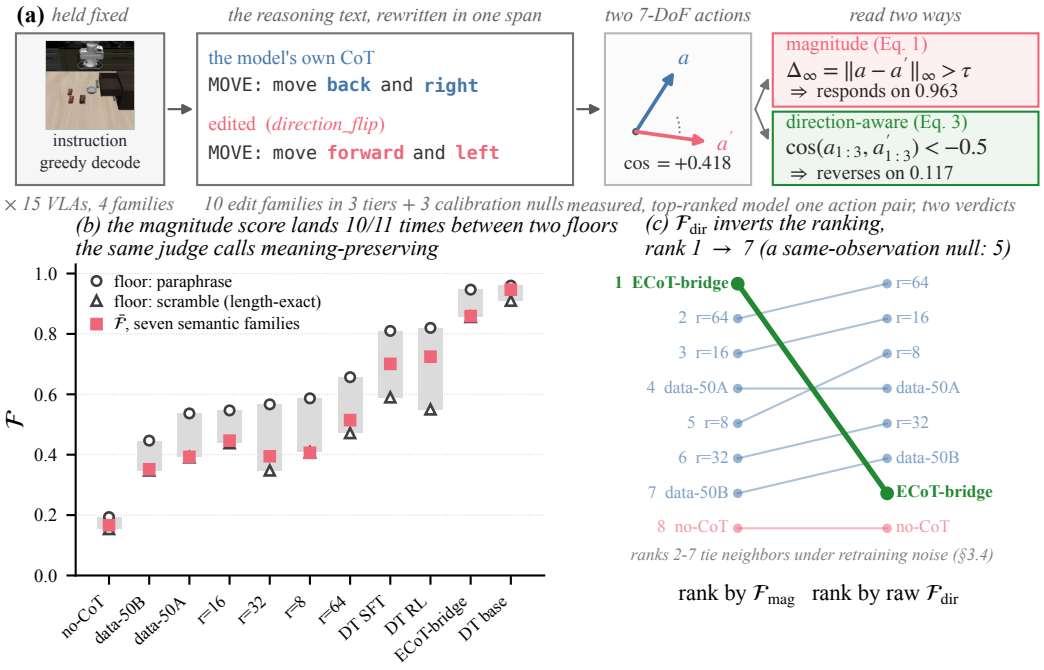

Figure 1: The protocol and the two disagreements this paper reports. *(a)* Observation, instruction and decode are held fixed, one span of the CoT is edited (here *direction_flip* on the `MOVE` line), and the two 7-DoF actions are compared. The frame is an illustrative LIBERO frame from the same task; the text-only CoT corpus stores no image for this pair. Arrows show `ECoT-bridge`'s mean translation cosine after the flip: it moves the same way (§3.3). *(b)* Per configuration, the two meaning-preserving floors (grey band) and the semantic mean, which lies between the floors on 10 of 11: the floor choice decides the sign (§3.1). *(c)* $\mathcal{F}_{\text{mag}}$ rank against raw $\mathcal{F}_{\text{dir}}$ rank on *direction_flip* (§3.3). Both ranks are on *direction_flip*, not the family-mean ordering of Table 10. Adjacent ranks 2–7 differ by less than *direction_flip*'s retraining movement (§3.6, measured on $\mathcal{F}_{\text{mag}}$, not $\mathcal{F}_{\text{dir}}$).

2.2 Edit Taxonomy

Thirteen families, grouped by what they probe (definitions and examples in Appendix B.1.1). *Semantic families* (seven), in two tiers: word-level edits (*direction_flip*, *gripper_flip*, *verb_swap*, *negation*) and object or causal-chain edits across tags (*subject_swap*, *location_swap*, *adversarial_plausible*). *Meaning-preserving CoT floors* (two): *paraphrase_null* (synonym substitution) and *syntactic_scramble* (length-exact word-order shuffle). *Further nulls* (two): *bbox_jitter_null* (bounding-box jitter) and *instr_random_sub* (random instruction tokens, outside the CoT). An *identity control*, *selfsplice_control*, which must score 0 and does on 11 of 11 configurations, and a *ceiling*, *cross_task_swap*, which replaces the whole CoT with another task's. We add two *size-matched nulls*: *paraphrase_null_matched_dir* substitutes synonyms in the tags *direction_flip* edits, at exactly as many word positions, and *paraphrase_null_matched_neg* makes *negation*'s two-word insertion at the same position ("you should" for "do not"). The families map onto six capability axes (Table 8).

2.3 Controls, Judge Validation and Statistics

A blind LLM judge (`Qwen2.5-7B-Instruct` (Qwen et al., 2024), from neither architecture family under test, following Zheng et al. (2023); Appendix B.4.6), gated on identity and negative controls, validates the generators; the judge scores 12 of the 13 families, all but *instr_random_sub*. It finds that *paraphrase_null* preserves meaning at 0.975, read against an 0.824 order-flip agreement floor, and rates both size-matched nulls meaning-preserving on every judged pair. It flags *verb_swap* as often meaning-preserving; dropping it changes no sign (Table 25). A second judge fails its own negative-control gate, so we do not treat it as corroboration.

Table 1: Models (per-model listing: Table 9). The 11 configurations in families B–D are scored. The 4 OpenVLA baselines have no CoT to edit and only validate the rollout harness. Base architectures: OpenVLA-7B (A, C, D) and PaliGemma (B).

Family	Base architecture	N	Reasoning target
A: OpenVLA non-CoT baselines	OpenVLA-7B	4	no CoT
B: DeepThinkVLA (base/SFT/RL)	PaliGemma	3	none / CoT (SFT) / CoT (RL)
C: ECoT-bridge (public)	ECoT-OpenVLA-7B	1	full 9-tag CoT
D: Ours (LoRA on ECoT-bridge)	ECoT-bridge	7	full CoT (no-CoT: action-only)

Primary statistic. For each semantic family with a size-matched null, the primary statistic is the paired difference $\mathcal{F}(f) - \mathcal{F}(\text{null}_f)$ on the samples where both apply, with an observation-clustered BCa bootstrap interval and Holm correction across the 11 configurations. The matched nulls were constructed and judged before the sweep; we adopted the matched contrast as primary after seeing it, because it is the comparison that controls edit size. It could fail: if the score measured only edit size, a semantic edit and a null of the same size should score alike. We also report the semantic mean $\bar{\mathcal{F}}$ over the seven semantic families against each unmatched floor, $\bar{\mathcal{F}}_{\text{diff}} = \bar{\mathcal{F}} - \mathcal{F}(\text{floor})$, as a test of whether a floor verdict is identified. Unmatched floor tests use the same paired, observation-clustered bootstrap (20,000 resamples of the shared observations), with Holm correction within each table.

2.4 Metrics

Besides $\mathcal{F}_{\text{mag}}$ (Eq. 2) and $\mathcal{F}_{\text{dir}}$ (Eq. 3), we report the exact-no-change rate $P(\Delta_\infty{=}0)$ beside every score, and a threshold-free divergence, $D_{\text{TV}}^{s,f} = \frac{1}{7}\sum_{d=1}^{7} \text{TV}\big(P_{\text{orig}}^{d}, P_{\text{edit}}^{d}\big)$, the total-variation distance between the policy's softmax over action bins under the original and edited CoT, averaged over dimensions. $\mathcal{F}_{\text{dir}}$ is read against three nulls, including the *family ceiling*, its largest value over families with no direction to reverse (Appendix B.6.2). In closed loop we report the change in task success, ΔSR, of an edited arm against the clean arm on the same episodes, with a cluster bootstrap over tasks and an exact McNemar test, and the divergence between the two arms' end-effector trajectories.

2.5 Models

We score 11 configurations (Table 1): the public ECoT-bridge, trained on Bridge V2 (Walke et al., 2023) from Open X-Embodiment (Collaboration et al., 2024); seven LoRA (Hu et al., 2022) fine-tunes of it on LIBERO-90 (Liu et al., 2023) varying rank, data split and training target; and three DeepThinkVLA (Yin et al., 2025) stages on a PaliGemma (Beyer et al., 2024) base. Two are controls without a CoT target: our action-only `no-CoT` LoRA and the pretrain-only `DT base`, which responds to almost every edit and is uninformative where it enters a count. They form two base-checkpoint lineages, not 11 independent replications. Four OpenVLA (Kim et al., 2024) LIBERO checkpoints only validate the rollout harness. Most policies do not solve their tasks: the LoRA variants reach $0/20$–$0/40$ closed-loop success, and ECoT-bridge is off its training distribution on LIBERO, so we also score it on Bridge V2 with its own normalization. DeepThinkVLA-SFT/RL are competent (closed-loop LIBERO success 0.965/0.960).

3 Experiments

3.1 The Magnitude Verdict Depends on the Control

A faithful model should move less under a meaning-preserving edit than under a meaning-changing one, so the semantic mean should exceed a meaning-preserving floor. Which floor? We have two that the same judge certifies: *paraphrase_null* and *syntactic_scramble*. They give opposite verdicts on the strongest models we have. On the two competent LIBERO policies, rerun at three seeds, the semantic mean is significantly below the paraphrase floor and significantly above the scramble floor: $\bar{\mathcal{F}}_{\text{diff}} = -0.088$ and $+0.147$ for `DT SFT`, -0.079 and $+0.169$ for `DT RL`. On ECoT-bridge in distribution on Bridge V2, its own training corpus, it is below the paraphrase floor in 4 of 4 runs and above the scramble floor in 4 of 4 (Table 3). The contradiction is therefore not an artifact of weak models. It holds across the cohort: under teacher-forced first-step scoring $\bar{\mathcal{F}}_{\text{diff}}$ is negative against the paraphrase floor on 11 of 11 configurations and positive against the scramble floor on 10 of 11, so the

Model	$\bar{\mathcal{F}}$	vs. paraphrase floor	vs. paraphrase Δ_{diff}	vs. scramble floor	vs. scramble Δ_{diff}
no-CoT	0.166	0.193	−0.027	0.154	+0.012
r=8	0.406	0.587	−0.180*	0.408	−0.002
r=16	0.447	0.547	−0.100*	0.438	+0.008
r=32	0.395	0.567	−0.172*	0.348	+0.047*
r=64	0.515	0.657	−0.142*	0.472	+0.043
data-50A	0.393	0.537	−0.144*	0.391	+0.002
data-50B	0.352	0.447	−0.095*	0.348	+0.004
ECoT-bridge	0.859	0.947	−0.088*	0.856	+0.003
DT base	0.947	0.960	−0.013	0.910	+0.037
DT SFT	0.701	0.810	−0.109*	0.590	+0.111*
DT RL	0.725	0.820	−0.095*	0.550	+0.175*
all 11 configurations			− 11/11		+ 10/11

Table 2: The semantic mean against each meaning-preserving floor. $\bar{\mathcal{F}}$ is the mean over the seven semantic families. *: 95% paired-bootstrap interval excludes zero (12 of 22 cells uncorrected). Both floors are judged meaning-preserving (0.975, 1.000). The signs agree only on r=8. A 4k Bridge-subset checkpoint calibrates but is excluded: its faithful samples move only the gripper bit, not comparable in scale.

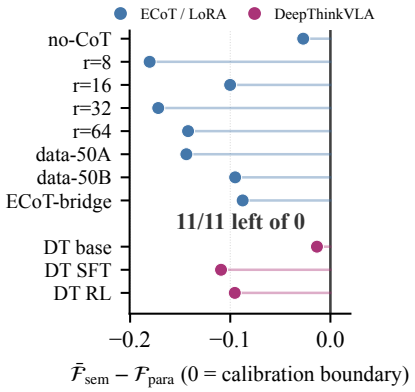

Figure 2: $\bar{\mathcal{F}}_{\text{diff}}$ against the paraphrase floor (Table 2), by lineage, teacher-forced first step. All 11 lie left of zero; against the scramble floor, 10 of 11 lie right of it.

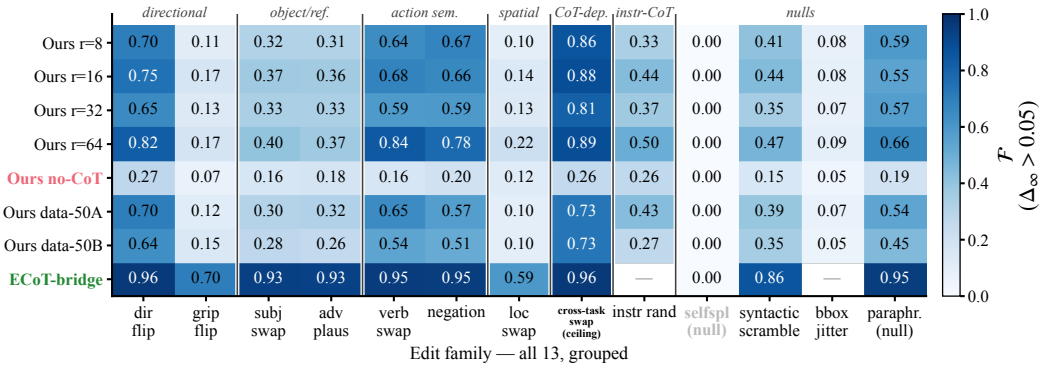

Figure 3: $\mathcal{F}$ at τ=0.05 for the eight three-seed configurations × 13 edit families, grouped by capability axis (Appendix B.1.1), with the four meaning-preserving and identity controls at right. *cross_task_swap* (bold) is the ceiling. ECoT-bridge's two empty cells were measured in a separate single-seed run (0.460/0.990, Table 15); they are not zero.

floors agree in sign only on r=8 (Table 2). By a paired, observation-clustered bootstrap, 9 of 11 are significant against the paraphrase floor (7 of 8 ECoT-lineage, 2 of 3 DeepThinkVLA-lineage) and 3 of 11 against the scramble floor; after Holm correction over the 22 tests of Table 2, 7 and 1 remain (all ECoT lineage, and DT RL; Appendix B.4.4). With the three-seed DeepThinkVLA rows in place of the single-seed ones, the Holm counts become 9 of 11 and 2 of 11. The same model thus looks unfaithful or faithful depending on the control the analyst picks. The instability recurs task by task (Table 16), across family sets (Appendix B.4.5), on-policy, where the paraphrase-floor sign reverses (§3.4), and against an instruction-side null, which moves the action more than the semantic CoT edits on 6 of 11 configurations (Table 15). We therefore report both floors everywhere and no single-floor verdict as a finding about faithfulness.

Self-generated reasoning. Every number above edits a teacher-forced annotation. On the model's own free-decoded CoT the paraphrase-floor comparison replicates on ECoT-bridge and DT-RL, points the same way on DT-SFT without reaching significance, reverses on r=32, depends on the

Table 3: The contradiction on the strongest models: $\bar{\mathcal{F}}_{\text{diff}}$ against each meaning-preserving floor, with BCa 95% intervals, for the two competent LIBERO policies at three seeds and ECoT-bridge in distribution on Bridge V2 (400 observations per run). Full tables: Table 17 and Table 13.

	vs. *paraphrase_null*		vs. *syntactic_scramble*	
Model, data	Δ	95%	Δ	95%
DT-SFT, LIBERO, 3 seeds	-0.088	$[-0.133, -0.040]$	$+0.147$	$[+0.093, +0.205]$
DT-RL, LIBERO, 3 seeds	-0.079	$[-0.122, -0.033]$	$+0.169$	$[+0.113, +0.226]$
ECoT-bridge, Bridge V2, self-generated	-0.096	$[-0.131, -0.061]$	$+0.052$	$[+0.008, +0.099]$
ECoT-bridge, Bridge V2, TF, first step, seed 0	-0.077	$[-0.128, -0.031]$	$+0.040$	$[-0.016, +0.094]$
ECoT-bridge, Bridge V2, TF, first step, seed 1	-0.072	$[-0.120, -0.027]$	$+0.074$	$[+0.020, +0.126]$
ECoT-bridge, Bridge V2, TF, random step	-0.056	$[-0.090, -0.023]$	$+0.081$	$[+0.033, +0.131]$

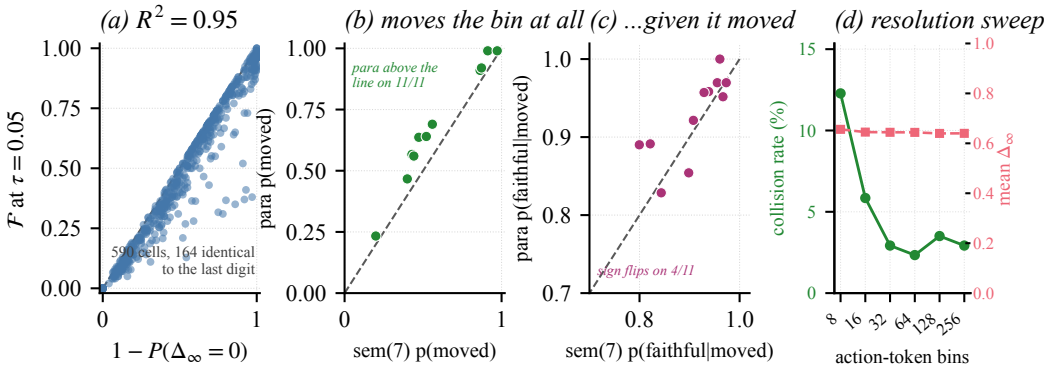

Figure 4: What $\mathcal{F}$ counts. *(a)* Over the 590 (run, family) cells with $n \geq 20$, $\mathcal{F}$ at τ=0.05 against $1 - P(\Delta_\infty{=}0)$: best-fit $R^2 = 0.946$; against the drawn $y{=}x$ line, $R^2 = 0.923$. *(b)* P(moved) for *paraphrase_null* against the pooled seven semantic families, one point per configuration: the paraphrase moves some argmax bin more often on all 11. *(c)* $P(\Delta_\infty{>}\tau \mid \text{moved})$ for the same pair: close to the diagonal, with the sign flipping on 4 of 11. *(d)* Regrouping `ECoT-bridge`'s 256-bin softmax into 8 to 256 bins changes the collision rate from 12.3% at 8 bins to 2.3%–3.5% from 32 bins on (2.9% at the native 256), while the mean continuous Δ_∞ stays at 0.640–0.656 (n=171 pooled over four families).

statistic on `r=64` and is inconclusive on `no-CoT` (Appendix B.6.1). The teacher-forced verdict does not reliably carry over.

3.2 What the Score Measures

Why would two meaning-preserving edits disagree? Because the score mostly records whether the action changes at all. 46.0% of edits leave the decoded action *exactly* unchanged, and almost every edit that changes it moves it past τ (the conditional rate has median 0.905; Table 23). $\mathcal{F}$ is therefore close to $1 - P(\Delta_\infty{=}0)$ ($R^2 = 0.946$ over 590 cells; Figure 4a), and τ carries little information. The same holds for ECoT-bridge in distribution on Bridge V2 ($R^2 = 0.961$–0.992). $P(\Delta_\infty{=}0)$ is computed for all 590 cells. Actions are per-dimension bins for the OpenVLA-based models and FAST tokens (Pertsch et al., 2025) for DeepThinkVLA, scored on the first action of each chunk. Most of the gap between the paraphrase floor and the semantic mean is this movement rate, not the size of movements that do occur (Figure 4b–c; Table 21), and per-family movement rate tracks which semantic families clear the floor (Spearman $\rho = 0.873$; Appendix B.4.5).

Whether the action changes depends on what the edit touches. The paraphrase floor changes 10.6 words on average, against 3.8 for *direction_flip* (Table 28), but word count is not the whole story: a length-exact paraphrase scores like *paraphrase_null* and far above *syntactic_scramble*, and a four-word in-vocabulary reorder moves the LoRA policies' actions more than any semantic edit, suggesting they key on the exact `MOVE` string. Edit type, token identity and position all matter, and we cannot yet disentangle them. Neither the policy's likelihood of the edited CoT nor judged fluency explains the gap (Appendix B.4.6).

	no-CoT	r=8	r=16	r=32	r=64	50A	50B	ECoT-br.	DT base	DT SFT	DT RL
direction_flip − matched	$+0.142^*$	$+0.092^*$	$+0.288^*$	$+0.124^*$	$+0.150^*$	$+0.223^*$	$+0.177^*$	$+0.054^*$	$+0.149^*$	$+0.190^*$	$+0.182^*$
negation − matched	-0.010	$+0.220^*$	$+0.207^*$	$+0.243^*$	$+0.293^*$	$+0.183^*$	$+0.143^*$	$+0.047$	$+0.020$	$+0.003$	$+0.020$
$\bar{\mathcal{F}}$ − *paraphrase_null*	-0.029	-0.184^*	-0.105^*	-0.174^*	-0.144^*	-0.145^*	-0.095^*	-0.096^*	-0.010	-0.086^*	-0.076^*

Table 4: Primary statistic: each semantic family against its size- and tag-matched null, paired on the samples where both apply, with the unmatched paraphrase comparison on the same records for contrast. *: Holm-significant across the 11 configurations (observation-clustered BCa bootstrap, 20,000 draws). Three seeds of 100 samples each, except DT base (one seed). Intervals, rates and the decomposition are in Table 29.

	magnitude			direction-aware				family ceiling		
Model	$\mathcal{F}_{\text{mag}}$	±sd	rank	$\mathcal{F}_{\text{dir}}$	±sd	cos	N	value	family	×
ECoT-bridge	0.963	.006	1	0.117	.032	$+\mathbf{0.418}$	299	0.047	*cross_task*	2.5
r=64	0.823	.016	2	0.779	.021	-0.510	299	0.070	*verb_swap*	11.1
r=16	0.749	.028	3	0.659	.046	-0.309	299	0.093	*verb_swap*	7.1
data-50A	0.699	.022	4	0.642	.004	-0.261	299	0.097	*cross_task*	6.6
r=8	0.696	.041	5	0.649	.020	-0.313	299	0.077	*cross_task*	8.5
r=32	0.652	.024	6	0.582	.034	-0.147	299	0.073	*cross_task*	7.9
data-50B	0.635	.017	7	0.532	.013	-0.097	299	0.057	*cross_task*	9.4
no-CoT	0.274	.032	8	0.087	.021	$+\mathbf{0.760}$	299	0.113	*cross_task*	**0.8**
DT RL	0.640	1 seed	n/a	0.010	1 seed	$+0.915$	100	0.020	*paraphrase*	**0.5**
DT SFT	0.650	1 seed	n/a	0.010	1 seed	$+0.902$	100	0.050	*negation*	**0.2**
DT base	0.970	1 seed	n/a	0.000	1 seed	$+0.900$	100	0.000	*paraphrase*	n/a

Table 5: $\mathcal{F}_{\text{mag}}$ against $\mathcal{F}_{\text{dir}}$ on *direction_flip*, and $\mathcal{F}_{\text{dir}}$ against its family ceiling, the largest $\mathcal{F}_{\text{dir}}$ over eight families with no direction to reverse, computed on each checkpoint's own de-quantization grid. cos is the mean original-vs-edited translation cosine, so reversal is negative. Rows are 3-seed means ± seed s.d., with N pooled across the three sampling seeds; DeepThinkVLA rows are the single-seed canonical runs (their three-seed reruns give $\mathcal{F}_{\text{mag}}$ 0.643/0.616 for SFT/RL on a partly different sample set, Table 29). × is $\mathcal{F}_{\text{dir}}$ over the ceiling. Seven rows above the rule clear; the four below do not, and no-CoT sits below its own ceiling, as a control should. Rank orders $\mathcal{F}_{\text{mag}}$ within the eight three-seed configurations. DT base's ratio is $0/0$.

We call an exact zero a collision as a description, not a cause. The total-variation divergence D_{TV} has no argmax and no threshold, yet on the six configurations rerun with softmax outputs it shows the same floor dependence (Table 12; after Holm correction, 2 of 6 and 5 of 6, so the evidence against the paraphrase floor alone is weak after correction). Nor is the dependence an artifact of broken models producing exact zeros. Exact zeros concentrate in the weak LoRA policies and are rare in ECoT-bridge and DeepThinkVLA (Table 24), which show the same floor dependence; where zeros occur they may be genuine non-response rather than quantization artifacts. Regrouping ECoT-bridge's own 256 action bins into coarser grids changes its collision rate while its continuous displacement barely moves (Figure 4d); we have not shown that resolution changes the floor verdict.

3.3 Size-Matched Controls and a Direction-Aware Diagnostic

If the verdict depends on what an edit touches, the control should touch the same thing. Against nulls matched in size and tag (Table 4), *direction_flip* exceeds its matched null on 11 of 11 configurations, all Holm-significant, including ECoT-bridge and all three DeepThinkVLA checkpoints. *negation* is positive on 10 of 11 and Holm-significant on the six CoT-trained LoRA variants; elsewhere both rates are near ceiling or the policy has no CoT target. On ECoT-bridge in distribution the contrast is positive in every run for both families (Holm-significant in 2 and 3), but these contrasts are weak evidence, because on Bridge V2 the identity control itself is not zero. On the same records the semantic mean still falls below the unmatched paraphrase floor on all 11. The matched contrast is necessary, not sufficient. It shows that direction words move the action more than size-matched synonyms, but it cannot separate reading the reasoning from lexical priming: the action-only no-CoT control, trained without a CoT target, clears the direction null too ($+0.142$). That control was fine-tuned from the CoT-trained ECoT-bridge and may inherit direction-word sensitivity. The contrast is still a magnitude-style score: a median 66% of the *direction_flip* gap is movement rate (Appendix B.4.8).

Table 6: Closed-loop success on the two competent policies: change in success rate against the clean arm over 400 paired episodes (all four LIBERO suites), with cluster-bootstrap 95% intervals. *cov.*: fraction of re-plans at which the edit applies (RL). †Applied at every re-plan by construction. All arms: Table 36.

		DT-SFT		DT-RL	
Arm	cov.	ΔSR	95%	ΔSR	95%
clean (success rate)	—	0.965		0.960	
cross-task CoT (positive control)	1.00	−0.158	[−0.253, −0.075]	−0.158	[−0.250, −0.075]
empty CoT	1.00†	+0.003	[−0.020, +0.025]	+0.000	[−0.015, +0.018]
negation	0.94	+0.005	[−0.013, +0.025]	+0.005	[−0.015, +0.025]
object_swap	0.91	+0.005	[−0.013, +0.022]	+0.022	[+0.003, +0.045]
direction_flip	0.08	+0.000	[−0.022, +0.022]	−0.003	[−0.020, +0.015]
paraphrase_null	0.93	−0.015	[−0.040, +0.010]	−0.003	[−0.025, +0.020]
syntactic_scramble	0.94	+0.003	[−0.020, +0.028]	−0.007	[−0.028, +0.013]

Moving is also not following. We score whether the action reverses after *direction_flip* (Eq. 3, Table 5). ECoT-bridge, the flagship public reasoning VLA, responds to 0.963 of flips on LIBERO but reverses on only 0.117 (mean cosine +0.418): after "left" becomes "right", it moves the *same* way. In distribution on Bridge V2 the same holds: $\mathcal{F}_{\text{dir}}$ is 0.075–0.274 with mean cosine +0.31 to +0.66 and no excluded samples (Appendix B.4.1), so the non-reversal is not an axis-convention or off-distribution artifact. The six CoT-trained LoRA variants reverse, but their action spaces are collapsed, and both competent DeepThinkVLA policies fail. Raw $\mathcal{F}_{\text{dir}}$ moves the top-ranked model from first of eight to seventh, and the two rankings correlate only weakly ($\rho = 0.476$). ECoT-bridge's low rank does not depend on the calibration against its three nulls, the *family ceiling*, an edited-permutation null and an unedited-pair null (Table 34), and its non-reversal persists when the unedited scene geometry is blanked or mirrored (Appendix B.6.2). $\mathcal{F}_{\text{dir}}$ shows that the two scoring rules disagree, not which is right, so we report it as a diagnostic beside $\mathcal{F}$.

3.4 Closed-Loop Evaluation on Competent Policies

The question the paper started from is whether the reasoning drives behaviour, so we test it where behaviour is meaningful: on the two competent policies, which succeed on 0.96 of episodes, in closed loop on all four LIBERO suites, with 400 paired episodes per arm and the edit applied at every re-plan where its target is present (coverage in Table 6; Appendix B.6.3). The rollouts can detect an effect: replacing the CoT with another task's lowers success by 0.158 on both policies (McNemar $p < 10^{-12}$). Against that positive control, the reasoning's content does not matter. An empty CoT, applied at every re-plan, leaves success unchanged, and negation and object swaps, which apply at $\geq 91\%$ of re-plans, do not lower it; every within-task edit and every null stays within 0.022 of clean, with 95% intervals reaching −0.040. These policies do not need their reasoning to succeed on LIBERO, consistent with Chen et al. (2025a); the cross-task drop may be distraction by an out-of-distribution trace rather than reliance on the right plan. On these policies' own states the open-loop controls agree that the action is sensitive to the reasoning at each step: at the same re-plans the semantic edits move the action ($\mathcal{F} = 0.23$–0.58), and the semantic mean lies *above* all three unmatched floors (+0.083 to +0.132), unlike under teacher-forced scoring. Yet the same edits, and even an empty CoT, leave closed-loop success unchanged. Sensitivity is not reliance: step-level sensitivity does not imply outcome-level reliance, possibly because the closed loop corrects perturbed steps, which we do not isolate, so open-loop scores are not evidence of reliance. Across edit families $\mathcal{F}$ ranks trajectory divergence at Spearman 0.98, which is close to tautological since both measure action change; its correlation with the loss of success (−0.42, −0.37) is over effects that are all within noise, so it is no evidence of prediction.

3.5 Continuous-Head Models

We also probe two continuous-head VLAs, open loop (Table 7; Appendix B.4.7). InstructVLA has a flow-matching head; we score it in distribution on Bridge V2 with the flow noise shared across each pair, so the identity edit scores ≤ 0.0002. No edit leaves its action unchanged, and all three floors agree in sign, but its floors change about as many words as its semantic edits (1.0–3.3 against 1.0–2.8), so the agreement does not isolate head type. Resampling the noise alone gives $\mathcal{F} = 0.99$,

so such heads need fixed noise. On MolmoAct2-Think-LIBERO depth-reasoning edits move the continuous-head action by at most 0.001, never past τ, unlike an instruction substitution: its depth reasoning is not load-bearing, a degenerate case rather than a calibration.

Table 7: Continuous-head probes (open loop; Appendix B.4.7, Tables 26 and 27). InstructVLA scores semantic mean minus each of three meaning-preserving floors; MolmoAct2 scores $\mathcal{F}$ over its depth-reasoning edits.

Model (format)	Head	$P(\Delta_\infty{=}0)$	$\bar{\mathcal{F}}-$ floors / $\mathcal{F}$ (reasoning edits)	instr. sub.	noise only
InstructVLA (primitive)	flow	0.00	$+0.64$ to $+0.92$	0.91–0.99	0.99
InstructVLA (free text)	flow	0.00	$+0.19$ to $+0.53$		0.99
MolmoAct2	flow	0.00	$\mathcal{F} = 0.000$	0.279	—
MolmoAct2	FAST	0.84–0.97	$\mathcal{F} = 0.005$–0.025	0.455	—

3.6 Retraining Variance

Every three-seed row carries Wilson 95% intervals, but sampling seeds hold the checkpoint fixed, so we retrained each of our seven LoRA configurations at the same seed and hyperparameters (six retrains have a full 13-family edit run; Appendix B.6). Per-family $\mathcal{F}$ moves by a median 0.037 (IQR 0.017–0.085 over 78 family–pair cells) and up to $\mathbf{0.317}$, which is about $4.8\times$ that widest Wilson half-width (*verb_swap* on `r=8`, the worst of 13 families), and $\bar{\mathcal{F}}_{\text{diff}}$ against the paraphrase floor by up to 0.068, more than most between-model gaps. Adjacent ranks 2–7 of Table 5 are therefore not informative. The bar measures fixed-seed noise only and is not measured for ECoT-bridge or DeepThinkVLA.

Positive control. To check that the pipeline detects a dependence when one exists, we score a rule-based oracle whose action is a hard-coded function of the `MOVE` text, on the judge-validation pairs (Appendix B.4.9). It scores $\mathcal{F}_{\text{dir}} = 0.925$ on *direction_flip* and 0.000 on *paraphrase_null*. This shows the pipeline detects a dependence that is built in; it does not calibrate any learned policy.

4 Conclusion

Editing a reasoning VLA's chain-of-thought and counting how often the action moves does not measure whether the reasoning drives the action. On token-decoded VLAs the score mostly records whether the action changes at all, so the analyst's choice of control sets the verdict. Controls matched to each edit in size and tag fix the sign, but a model without a CoT target passes them too, and moving is not following. On competent policies the open-loop controls agree that the action is sensitive to the reasoning, yet editing or even emptying it leaves closed-loop success unchanged: sensitivity is not reliance.

For edit-based faithfulness scores, our results suggest comparing each semantic family against a meaning-preserving null that edits the same tags and number of words; reporting the exact-no-change rate $P(\Delta_\infty{=}0)$ beside every score, with sampling noise shared across each pair on stochastic heads; adding a direction-aware score with its own null for signed edits; using a same-recipe retraining replicate as the error bar for between-model gaps; and, wherever the policy is competent, measuring closed-loop success under the same edits, including an empty CoT (Stages 1–5 and 6b of the checklist in Appendix B.1.2). Open-loop scores are cheap, but only closed-loop success on a competent policy answers whether a policy relies on its reasoning, so open-loop instruments should be validated against it. The next step is a known-ground-truth extension of the benchmark, where a valid score must separate policies trained to use their reasoning from policies trained to ignore it.

Limitations. The token-decoded cohort is mostly weak, so the claims concern the instrument, not any policy; open-loop scores are first-step and teacher-forced. The matched contrast was adopted as primary after the sweep, and it cannot separate reading from lexical priming; a token-identity-controlled null is future work. On Bridge V2 the identity control is not zero, so the in-distribution rate contrasts there are weak. In closed loop, *direction_flip* and *gripper_flip* apply at only 8% and 38% of re-plans, clean success near ceiling leaves little headroom, and a same-task, other-episode CoT was not run. Continuous heads are only probed (§3.5), and a dose–response over edit sizes $k{=}1, 2, 4, 8$, a resolution sweep of the floor verdict and the known-ground-truth test are not yet run.

AI USE STATEMENT

In this work, we used generative AI tools to help write code and to check the grammar of the paper's text. We did not use generative AI for research ideation, experimental design, or the interpretation of results, and all AI-assisted code is checked against the raw per-sample data (2,057 assertions against the raw data, non-zero exit on any mismatch) rather than trusted on its own output. We take responsibility for the final content of this work.

ETHICS STATEMENT

The motivation is a safety concern: CoT-based monitors and CoT-edit interventions are proposed for robots on the assumption that the reasoning is load-bearing, and our results indicate that a common test of that assumption largely records whether the decoded action changes at all, not whether success depends on the reasoning. All experiments are in simulation (LIBERO) or on public offline corpora: no physical robot was operated, and no human subjects or personal data were involved. Upstream licenses for every asset are resolved from the Hugging Face Hub API; 5 of the 19 assets declare no license upstream; we assume none for them, use them for research evaluation only and redistribute none. No open-loop score here is conditioned on task success; closed-loop success is reported separately for the two competent DeepThinkVLA policies. The blind judge that validates both floors is an LLM run locally as a method component, under a fixed prompt gated on two controls (§2).

REPRODUCIBILITY STATEMENT

Every number in this paper is derived from per-sample records by a script, and an audit script that asserts 2,057 claims re-derives each one from those records rather than from the printed text, exiting non-zero on any mismatch (Appendix B.5). The judge prompt, seed and both presentation orders are fixed; the judge verdicts are not reproducible token-for-token because the judge weights are not pinned to a specific commit, so a re-run should match the reported rates to within the order-flip noise floor (§2).

REFERENCES

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps. *NeurIPS*, 2018. arXiv:1810.03292.

I. Arcuschin, J. Janiak, R. Krzyzanowski, S. Rajamanoharan, N. Nanda, and A. Conmy. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. arXiv:2503.08679, 2025.

P. Atanasova, O. Camburu, C. Lioma, T. Lukasiewicz, J. G. Simonsen, and I. Augenstein. Faithfulness Tests for Natural Language Explanations. *ACL*, 2023. arXiv:2305.18029.

O. Bentham, N. Stringham, and A. Marasović. Chain-of-Thought Unfaithfulness as Disguised Accuracy. *TMLR*, 2024. arXiv:2402.14897.

L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, et al. PaliGemma: A versatile 3B VLM for transfer. arXiv:2407.07726, 2024.

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *CoRL*, 2023. arXiv:2307.15818.

C.-S. Chan, H. Kong, and G. Liang. A Comparative Study of Faithfulness Metrics for Model Interpretability Methods. *ACL*, 2022.

W. Chen, S. Belkhale, S. Mirchandani, O. Mees, D. Driess, et al. Training Strategies for Efficient Embodied Reasoning. arXiv:2505.08243, 2025a.

Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, et al. Reasoning Models Don't Always Say What They Think. arXiv:2505.05410, 2025b.

O. X. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. *ICRA*, 2024. arXiv:2310.08864.

R. Cui, Z. Zhang, J. Pang, H. Chi, J. Guo, et al. LIBERO-Safety: A Comprehensive Benchmark for Physical and Semantic Safety in Vision-Language-Action Models. arXiv:2606.23686, 2026.

M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, et al. The Benchmark Lottery. arXiv:2107.07002, 2021.

J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. ERASER: A Benchmark to Evaluate Rationalized NLP Models. *ACL*, 2020.

J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith. Show Your Work: Improved Reporting of Experimental Results. *EMNLP*, 2019. arXiv:1909.03004.

S. Fei, S. Wang, J. Shi, Z. Dai, J. Cai, P. Qian, L. Ji, X. He, S. Zhang, Z. Fei, J. Fu, J. Gong, and X. Qiu. LIBERO-Plus: In-depth Robustness Analysis of Vision-Language-Action Models. arXiv:2510.13626, 2025.

M. Foutter, M. Cercola, L. Wild, Y. Wang, M. Li, et al. Do vision-language-action models mean what they say? On the role of faithfulness in embodied reasoning. arXiv:2607.04681, 2026.

D. Friedman, A. Lampinen, L. Dixon, D. Chen, and A. Ghandeharioun. Interpretability Illusions in the Generalization of Simplified Models. *ICML*, 2024. arXiv:2312.03656.

T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

A. Geiger, Z. Wu, H. Lu, J. Rozner, E. Kreiss, T. Icard, N. D. Goodman, and C. Potts. Inducing Causal Structure for Interpretable Neural Networks. *ICML*, 2022.

Y. Gur-Arieh, A. Marasović, and M. Geva. Faithfulness Metrics Don't Measure Faithfulness: A Meta-Evaluation with Ground Truth. arXiv:2605.25052, 2026.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep Reinforcement Learning that Matters. *AAAI*, 2018. arXiv:1709.06560.

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. *NeurIPS*, 2019.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022. arXiv:2106.09685.

Z. Huang, W. Zhu, H. Qiu, X. Ji, and W. Xu. TRAP: Hijacking VLA CoT-Reasoning via Adversarial Patches. *ICML*, 2026. arXiv:2603.23117.

A. Z. Jacobs, and H. Wallach. Measurement and Fairness. *FAccT*, 2021. arXiv:1912.05511.

A. Jacovi, and Y. Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *ACL*, 2020.

S. Jain, and B. C. Wallace. Attention is not Explanation. *NAACL*, 2019. arXiv:1902.10186.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, et al. Mistral 7B. arXiv:2310.06825, 2023.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *CoRL*, 2024. arXiv:2406.09246.

M. J. Kim, C. Finn, and P. Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. arXiv:2502.19645, 2025. Referred to as OpenVLA-OFT.

C. Kim, M. Kim, M. Kang, H. Kim, and D. Jung. LIBERO-Para: A Diagnostic Benchmark and Metrics for Paraphrase Robustness in VLA Models. *EMNLP*, 2026.

S. Krishna, T. Han, A. Gu, S. Wu, S. Jabbari, and H. Lakkaraju. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. arXiv:2202.01602, 2022.

T. Lanham et al. Measuring faithfulness in chain-of-thought reasoning. arXiv:2307.13702, 2023.

X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, et al. Evaluating Real-World Robot Manipulation Policies in Simulation. *CoRL*, 2024. arXiv:2405.05941.

Z. C. Lipton. The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*, 2016. arXiv:1606.03490.

B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. *NeurIPS Datasets & Benchmarks Track*, 2023.

Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *ACL*, 2022.

Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, et al. Faithful Chain-of-Thought Reasoning. *IJCNLP-AACL*, 2023. arXiv:2301.13379.

Q. Lyu, M. Apidianaki, and C. Callison-Burch. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 2024. arXiv:2209.11326.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT. *NeurIPS*, 2022. arXiv:2202.05262.

M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky. State of What Art? A Call for Multi-Prompt LLM Evaluation. arXiv:2401.00595, 2024.

T. Nguyen, D. Guo, T. Lampo, F. Wu, and B. Yaman. VLADriveBench: Evaluating CoT-Action Relationship in VLA for Autonomous Driving. arXiv:2606.12706, 2026.

L. Parcalabescu, and A. Frank. On Measuring Faithfulness or Self-consistency of Natural Language Explanations. *ACL*, 2024. arXiv:2311.07466.

K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models. arXiv:2501.09747, 2025.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, et al. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *JMLR*, 22(164):1–20, 2021. arXiv:2003.12206.

W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox. The Colosseum: a benchmark for evaluating generalization for robotic manipulation. *RSS*, 2024.

Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

A. Radhakrishnan, K. Nguyen, A. Chen, C. Chen, C. Denison, D. Hernandez, et al. Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. arXiv:2307.11768, 2023.

I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna. AI and the Everything in the Whole Wide World Benchmark. *NeurIPS Datasets & Benchmarks Track*, 2021. arXiv:2111.15366.

R. Schaeffer, B. Miranda, and S. Koyejo. Are Emergent Abilities of Large Language Models a Mirage? *NeurIPS*, 2023. arXiv:2304.15004.

M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design, or: How I Learned to Start Worrying about Prompt Formatting. *ICLR*, 2024.

S. Serrano, and N. A. Smith. Is Attention Interpretable? *ACL*, 2019. arXiv:1906.03731.

Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, et al. Octo: An Open-Source Generalist Robot Policy. *RSS*, 2024. arXiv:2405.12213.

T. D. Trinh, N. Akhtar, and B. Azam. Altered Thoughts, Altered Actions: Probing Chain-of-Thought Vulnerabilities in VLA Robotic Manipulation. arXiv:2603.12717, 2026.

T. D. Trinh, N. Akhtar, and B. Azam. Reasoning as a double-edged sword: architecture and cross-stage robustness in vision-language-action models. arXiv:2607.17786, 2026.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *NeurIPS*, 2023.

H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, et al. BridgeData V2: A Dataset for Robot Learning at Scale. *CoRL*, 2023. arXiv:2308.12952.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*, 2022. arXiv:2201.11903.

Y. Wu, A. Li, T. Hermans, F. Ramos, A. Bajcsy, and C. Pérez-D'Arpino. Do What You Say: Steering Vision-Language-Action Models via Runtime Reasoning-Action Alignment Verification. arXiv:2510.16281, 2025.

X. Ye, and G. Durrett. The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning. *NeurIPS*, 2022. arXiv:2205.03401.

C. Yin, Y. Lin, W. Xu, S. Tam, X. Zeng, et al. DeepThinkVLA: Enhancing Reasoning Capability of Vision-Language-Action Models. arXiv:2511.15669, 2025.

M. S. Zaman, and M. Srivastava. A Causal Lens for Evaluating Faithfulness Metrics. *EMNLP*, 2025. arXiv:2502.18848.

M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. *CoRL*, 2024. arXiv:2407.08693.

F. Zhang, and N. Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. *ICLR*, 2024. arXiv:2309.16042.

S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, et al. VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks. arXiv:2412.18194, 2024.

Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, et al. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. *CVPR*, 2025.

L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS Datasets & Benchmarks Track*, 2023. arXiv:2306.05685.

X. Zhou, Y. Xu, G. Tie, Y. Chen, G. Zhang, D. Chu, P. Zhou, and L. Sun. LIBERO-PRO: Towards Robust and Fair Evaluation of Vision-Language-Action Models Beyond Memorization. arXiv:2510.03827, 2025.

A Related work

CoT faithfulness in language models. Chain-of-thought prompting (Wei et al., 2022) is read as a window on the computation, on a definition of faithfulness that requires the stated reasoning to be the actual reason for the output (Jacovi and Goldberg, 2020), and the literature that tests that reading finds it does not hold: stated reasoning often does not determine the output (Turpin et al., 2023; Lanham et al., 2023; Ye and Durrett, 2022), reasoning-trained models are no exception (Chen et al., 2025b), and unfaithfulness appears on natural rather than adversarially constructed prompts (Arcuschin et al., 2025). The proposed repairs are architectural (decomposition (Radhakrishnan et al., 2023), constrained decoding (Lyu et al., 2023)) rather than metrological, and the field's own faithfulness tests are themselves contested (Atanasova et al., 2023; Parcalabescu and Frank, 2024; Lyu et al., 2024; Zaman and Srivastava, 2025): against ground-truth-labelled traces, the best current faithfulness metrics reach only 0.70 AUROC (Gur-Arieh et al., 2026), and a widely-cited unfaithfulness result

is itself confounded with answer accuracy (Bentham et al., 2024). The observation that a meaning-preserving edit can move the output as reliably as a meaning-changing one is that field's prompt-order and prompt-format sensitivity result (Lu et al., 2022; Sclar et al., 2024; Mizrahi et al., 2024) read out in action space, not a new phenomenon; Lanham et al. (2023) already use CoT paraphrasing as a faithfulness control. What we add for token-decoded VLAs is a description of what the score records (§3.2) and the observation that the floor choice decides the verdict. Our contribution is on the measurement side of that literature rather than the model side: we ask what an edit-magnitude score can license, not how to make a model more faithful.

The same question in VLAs. Transferring the perturb-and-observe method to VLAs (Kim et al., 2024; Brohan et al., 2023; Team et al., 2024; Zawalski et al., 2024; Zhao et al., 2025; Chen et al., 2025a) is attractive because the "answer" is a continuous action rather than a token, so the effect size is directly measurable. Nguyen et al. (2026) do this for autonomous driving: they pair observational CoT metrics with an intervention protocol scored by the magnitude of the induced trajectory change, and find the two disagree – ORION scores highest on observational alignment while its CoT is epiphenomenal, and Alpamayo v1.5 scores lower while its CoT is strongly causal. That separation is the kind of result a magnitude protocol can deliver when the difference between models is large, and our results do not contest it: both models have continuous trajectory heads, which exact-no-change counting (§3.2) does not describe; on the one continuous head we probe, the floors agree but do not isolate head type. What carries over is the floor question. On their highest-effect meaning-bearing prompt (*origbrake*), their own single-run pilot calls a word-shuffled control "an indistinguishable effect" from the original and a repeated-nonsense-token control "nearly twice" its effect, and their 9-seed aggregate (their Table 13) gives -0.417 m vs. -0.426 m mean displacement for the shuffle and -0.818 m for the repeated token. Meaning-destroying and meaning-free edits there move a continuous output at least as much as a meaningful one; they report this as a property of those prompts, and a two-floor check of the kind §2 proposes would say whether it is more. Huang et al. (2026) hijack a manipulation VLA's CoT with an adversarial visual patch and read the resulting action change, on three reasoning architectures, as evidence that CoT "strongly governs" the decode; §3.2 is why action change alone does not separate a semantically driven response from a collision-driven one on a tokenized head. Trinh et al. (2026a) corrupt a LIBERO reasoning VLA's CoT under seven corruptions in three attacker tiers and find that swapping entity names lowers success by 8.3 points while reordering, direction reversal, token noise and LLM-written plausible plans stay within about ± 4 points; scored by task success on a competent policy, this is the closest behavioural counterpart to our finding that much of the reasoning text is not what the action responds to. Trinh et al. (2026b) perturb three models spanning a reasoning spectrum at the vision, reasoning and action stages on LIBERO and SimplerEnv, and find that plan–action consistency monitors that look near-perfect under naive evaluation fall to chance under adaptive attack; their reasoning-stage perturbation is scored by detection metrics, not action magnitude. Foutter et al. (2026), also driving, train a critic to classify reasoning-graph consistency and use it as a reward, a different paradigm from perturbation scoring. Wu et al. (2025) treat the same reasoning-action mismatch ("a lack of embodied CoT faithfulness") as something to fix at runtime rather than score. Separately, VLAs are reported to be sensitive to instruction rewording at the task-success level on some configurations (22–52 points, Kim et al. (2026)) and nearly insensitive on others (Zhou et al., 2025; Fei et al., 2025); none of these edits reasoning text or reads a continuous action. The manipulation benchmarks these models are graded on measure task success under distribution shift (Pumacay et al., 2024; Zhang et al., 2024), in simulation standing in for hardware (Li et al., 2024), or safety-relevant behaviour (Cui et al., 2026).

Attention as a proxy, and why we do not use it. Attention mass on the CoT segment is the obvious cheap proxy for this kind of interpretability claim, though no VLA work we know of has validated it as one. That attention weight is not by itself an explanation is argued widely in NLP (Jain and Wallace, 2019; Serrano and Smith, 2019), as is the more general point that a plausible-looking attribution can survive controls it should fail (Adebayo et al., 2018; Friedman et al., 2024). We do not add a further measurement here, relying instead on the causal edit protocol §2 defines directly. Causal intervention on the intermediate representation (Meng et al., 2022; Zhang and Nanda, 2024; Geiger et al., 2022) is the closer analogue of what we do, and our Δ_∞ is that intervention read out in action space rather than in a hidden representation.

Construct validity in benchmarks. Our claim is of a kind that recurs: that a widely reported number does not measure the thing its name says. It has been made about fairness metrics (Jacobs and Wallach, 2021), about benchmark choice (Dehghani et al., 2021; Raji et al., 2021), and about experimental reporting in RL (Henderson et al., 2018; Dodge et al., 2019; Pineau et al., 2021). The closest structural analogy is emergence: Schaeffer et al. (2023) show that several claimed sudden capability jumps are an artifact of which metric, discontinuous or continuous, is used to read the same underlying model outputs, not a discontinuity in the outputs themselves. The diagnosis here is the same move with the floor in place of the metric: which edit is nominated as meaning-preserving, not anything about the model, decides whether a semantic edit "passes." Perturbation-based interpretability metrics carry the same problem and the same fix: comprehensiveness and sufficiency (DeYoung et al., 2020) and retrain-and-evaluate calibration (Hooker et al., 2019) pair a perturbation score with an explicit null, the design our two floors extend to a continuous action space, and that different faithfulness metrics disagree with each other by more than chance is itself a documented, general phenomenon (Krishna et al., 2022; Chan et al., 2022) rather than a defect specific to our instrument. The remedy in each case is the same one we follow here: report the null and the variance alongside the score rather than the ranking alone (Lipton, 2016; Gebru et al., 2021). Zhao et al. (2025) is cited but not evaluated: as of this submission its authors host no released checkpoint, so it is not a lineage we could add without training one from scratch. Chen et al. (2025a) shares Zawalski et al. (2024)'s template and supplies the LIBERO-90 CoT annotations we use (§B.1), so it is the same lineage, not independent evidence.

The mechanism is not specific to faithfulness scoring, either. §3.2 describes that on token-decoded heads a magnitude score comparing two decoded actions is close to the rate at which the action changes at all (Pertsch et al., 2025), whatever perturbation produced the two decodes. A perturbation-based VLA robustness or out-of-distribution-sensitivity score built the same way (perturb the input, measure how far the action moved) would need the same per-model null this paper argues a faithfulness score needs, and for the same reason. We do not test this: no such score is evaluated here, and the point concerns the shared decoding, not any specific robustness benchmark's numbers.

B SUPPORTING EVIDENCE

This appendix follows the body. §B.1–B.2 give the protocol, edit taxonomy, evaluation checklist and models (§2). §B.3–B.4 give the full score tables and the analyses behind §3.1–3.3: the threshold-free divergence, ECoT-bridge on Bridge V2, per-task, per-family and multiplicity analyses, the judge validation, the size-matched nulls, the continuous-head probes and the oracle. §B.5 and §B.6 cover verification, the limitations the body lists, and the closed-loop rollouts (§3.4).

B.1 THE MEASUREMENT PROTOCOL

CoT-Faith is the evaluation protocol and benchmark we introduce to audit edit-based faithfulness scores: it applies CoT edits, scores the action change, and reports every score against nulls and controls.

Task suite and prompt template. We evaluate primarily on the Embodied-CoT LIBERO-90 dataset (Chen et al., 2025a), which contains dense reasoning annotations across 90 LIBERO tasks spanning the four LIBERO suites (Spatial, Object, Goal, 10/Long). Each sample is a triple $(\text{img}, \text{instr}, \text{cot})$. The assistant turn follows the standard nine-tag ECoT template of Zawalski et al. (2024):

```
USER: What action should the robot take to {instr}?  ASSISTANT: TASK: {
    ...} PLAN: {...} ... GRIPPER POSITION: {...} ACTION: {7 tokens}
```

Non-CoT baselines use the OpenVLA template, leaving the CoT segment empty. *Sampling.* Each condition scores 100 first-step samples per sampling seed, with 3 seeds for every row of Table 10 and the DeepThinkVLA reruns; Wilson intervals on three-seed rows use counts pooled over seeds.

B.1.1 EDIT TAXONOMY

CoT-Faith organizes edits by *semantic depth* of the intervention:

Controls ($\times 2$). Interventions that should have zero or maximal effect if the metric is well-behaved. *selfsplice_control:* identity substitution (X→X), expected and observed 0.00. *cross_task_swap:* replace the entire CoT with the reasoning from an unrelated LIBERO sample, the ceiling.

Tier 1: word-level semantic ($\times 4$). Single-token or short-phrase swaps that mutate one semantic axis at a time. *direction_flip:* left↔right, up↔down, in↔out within `MOVE`. *gripper_flip:* grasp↔release, open↔close within `PLAN`/`SUBTASK`. *verb_swap:* action verbs → unrelated verbs, though an LLM judge finds the swap preserves meaning on 0.575 of pairs (Section B.4.6) rather than changing it, so this family is weaker than its name suggests. *negation:* prepend "do not" to `SUBTASK`/`MOVE`.

Tier 2: object / causal-chain ($\times 3$). Higher-order edits requiring cross-tag consistency. *subject_swap:* primary object referent → visible distractor across all reasoning tags. *location_swap:* spatial adverbs and 2-word location phrases within `PLAN`/`SUBTASK`/`TASK` (left/right, top/bottom, front/back, near/far, inside/outside, 'left compartment' etc.). *adversarial_plausible:* object → second-most visible object, *intended* to be visually plausible but wrong. An LLM judge confirms the referent changes (0.958) but rejects the plausibility half of that description: only 0.125 of its edits are judged plausible (Section B.4.6). We keep the family and the name, and readers should treat it as an unconstrained object substitution rather than as an adversarially plausible one.

Nulls ($\times 4$). Meaning-preserving perturbations whose score defines each model's *floor*. Every reported $\mathcal{F}$ must be read against them, and we treat them as first-class families rather than sanity checks. *paraphrase_null:* synonym substitution within `MOVE`/`SUBTASK` that preserves the instruction's meaning. *syntactic_scramble:* word-order shuffle within `MOVE` and `SUBTASK`, preserving content words and length; it is judge-validated (Section B.4.6) and is the second meaning-preserving CoT floor of the decision rule (§2). *bbox_jitter_null:* ± 1 px perturbation of the `VISIBLE OBJECTS` bounding-box coordinates, below the resolution the reasoning operates at. *instr_random_sub:* 5-token random substitution in the *instruction* rather than the CoT, which measures each model's global prompt sensitivity from *outside* the CoT segment and so bounds how much of any CoT-edit score is CoT-specific at all. (*selfsplice_control* is not a null in this sense: it is an identity control with nothing to calibrate a floor against.)

In total: **7** semantic families (Tiers 1–2), **4** nulls, **1** identity control and **1** ceiling – **13** families in all. Applied to N=58–100 LIBERO samples per model, with *location_swap* at $N \approx 58$–74 after an annotation fix (a generator regex that conflated multi-word location phrases with single-word boundary patterns, undercounting applicable samples at $N \approx 12$ before the split). The calibration nulls run on all **11** configurations in both architecture families (the seven "ours" rows at 3 sampling seeds each, ECoT-bridge and the three DeepThinkVLA checkpoints), plus an independent no-CoT retraining replicate, so every $\mathcal{F}$ in Table 10 has a measured floor and ceiling from its own run, and the out-of-CoT specificity check is defined on every model (Table 15). One family, *bbox_jitter_null*, is inapplicable to DeepThinkVLA, whose CoT renders no bounding boxes (Table 15).

Capability axes. The thirteen families above are not thirteen unrelated perturbations; each targets one of six axes a CoT-conditioned policy's control could depend on, and a model's row in Table 10/Table 11 is a profile over these axes, not a single number: *directional grounding* (*direction_flip*, *gripper_flip*: left/right, up/down, in/out, grasp/release); *object and referent grounding* (*subject_swap*, *adversarial_plausible*: which visible object the reasoning names); *spatial relation* (*location_swap*: which named place the action targets); *action semantics* (*verb_swap*, *negation*: which verb, and whether it is negated); *full-CoT dependence* (*cross_task_swap*: an entire unrelated reasoning trace); and *instruction-vs-CoT attribution* (*instr_random_sub*: whether a response comes from the instruction rather than the CoT segment at all). The null and control families (§B.1.1) are not a seventh capability axis; they are the null every one of the six is read against. This grouping is descriptive of the existing edit families, not a new experiment, and we report it so that a future comparison can say a model is strong on one axis and weak on another rather than citing a single $\mathcal{F}$.

Figure 5 draws the resulting per-model profile over these axes.

Table 8: The six capability axes, as a count rather than only a list. Three axes are triangulated by two families; three rest on one family each – a per-axis score on a one-family axis is one edit's result, not yet a triangulated finding, and *instr_random_sub* is also one of the four nulls (§B.1.1), not a family exclusive to this axis.

Axis	Families	n_{fam}
Directional grounding	*direction_flip*, *gripper_flip*	2
Object/referent grounding	*subject_swap*, *adversarial_plausible*	2
Action semantics	*verb_swap*, *negation*	2
Spatial relation	*location_swap*	1*
Full-CoT dependence	*cross_task_swap*	1*
Instruction-vs-CoT attribution	*instr_random_sub*	1*

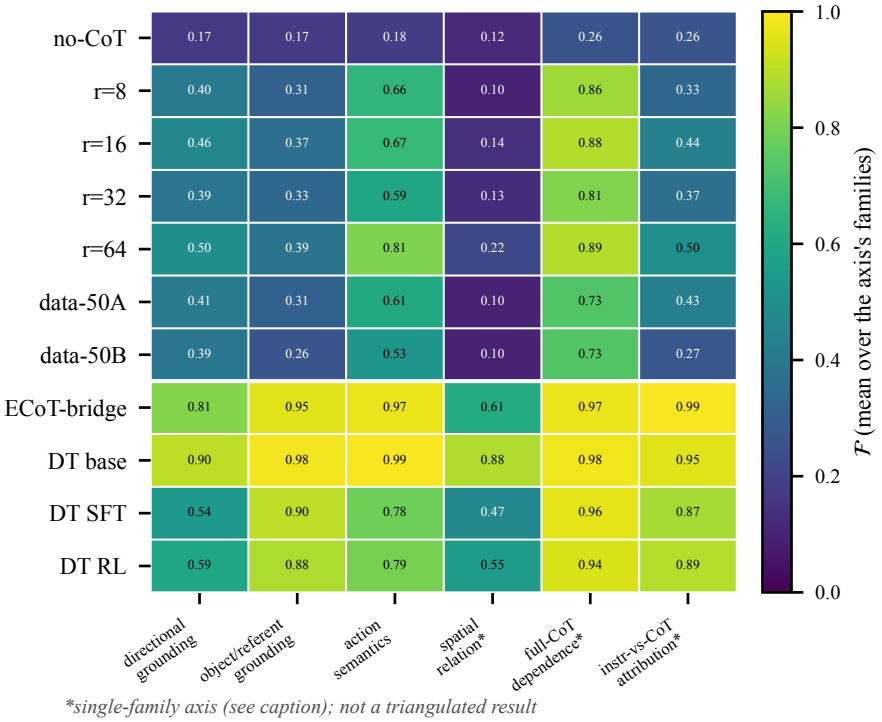

Figure 5: The capability profile Table 8 describes, drawn directly: one cell per (model, axis), the mean $\mathcal{F}$ over that axis's families, on all 11 calibrated configurations. No column or row is averaged into a single per-model number, so a model's row is read as a profile, not ranked by a mean. Starred columns are single-family axes (Table 8): read them as that one family's result, not a triangulated one. The white rule separates the ECoT/LoRA lineage from DeepThinkVLA.

B.1.2 Evaluation checklist

The controls and diagnostics of §2–4, collected as one checklist for a new model or edit family.

Stage 1 – Null calibration. Score the model on the null-like controls this protocol requires (*selfsplice_control*, *paraphrase_null*, *syntactic_scramble*, *instr_random_sub*; *bbox_jitter_null* where the CoT renders bounding boxes) – a protocol-specific list, not the same set as Figure 3's four trailing controls. No stage below is interpretable without this one: §3.1 is the paper-length version of what happens when it is skipped.

Stage 2 – Size-matched selectivity. For each semantic family, require $\mathcal{F}$ to exceed that of a meaning-preserving null that edits the same tags and the same number of words (the primary statistic of §2, item 1 of §4), with Holm correction across configurations. Unmatched floors (*paraphrase_null*, *syntactic_scramble*) are reported beside it but are not a criterion, because their sign depends on edit size and on the state distribution (§3.1, §3.3). Here *direction_flip* clears its matched null on 11 of 11

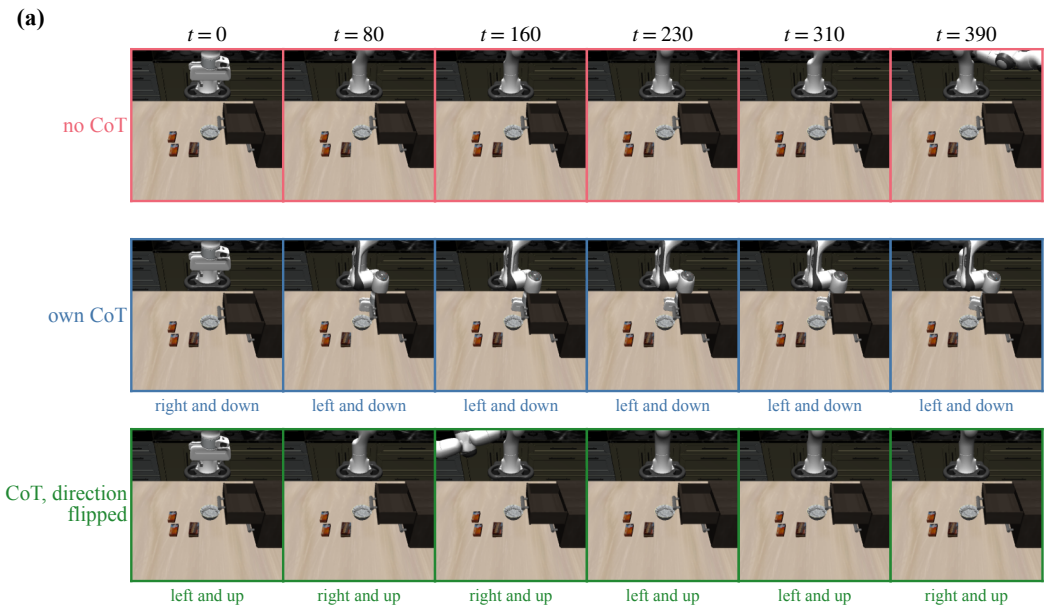

Figure 6: What a first-step probe and a rollout probe see differently, on one episode – illustrative, not a faithfulness measurement (the eight-scene replay behind Stage 6a is the measured version). One LIBERO episode, three arms (no-CoT, own CoT, direction-flipped), replayed from the same init state for 400 steps, shown at $t = 0, 80, 160, 230, 310, 390$; the MOVE phrase each arm's CoT carried that step is below its row. All three start pixel-identical at t=0 ($|\Delta\text{pixel}| = 0.0$). All three score 0 rollout success, so no ΔSR is defined: what it shows instead is that an edit reaches the actuator and keeps moving the arm for hundreds of steps, which a first-step Δ_∞ cannot show. The own-CoT arm barely travels, its whole path inside a 21.8 cm box, while the flipped and no-CoT arms reach 139.7 cm and 128.8 cm from it.

configurations and *negation* on 6 of 11 (Table 4). Clearing this stage is necessary, not sufficient: the `no-CoT` control clears it for *direction_flip* too.

Stage 3 – Directional consistency. For families with a signed semantic (*direction_flip*, *gripper_flip*, *negation*), require $\mathcal{F}_{\text{dir}}$ (Eq. 3, $\cos < -0.5$) to clear the model's own null ceiling (§3.3). 7 of 11 configurations clear it here.

Stage 4 – Decoder diagnostic. Report the action-space collapse diagnostic (§B.6.2) and, where feasible, the threshold-free $\overline{\text{TV}}$ (§3.2) beside every score in Stages 2–3, since both can inflate or deflate a magnitude score independent of semantic dependence.

Stage 5 – Stability. Require $|\Delta\mathcal{F}|$ between two same-recipe retraining replicates to be smaller than the model-to-model gap being reported (§3.6); 0.068 is this cohort's own measured reference, from the six `Ours` retrains with full edit runs, not a universal constant for a model retrained on a different recipe.

Stage 6a – Trajectory correspondence (descriptive). Report whether the step-level effect reappears as trajectory divergence in closed loop. On `r=32` an eight-scene replay reproduces the family ranking ($\rho = +0.964$), and on DT-SFT/RL $\mathcal{F}$ ranks trajectory divergence at ρ=0.98. But the raw movement rate explains the `r=32` replay (first 80 steps of 8 LIBERO-90 scenes) at least as well (ρ=1.000, p=0.0002), so it does not add evidence independent of the decode-collision mechanism: this stage says the score measures movement that persists, not that the movement matters. It is not evidence of faithfulness and no claim rests on it. Figure 6 illustrates what a rollout shows that a first-step probe cannot.

Stage 6b – Task-level rollout faithfulness. Confirm that a policy's success rate, not just its trajectory, diverges under the same edit in the direction faithfulness predicts. Our LoRA variants cannot be tested this way (in-suite success is $0/20$ for every arm of the replay and $0/40$ in both arms of the paired rollout). On the two competent DeepThinkVLA policies (§B.6.3), only a cross-task CoT changes success; no within-task content edit does, so neither policy clears this stage.

Table 9: The 15 models in CoT-Faith across 4 model families, one row per model (Table 1 in the main text is the 4-row family summary). "Ours" variants LoRA-fine-tuned from `Embodied-CoT/ecot-openvla-7b-bridge`.

Model	Base architecture	Training data	LoRA rank	Reasoning target
Family A: OpenVLA non-CoT baselines (4)				
OpenVLA-libero-spatial	OpenVLA-7B	LIBERO-Spatial	(public FT)	no CoT
OpenVLA-libero-object	OpenVLA-7B	LIBERO-Object	(public FT)	no CoT
OpenVLA-libero-goal	OpenVLA-7B	LIBERO-Goal	(public FT)	no CoT
OpenVLA-libero-10	OpenVLA-7B	LIBERO-10	(public FT)	no CoT
Family B: DeepThinkVLA (3)				
DeepThinkVLA-base	PaliGemma	(pretrain only)	—	—
DeepThinkVLA-SFT	PaliGemma	LIBERO-CoT	(public)	CoT (SFT)
DeepThinkVLA-RL	PaliGemma	LIBERO-CoT	(public)	CoT (RL)
Family C: ECoT (1 public)				
ECoT-bridge	ECoT-OpenVLA-7B	Bridge-V2	(public)	full 9-tag CoT
Family D: Ours (7 LoRA variants)				
Ours r=8/16/32/64	ECoT-bridge	LIBERO-90	{8,16,32,64}	full CoT
Ours no-CoT	ECoT-bridge	LIBERO-90	32	action-only
Ours data-50A/B	ECoT-bridge	LIBERO-90 (50% split A/B)	32	full CoT

The five recommendations of §4 are Stages 2, 4 (with the noise sharing of Stage 1), 3, 5 and 6b of this checklist. They resolve into three tiers. *Required for reporting any $\mathcal{F}$ or $\mathcal{F}_{dir}$ at all*: Stage 1, including shared sampling noise on stochastic heads (§B.4.7). *Required for a comparative claim about reliance on the CoT*: Stages 2 and 3 cleared, Stage 4 reported, Stage 5 satisfied, and, wherever the policy is competent, Stage 6b. No configuration here supports such a claim. The six CoT-trained LoRA variants clear Stages 2–3 but cannot be tested at Stage 6b, and the `no-CoT` control fails Stage 3; ECoT-bridge clears Stage 2 for *direction_flip* (its *negation* contrast is not Holm-significant) and only weakly Stage 3. DT-SFT/RL clear Stage 2 for *direction_flip*, fail Stage 3, and show no success dependence at Stage 6b. Stage 6a is descriptive only.

B.2 Models

Table 1 (main text) gives the 4-family summary; Table 9 below is the full per-model listing. Both span 2 base architectures: OpenVLA-7B for A, C and D, PaliGemma for B. Only the *no-CoT* variant differs from our main model in the training *target* (action-only); all other "ours" variants share the same nine-tag target. This isolates the effect of reasoning-target training with all other factors held constant. *Training compute*: every "ours" checkpoint is a LoRA fine-tune of the same 7B ECoT-OpenVLA (Zawalski et al., 2024), $15{,}000$ steps, batch size 2, bf16, 14 runs total (§3.6's seven configurations, each trained twice).

B.3 Full score tables

Table 10 reports the magnitude score $\mathcal{F}(m, f)$ for every model×family cell we measured: rows are the 8 CoT-VLA models (non-CoT baselines have no CoT to edit and are not part of this table), columns are ten edit families, including *syntactic_scramble*, plus *paraphrase_null* (*para*), organized by tier. Every row is a 3-seed mean±std (13-family runs for the `Ours` rows, an 11-family sweep for ECoT-bridge), so every cell carries an error bar and the *para* floor is populated on all 8 rows.

The table is a diagnostic, not a ranking: on every row the mean over the seven semantic families falls below that row's own paraphrase floor ($\bar{\mathcal{F}}_{\text{diff}}$ from -0.027 to -0.180), which is why the checklist of §B.1.2 uses the size-matched contrast rather than this floor; no cell is bolded, and the row order is the configuration order of Table 1. Figure 3 draws these cells and those of Table 11.

Figure 7 shows that neither LoRA rank nor the data subset orders these scores beyond retraining noise.

B.4 Analysis

Table 12 gives the threshold-free divergence of §3.2.

Table 10: Magnitude scores $\mathcal{F}(m, f)$: a diagnostic table, not a ranking (1 of 2: control and word-level tiers; Table 11 continues with the object/causal tier, the paraphrase floor and $\mathcal{F}_{\text{dir}}$). 8 CoT-VLAs $\times$ 11 edit families (ten families including *syntactic_scramble*, plus *paraphrase_null*), split across two tables so neither needs the column count that made a single one shrink past legibility. Every row is a mean$\pm$std across 3 sampling seeds (13-family runs for the `Ours` rows, an 11-family sweep for ECoT-bridge) at N as given in the header, so every cell carries a floor from its own run; the two out-of-CoT floors are in Table 15 rather than here. Nothing is bolded; on *every* row the mean over the seven non-control families falls *below* that row's own *para* floor ($\bar{\mathcal{F}}_{\text{diff}}$ from -0.027 to -0.180; §3.1), so a column maximum identifies the model that reacts most to *any* rewording, not the model that is most faithful. *ss* = selfsplice_control (null; observed 0.00 on 8/8 here, and on 11/11 including the DeepThinkVLA family). This table is the ECoT architecture family only; the second family's full 13-family protocol is the lower block of Table 15, which is *not* merged in because its floor, ceiling, and decode conventions all differ. Wilson 95% CI half-widths on the pooled 3-seed counts are ≤ 0.067 (widest: *subject_swap* on r=64), but they bound the *sample* draw only, and they are the *smaller* term: retraining r=32 at the same configuration moves a single family by up to 0.193, $\sim$2.9$\times$ the widest interval here, and across all six retraining pairs the maximum is 0.317 (*verb_swap*, r=8; §3.6). *Retrain median/max*: median and maximum $|\Delta\mathcal{F}|$ for that family across the six same-config retraining pairs (§3.6); *verb_swap*'s 0.112/0.317 is the noisiest cell in the benchmark by a wide margin, consistent with it also being the least semantically reliable generator (§B.4.6).

	identity / floor / ceiling			Tier 1 (word-level)			
Model	ss	scram.	cross	dir.	grip.	verb	neg.
N =	69	100	100	100	100	100	100
Ours r=8	0.00	.41±.03	.86±.01	.70±.04	.11±.01	.64±.05	.67±.04
Ours r=16	0.00	.44±.01	.88±.03	.75±.03	.17±.03	.68±.03	.66±.04
Ours r=32	0.00	.35±.04	.81±.02	.65±.02	.13±.01	.59±.01	.59±.02
Ours r=64	0.00	.47±.05	.89±.05	.82±.02	.17±.06	.84±.04	.78±.03
Ours no-CoT	0.00	.15±.05	.26±.03	.27±.04	.07±.01	.16±.04	.20±.07
Ours data-50A	0.00	.39±.04	.73±.03	.70±.02	.12±.03	.65±.04	.57±.03
Ours data-50B	0.00	.35±.03	.73±.07	.64±.02	.15±.01	.54±.02	.51±.02
ECoT-bridge	0.00	.86±.04	.96±.01	.96±.01	.70±.02	.95±.04	.95±.01
retrain median	—	—	—	0.032	0.025	0.112	0.030
retrain max	—	—	—	0.081	0.093	0.317	0.170

Table 11: Magnitude scores $\mathcal{F}(m, f)$ (2 of 2: object/causal tier, the paraphrase floor, and the direction-aware score). Same 8 rows and run as Table 10. *para* = paraphrase_null, the per-model surface-form floor. $\mathcal{F}_{\text{dir}}$ (Eq. 3) requires the action to reverse, not merely to move; its ranking agrees only weakly with the magnitude columns ($\rho = +0.476$), and it moves `ECoT-bridge` from first to seventh (Section 3.3). $\mathcal{F}_{\text{dir}}$ is reported on each checkpoint's native de-quantization grid; the magnitude columns use each model's native decoded actions. The separate resolution sweep in §3.2 shows that regrouping the underlying discrete action grid can nevertheless change the collision-driven component of $\mathcal{F}$. *Retrain median/max*: median and maximum $|\Delta\mathcal{F}|$ for that family across the six same-config retraining pairs, as in Table 10; no cell here reaches *verb_swap*'s.

	Tier 2 (object/causal)				
Model	subj.	loc.	adv.	para.	$\mathcal{F}_{\text{dir}}$
N =	69	74	69	100	100
Ours r=8	.32±.09	.10±.06	.31±.07	.59±.03	0.65±.02
Ours r=16	.37±.06	.14±.02	.36±.06	.55±.04	0.67±.05
Ours r=32	.33±.03	.13±.04	.33±.01	.57±.02	0.58±.03
Ours r=64	.40±.05	.22±.02	.37±.02	.66±.02	0.78±.02
Ours no-CoT	.16±.03	.12±.05	.18±.06	.19±.05	0.09±.02
Ours data-50A	.30±.04	.10±.02	.32±.03	.54±.06	0.64±.01
Ours data-50B	.28±.05	.10±.03	.26±.04	.45±.03	0.54±.01
ECoT-bridge	.93±.03	.59±.05	.93±.03	.95±.02	0.12±.04
retrain median	0.081	0.044	0.054	—	—
retrain max	0.128	0.103	0.134	—	—

B.4.1 ECoT-bridge in distribution on Bridge V2

The LIBERO rows score ECoT-bridge off the corpus it was trained on and without its own normalization statistics. Here we score it on Bridge V2 itself, with its `bridge_orig` normalization, joining the Bridge V2 episodes to their ECoT reasoning annotations. There are four runs of 400 observations

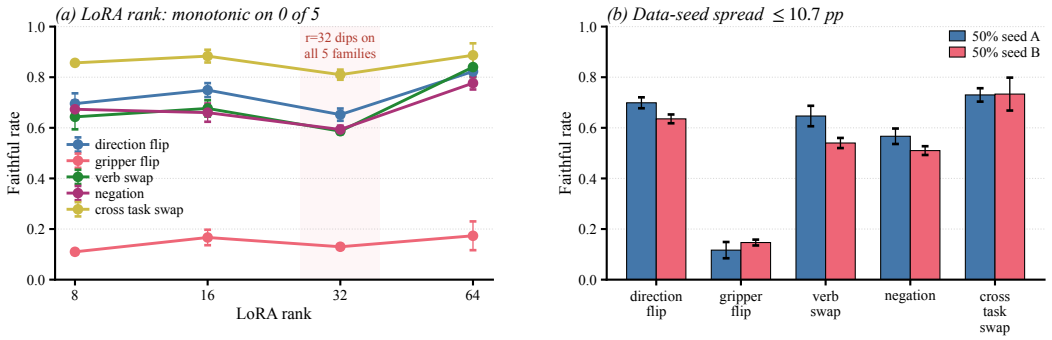

Figure 7: The retraining-variance figure §3.6 names. *(a)* Faithful rate vs. LoRA rank, five families: r=32 dips below both neighbors on all five, so rank does not order faithfulness monotonically. *(b)* The two 50%-data seeds, same five families: spread up to 10.7 pp, the same order of magnitude as (a)'s non-monotonicity. Both spreads are within the per-family retraining movement of §3.6 (median 0.037, up to 0.317), which is why neither licenses a rank- or seed-based ordering.

	floor $\overline{\text{TV}}$			$\overline{\text{TV}}_{\text{diff}}$ vs. floor	
Model	para	scram	$\overline{\text{TV}}$	para	scram
`ECoT-bridge`	0.663	0.538	0.590	-0.072^{*}	$+0.052$
`r=32`	0.136	0.063	0.105	-0.031^{*}	$+0.043^{*}$
`r=64`	0.141	0.080	0.127	-0.014^{*}	$+0.047^{*}$
`DT base`	0.589	0.459	0.557	-0.032^{*}	$+0.098^{*}$
`DT SFT`	0.287	0.169	0.269	-0.018^{*}	$+0.101^{*}$
`DT RL`	0.288	0.167	0.267	-0.021^{*}	$+0.101^{*}$
all 6 configurations				− 6/6	+ 6/6

Table 12: The floor collapse survives a threshold-free metric, on the six configurations rerun this way. $\overline{\text{TV}}$ is the mean of $D_{\text{TV}}^{s,f}$ (defined in the text) over the same seven non-control semantic families as $\bar{\mathcal{F}}$; * marks the 11 of 12 $\overline{\text{TV}}_{\text{diff}}$ cells whose 95% paired-bootstrap interval (same clustered-resampling procedure as Table 2) excludes zero: 6 of 6 against *paraphrase_null*, 5 of 6 against *syntactic_scramble*, the one exception `ECoT-bridge`. This is a smaller cohort than Table 2's eleven configurations (`r=8`, `r=16`, `data-50A/B` and `no-CoT` were not rerun under this metric) but it is the first continuous-metric evidence from the DeepThinkVLA family, whose Table 2 rows all come from the thresholded Δ_{∞} statistic.

each: teacher-forced reasoning at the first step with sampling seeds 0 and 1, teacher-forced reasoning at a uniformly random step, and self-generated reasoning at the first step. All families of the LIBERO protocol and both size-matched nulls are scored within each run (Table 14). The self-generated run admits no *subject_swap* or *adversarial_plausible* edits, so its semantic mean is over five families. The primary statistic uses the same statistic, bootstrap and Holm correction as Table 4, with Holm applied across the four runs.

Reading. Every qualitative result of the LIBERO rows holds in distribution:

- $\mathcal{F}$ is close to the rate at which an edit changes the action at all.
- The two unmatched floors straddle the semantic mean.
- Both semantic families beat their size-matched nulls in every run, though less decisively than on LIBERO (2 and 3 of 4 Holm-significant).
- ECoT-bridge moves without reversing. Its $\mathcal{F}_{\text{dir}}$ stays low with a positive mean cosine, while its ceilings stay near zero. Its LIBERO non-reversal is therefore not an artifact of the Bridge and LIBERO action frames using different axis conventions.

The random-step run gives the highest $\mathcal{F}_{\text{dir}}$ (0.274) and the lowest cosine (+0.310), still far from reversal.

Table 13: ECoT-bridge in distribution on Bridge V2, collision relation and unmatched floors. R^2: $\mathcal{F}$ against $1 - P(\Delta_\infty{=}0)$ over the run's families; $\bar{\mathcal{F}}$: semantic mean; Δ: $\bar{\mathcal{F}}$ minus the floor with its observation-clustered BCa 95% interval.

			vs. *paraphrase_null*		vs. *syntactic_scramble*	
Run	R^2	$\bar{\mathcal{F}}$	Δ	BCa 95%	Δ	BCa 95%
self-gen., first step	0.977	0.740	-0.096	$[-0.131, -0.061]$	$+0.052$	$[+0.008, +0.099]$
TF, first step, seed 0	0.986	0.701	-0.077	$[-0.128, -0.031]$	$+0.040$	$[-0.016, +0.094]$
TF, first step, seed 1	0.961	0.674	-0.072	$[-0.120, -0.027]$	$+0.074$	$[+0.020, +0.126]$
TF, random step, seed 0	0.992	0.834	-0.056	$[-0.090, -0.023]$	$+0.081$	$[+0.033, +0.131]$

Table 14: ECoT-bridge in distribution on Bridge V2, primary statistic and direction-aware score. Paired differences against the size-matched nulls, * Holm-significant across the four runs; $\mathcal{F}_{\text{dir}}$ and the mean translation cosine after *direction_flip*; *ceiling*: the largest $\mathcal{F}_{\text{dir}}$ over families with no direction to reverse; *excl.*: samples with an undefined cosine.

Run	*dir* − m.	BCa 95%	*neg* − m.	BCa 95%	$\mathcal{F}_{\text{dir}}$	cos	ceiling	excl.
self-gen., first step	$+0.098^*$	$[+0.041, +0.159]$	$+0.045^*$	$[+0.010, +0.083]$	0.099	$+0.566$	0.025	0
TF, first step, seed 0	$+0.045$	$[-0.027, +0.118]$	$+0.085^*$	$[+0.050, +0.123]$	0.083	$+0.646$	0.015	0
TF, first step, seed 1	$+0.095$	$[+0.014, +0.176]$	$+0.065^*$	$[+0.020, +0.105]$	0.075	$+0.658$	0.033	0
TF, random step, seed 0	$+0.119^*$	$[+0.060, +0.185]$	$+0.020$	$[-0.015, +0.050]$	0.274	$+0.310$	0.090	0

Identity floor. selfsplice_control scores 0.140, 0.074 and 0.227 in the three teacher-forced runs, against 0.00 on every LIBERO run, possibly because its case-insensitive substitution changes capitalization in the annotations; the Bridge V2 rate contrasts are therefore weak evidence, as the body notes.

The instruction-side null. Corrupting five instruction tokens, outside the CoT, moves the action more than the semantic CoT edits on 6 of 11 configurations; `ECoT-bridge` is below the control (0.878), and on the other five the margins ($+0.082$, $+0.074$, $+0.025$, $+0.014$ and $+0.009$ in $\bar{\mathcal{F}}$) are within retraining noise, so no configuration clears its own out-of-CoT control by a margin clearly larger than its own retraining noise.

Threshold sweep. The verdict also depends on the threshold, another analyst choice: the paraphrase-floor sign count is 11 negative at $\tau \leq 0.10$, 8/3 at $\tau{=}0.20$ and 6/5 at $\tau{=}0.30$, while the scramble-floor count stays at 10–9 positive.

For DeepThinkVLA, which de-quantizes to its own action quantiles, $\tau{=}0.05$ is 1.0–1.22$\times$ stricter on translation and gripper and 3.6–8.1$\times$ stricter on rotation than for the ECoT rows, a bias against, not for, its paraphrase-floor rows.

B.4.2 Per-task decomposition

Every $\bar{\mathcal{F}}$ here averages over samples drawn from LIBERO-90 episodes, and an average can hide its shape in two ways: the unmatched paraphrase-floor comparison could come from a minority of tasks where the paraphrase edit happens to move the action a lot, leaving the majority above their floor; or $\mathcal{F}$ could be so task-dependent that a per-model number is not a property of the model at all. Neither question needs new inference, since every edit record carries `file_base`, the LIBERO episode file the observation came from, so $\mathcal{F}$ can be recomputed *within* each task from the artifacts on disk.

Protocol. Pooling the three sampling seeds of each 3-seed run gives **85** distinct LIBERO-90 episode files, the *same* 85 for all eight models (every run draws from the same sampler seeds), so the test is paired across tasks. Within each task we compute $\mathcal{F}_{\text{sem}}$, the macro average over the 7 non-control families present for that task (at least 4 of the 7 on every task, from at least 5 scored records), and that task's own $\mathcal{F}(\textit{paraphrase_null})$, and report the *sign* of the difference with an exact two-sided binomial test, ties dropped. The regrouping script first re-derives each model's published $\bar{\mathcal{F}}$ and paraphrase floor from the same records under the same convention and refuses to run if either disagrees by more than 10^{-9}; all eight reproduce their three-seed values exactly (for ECoT-bridge its 11-family sweep, 0.860/0.947, not the single-seed 0.869/0.960 of Table 15).

Table 15: Two-sided calibration on all eleven models we can calibrate, spanning both architecture families. Every quantity in a row is computed only against families from that same run, so no cross-run offset enters. The seven *Ours* three-seed rows are 3-sampling-seed means (per-family F averaged over seeds, Wilson intervals on pooled counts); the remaining rows (ECoT-bridge, the no-CoT replicate, and the three DeepThinkVLA checkpoints) are single-seed calibration runs and are marked †. ECoT-bridge's $0.869/0.960$ therefore differ from the $0.860/0.947$ in Table 16: same model and protocol, this row's single 13-family run against that table's 3-seed 11-family sweep, and $\bar{\mathcal{F}}$ sits below the floor in both. For the same reason its ceiling is 0.970 here and 0.963 in the 3-seed sweep. The *Ours* rows r=8, r=16 and r=64 differ from Table 2 by 0.001 in $\bar{\mathcal{F}}$ (as does `DT RL`, 0.724 vs. 0.725; data-50B is identical), from independent per-family aggregation over the same per-sample records; no sign count anywhere in the paper turns on that third decimal. *Ours no-CoT* appears twice and the two rows are not interchangeable: *(primary)* is the checkpoint every other table uses, *(replicate)* is an independent training of the same configuration, and the gap between their two-sided scores (-0.416 vs. $+0.127$) is itself the retraining-run variance §3.6 measures. The *bbox* cell is empty on the DeepThinkVLA rows because that family is inapplicable to them, not unmeasured: their CoT renderer emits no bounding boxes, so the edit produces a byte-identical trace and the harness records $n{=}0$ with all 100 samples skipped rather than scoring an identity edit as $\mathcal{F} = 0.000$. Dynamic range = ceiling − floor; two-sided = $(\bar{\mathcal{F}} - \text{floor})/\text{range}$, so 0 is at the floor and 1 at the ceiling and a negative value is below the floor by that fraction of the range: one signed number in place of reporting floor and ceiling separately. A range below 0.05 makes it undefined and is marked *degen.* Specificity ratio = $\bar{\mathcal{F}}/\mathcal{F}(\text{instr_rand})$, where $\bar{\mathcal{F}}$ is the mean over the seven non-control families; >1 means the CoT families move the action more reliably than corrupting five *instruction* tokens does.

model	$\bar{\mathcal{F}}$	$\mathcal{F}$(para)	$\mathcal{F}$(bbox)	$\mathcal{F}$(instr_rand)	ceiling	range	two-sided	ratio
Ours r=8	0.407	0.587	0.077	0.333	0.857	0.270	−0.665	1.222
Ours r=16	0.446	0.547	0.083	0.437	0.883	0.337	−0.299	1.022
Ours r=32	0.395	0.567	0.073	0.370	0.810	0.243	−0.706	1.067
Ours r=64	0.514	0.657	0.087	0.500	0.887	0.230	−0.621	1.028
Ours no-CoT (primary)	0.166	0.193	0.047	0.263	0.260	0.067	−0.416	0.629
Ours no-CoT (replicate)†	0.124	0.110	0.050	0.190	0.220	0.110	**+0.127**	0.653
Ours data-50A	0.393	0.537	0.067	0.433	0.730	0.193	−0.743	0.907
Ours data-50B	0.352	0.447	0.053	0.270	0.733	0.287	−0.329	1.305
ECoT-bridge†	0.869	0.960	0.460	0.990	0.970	**0.010**	*degen.*	0.878
Second architecture family (PaliGemma / `pi0fast_base`, single `<think>` block, $(10, 7)$ chunk decode)								
DeepThinkVLA-base†	0.947	0.960	—	0.950	0.980	**0.020**	*degen.*	0.996
DeepThinkVLA-SFT†	0.701	0.810	—	0.870	0.960	0.150	−0.728	0.806
DeepThinkVLA-RL†	0.724	0.820	—	0.890	0.940	0.120	−0.796	0.814

Result. Table 16 gives the decomposition: on seven of the eight models the majority of tasks sit *below* their own paraphrase floor, with median per-task $\mathcal{F}_{\text{diff}}$ from -0.167 to -0.050 – a property of most individual tasks, not of the average. The strongest cases are not the saturated one: *r=32* is below on 64 of 74 untied tasks and *r=8* on 60 of 74, against ECoT-bridge's 60 of 75. Restricting to the 53 tasks whose floor rests on at least 3 samples keeps six of the eight significant at $p < 0.05$ and reverses no sign; *data-50B* is the one row that loses significance ($0.020 \to 0.085$), reported as weakened rather than as surviving.

Over the seven full-CoT rows, p from 9.0×10^{-11} to 0.020, and all seven survive Holm correction across the eight tests (the `no-CoT` control is below). The one positive paraphrase-floor entry in the calibration set is off this table: the no-CoT *replicate* ($+0.014$), a retraining of the action-only control.

The no-CoT control. *Ours no-CoT* is the model whose per-task test is null (25 below, 29 above, 31 tied, $p = 0.683$): the checkpoint trained on an action-only target, whose whole-run margin is the smallest in the benchmark ($\bar{\mathcal{F}}_{\text{diff}} = -0.027$) – a model sitting essentially *on* its own floor in aggregate should scatter around it per task, and it does. A coherence check, stated as post hoc, not a counterexample.

Task dependence. The within-model IQR of $\mathcal{F}_{\text{sem}}$ across tasks runs from 0.129 (ECoT-bridge) to 0.343 (*r=16*), and on three of the eight models the task range is the entire unit interval: the same checkpoint scores 0.00 on some LIBERO-90 tasks and 1.00 on others. Ordering the eight models by

Table 16: Per-task decomposition of the unmatched paraphrase-floor comparison, over the same 85 LIBERO-90 episode files for all eight models. Recomputed from the 3-seed edit records with no new inference; the derivation reproduces each row's $\bar{\mathcal{F}}$ and floor from those records before regrouping them, and aborts if either disagrees with Table 15. *Below/above/tied* counts tasks whose $\mathcal{F}_{\text{sem}}$ (macro average over the 7 non-control families) is under, over, or exactly at that task's own *paraphrase_null* rate; p is an exact two-sided binomial test on the untied tasks. The *floor* ≥ 3 columns repeat the test on the 53 tasks whose floor rests on at least three samples, so that no verdict comes from a single paraphrase draw. Task IQR is the interquartile range of $\mathcal{F}_{\text{sem}}$ across the 85 tasks, the within-model heterogeneity that 6 of the 7 adjacent between-model gaps fall inside; per-task minima and maxima are in the text and the artifact. The one null row is *Ours no-CoT*, the action-only-target control, whose aggregate margin is also the smallest in the benchmark ($\bar{\mathcal{F}}_{\text{diff}} = -0.027$).

			all 85 tasks			floor ≥ 3 (53)		
model	$\bar{\mathcal{F}}$	floor	below/above/tied	p	median $\mathcal{F}_{\text{diff}}$	below/above/tied	p	task IQR
Ours r=8	0.407	0.587	60/14/11	6.2×10^{-8}	−0.160	38/9/6	2.5×10^{-5}	0.286
Ours r=16	0.446	0.547	49/25/11	0.007	−0.080	34/14/5	0.006	0.343
Ours r=32	0.395	0.567	64/10/11	9.0×10^{-11}	−0.167	43/6/4	5.7×10^{-8}	0.333
Ours r=64	0.514	0.657	57/23/5	1.8×10^{-4}	−0.143	39/11/3	9.0×10^{-5}	0.211
Ours no-CoT	0.166	0.193	**25/29/31**	**0.683**	+0.000	22/19/12	0.755	0.286
Ours data-50A	0.393	0.537	57/16/12	1.5×10^{-6}	−0.133	39/11/3	9.0×10^{-5}	0.314
Ours data-50B	0.352	0.447	48/27/10	0.020	−0.050	31/18/4	**0.085**	0.298
ECoT-bridge	0.860	0.947	60/15/10	1.6×10^{-7}	−0.083	44/7/2	1.2×10^{-7}	0.129

$\bar{\mathcal{F}}$, 6 of the 7 adjacent gaps are smaller than both models' task-level IQR, the largest such gap being 0.187; the only adjacent pair that separates at task granularity is *r=64* versus ECoT-bridge at 0.346. So, with one exception, differences between rows are smaller than the variation a single row already contains across its own tasks – independent support for Section B.3's position – and it bounds the decomposition: the sign test is a statement about a fixed set of 85 tasks, and that spread is why we do not read the per-task medians as effect sizes.

Scope. It decomposes the *edit* protocol only. A per-task breakdown of rollout success rate would answer a different question and is not derivable from the per-sample records alone: the LoRA variants never succeed, and per-task closed-loop outcomes for the two competent policies are in §B.6.3 (10 tasks $\times$ 10 episodes per suite).

B.4.3 Pseudo-replication and the three-seed DeepThinkVLA rerun

§2 notes that the eleven calibrated configurations are two base-checkpoint lineages, not eleven independent replications: eight rows are `ECoT-bridge` or a LoRA adapter on it, three are `DeepThinkVLA-base` stages. Each configuration's own paired, observation-clustered bootstrap validly tests *that* configuration, but the eight ECoT-lineage configurations share a base checkpoint, a training corpus, and, for the seven `Ours` rows, the same 85 LIBERO-90 episode files: their results are not eight independent pieces of evidence, only as many as there are independent lineages – two.

A third, weaker but genuinely independent data point comes from outside this benchmark entirely: Nguyen et al. (2026)'s own Table 13, re-read in §A, shows the identical unmatched-floor pattern (a meaning-destroying control moving their driving VLA's continuous output at least as much as a meaningful prompt) on a different model, modality, and research group's own numbers. It does not raise the lineage count to three: it is secondhand, reported by its authors as a property of their specific prompts rather than the scoring rule, and not re-derived under §3.1's calibrated-floor methodology. It does make "a quirk of these two lineages" a harder claim to sustain than the pooled count alone would suggest.

Made explicit rather than reconstructed: within the ECoT lineage, 7 of 8 configurations are individually significant (every `Ours` LoRA variant and `ECoT-bridge`; `no-CoT` the exception); within DeepThinkVLA, 2 of 3 are (`DT SFT`, `DT RL`; `DT base` the exception). Pooled, that is the manuscript's 9 of 11; by lineage, it is two independent confirmations, not nine, and each lineage's exception is the same kind: the row closest to its own floor, exactly what a null control (`no-CoT`) and what behaves like one (`DT base`, §3.1) should look like.

The two confirmations are not equally precise, either. The ECoT lineage's significance rests on 3 sampling seeds per configuration (Table 10); DeepThinkVLA's rests on one seed per checkpoint in Table 2 (†, e.g. Table 15), so its Wilson intervals bound the sample draw only, and run-to-run training variance (which §3.6 shows matters more than sampling for the ECoT lineage) is not measured for DeepThinkVLA at all. "2 of 3" is the finer-grained count for those single-seed rows, not a confirmation of comparable statistical power to the ECoT lineage's "7 of 8"; the three-seed rerun below narrows the gap in sampling precision, not in training variance.

The two competent LIBERO policies at three seeds. DT-SFT and DT-RL are rerun with sampling seeds 1 and 2 added to seed 0, with the harness and configuration unchanged; only the drawn episodes change. Pooled over 300 observations, both are significant against both floors, in opposite directions (Table 17). Per seed, the paraphrase sign is negative on all 6 (the interval excludes zero on 3), and the scramble sign is positive on all 6 (excludes zero on all 6). Replacing the single-seed rows of Table 2 with these pooled rows in the Holm family of §B.4.4 gives 9 of 11 Holm-significant against the paraphrase floor, both DeepThinkVLA rows included, and 2 of 11 against the scramble floor, both of them DeepThinkVLA. With three seeds, then, the floor dependence is significant within the DeepThinkVLA lineage on both floors, not only in the pooled count.

Table 17: $\bar{\mathcal{F}}_{\text{diff}}$ for DT-SFT/RL pooled over three seeds (300 observations), with BCa 95% intervals clustered by observation and, alongside, by LIBERO episode file (85 files).

	vs. *paraphrase_null*			vs. *syntactic_scramble*		
Model	Δ	obs. BCa	file BCa	Δ	obs. BCa	file BCa
DT SFT	-0.088	$[-0.133, -0.040]$	$[-0.137, -0.038]$	$+0.147$	$[+0.093, +0.205]$	$[+0.090, +0.202]$
DT RL	-0.079	$[-0.122, -0.033]$	$[-0.128, -0.025]$	$+0.169$	$[+0.113, +0.226]$	$[+0.119, +0.222]$

We do not fit a hierarchical or cluster-robust model with lineage as a random effect: such a model needs enough clusters to estimate a between-cluster variance component, and two clusters supply two data points, which is not enough to fit anything rather than merely narrate one. A lineage-level p-value would manufacture precision this design cannot support; the within-lineage sign and significance count above is the finest-grained statement the data honestly permit.

B.4.4 Multiplicity and bootstrap

Table 2's "9 of 11"/"3 of 11" and Table 12's "6 of 6"/"5 of 6" are each a family of paired, observation-clustered bootstraps (B=20,000 resamples, percentile 95% interval): 22 tests for the first table, 12 for the second, uncorrected for multiplicity except the 8 per-task tests of §B.4.2. We check three things: whether a bias-and-acceleration-corrected (BCa) interval disagrees with the plain percentile one near the statistic's $[-1, 1]$ boundary; whether resampling at a coarser, demonstrably-independent unit changes any verdict; and what survives once each table's own family-wise error rate is controlled.

Neither of the first two moves a single verdict. Re-deriving all 34 tests with a BCa interval (bias $z_0 = \Phi^{-1}\big(\hat{P}(\theta^* < \hat{\theta})\big)$, acceleration from a leave-one-observation-out jackknife) matches the plain-percentile call on all 34: the same 9/11 and 3/11 (Table 2), 6/6 and 5/6 (Table 12) – the boundary-proximity concern motivating BCa does not materialize here. Table 2's 22 tests resample observations, that is (seed, sample) pairs (the observation-clustered bootstrap of §3.1), which §B.4.2 shows are not exchangeable (within-model task-level $\mathcal{F}_{\text{sem}}$ IQR up to 0.343), so we re-derived at the LIBERO episode-file level (61–85 distinct files per configuration) with the same statistic, the same τ, and the same B=20,000, and reproduce the identical significance verdict on all 22 tests, configuration by configuration – the exact same nine and three rows, not merely the same counts: the within-file correlation §B.4.2 documents is real but too small, relative to each test's between-condition gap, to cross a significance boundary.

Why $\mathcal{F}$ is insensitive to where τ sits, and the two R^2 values in §3.2. The three bins spanning τ are unequal-width (0.005/0.045/0.05); it is the 90.6% combined density share on either side of τ, not a density gap at τ itself, that drives the insensitivity, and the relation is one-sided by construction ($\{\Delta_\infty > \tau\} \subseteq \{\Delta_\infty \neq 0\}$): of the 590 (run, family) cells, 0 lie above the drawn $y{=}x$ line, 426 below it, 164 exactly on it. The two R^2 values differ only in convention: $R^2{=}0.946$ is the best-fit-line correlation between $\mathcal{F}$ and $1{-}P(\Delta_\infty{=}0)$; $R^2{=}0.923$ anchors instead on the drawn $y{=}x$ line with $\mathcal{F}$ as the response (SS_{tot} from $\mathcal{F}$'s own variance); anchoring SS_{tot} on $1{-}P(\Delta_\infty{=}0)$'s variance instead

Table 18: Every bootstrap test behind Table 2 and Table 12, corrected three ways. *raw*: the percentile-CI verdict the main text reports. *BCa*: bias-and-acceleration-corrected interval, same draws. *Holm*: raw p-values Holm-corrected within that row's table (m=22 or m=12) at α=0.05. BCa never disagrees with raw; Holm is where the counts move.

Table	Floor	raw	BCa	Holm
tab:floors (m=22)	paraphrase	9/11	9/11	7/11
	scramble	3/11	3/11	1/11
tab:tvfloors (m=12)	paraphrase	6/6	6/6	2/6
	scramble	5/6	5/6	5/6

gives 0.924, an equally defensible convention not used here because §3.2 treats $\mathcal{F}$ as the thing being explained.

Table 18 Holm-corrects each table's own family at α=0.05 (two-sided bootstrap $p = 2\min(\hat{P}(\theta^*\leq 0), \hat{P}(\theta^*\geq 0))$, ceilinged at 1; the same step-down already applied to the eight per-task tests). This lowers Table 2's 22 tests to 7/11 (paraphrase) and 1/11 (scramble), and Table 12's 12 to 2/6 (paraphrase) and 5/6 (scramble, unchanged). The scramble counts were already a small minority and stay one; paraphrase moves most visibly in Table 12 ("all six significant" becomes two of six). What correction does not touch is the sign pattern §3.1 actually turns on: 11 of 11 (Table 2) and 6 of 6 (Table 12) configurations have the paraphrase floor at or above the semantic mean, before any significance test runs on top of that comparison. The seven paraphrase-floor Holm survivors are exactly the seven `ECoT`-lineage rows (`r=8/16/32/64`, `data-50A/50B`, `ECoT-bridge`); all three `DeepThinkVLA` configurations, two of which were significant before correction, fail Holm correction on this floor. The one scramble-floor survivor is the other lineage's `DT RL`. With the single-seed DeepThinkVLA rows, the two floors' Holm-corrected evidence therefore comes from disjoint lineages. Rerunning DT-SFT/RL at three seeds removes that: both then survive Holm on both floors, giving 9 of 11 and 2 of 11, and the paraphrase-floor survivors span both lineages (Table 17).

B.4.5 Per-family heterogeneity and the movement-rate decomposition

$\bar{\mathcal{F}}$ averages over seven semantic families as well as over samples, and a reader who computes each family's own rate from the per-model tables (Table 10, Table 11, Table 15) will find a sharp split: across the eleven calibrated configurations, *negation* clears the paraphrase floor on all eleven and *direction_flip* on nine, while *gripper_flip* and *location_swap* clear it on none and *subject_swap* on three. Read alone, that could put the unmatched paraphrase-floor comparison at the mercy of which families are pooled – a family-averaging artifact competing with §3.1's claim rather than supporting it.

It does not, for the same reason §3.2 gives for a different pattern in the same data. We computed each family's own share of records where the policy's argmax action bin moves at all ($1 - P(\Delta_\infty$=0), from the same 590 (run, family) cells §3.2 correlates against $\mathcal{F}$ at R^2=0.946) and compared it against how often that family clears the paraphrase floor. The two track closely: Spearman $\rho = 0.873$ (n=7 families, average-rank tie-breaking, Fisher-z 95% CI $[0.350, 0.981]$; Table 20). *negation* (clears 11 of 11, mean non-collision 0.686) and *direction_flip* (9 of 11, 0.747) are exactly the families whose edits reliably flip the argmax bin; *gripper_flip* (0 of 11, 0.331) and *location_swap* (0 of 11, 0.266) rarely do. A family that mechanically produces a large action-space perturbation clears the floor regardless of how much meaning it changes, and vice versa, independent of meaning – the description of §3.2 one level down.

Tag-locality-matched subset. *paraphrase_null* substitutes synonyms within `MOVE`/`SUBTASK` only (§B.1); of the seven semantic families pooled into $\bar{\mathcal{F}}_{\text{sem}}$, only *negation* (`SUBTASK`/`MOVE`) and *direction_flip* (`MOVE`) share that exact tag locality, while the other five additionally touch `PLAN`, `TASK`, or `VISIBLE OBJECTS`. Table 19 already has what this needs: averaging just its *negation* and *dir. flip* columns per row, in place of all seven, puts **9 of 11** configurations at or above their own paraphrase floor – the opposite sign from the pooled-seven headline, on the same nine rows. All nine are strict reversals, not eight-plus-a-tie: on `ECoT-bridge` specifically, *negation* ties its own floor exactly (0.947 against 0.947) but *direction_flip* clears it outright (0.963), so the pair's mean (0.955) clears too, by the narrowest margin of the nine (+0.008, against +0.020 for the next-narrowest, `DeepThinkVLA-base`). The two that do not flip, `DeepThinkVLA-SFT` and

Table 19: The full per-config, per-family matrix behind the split, and why it splits. Each cell is that family's $\mathcal{F}$ for that configuration; bold marks a cell at or above the row's own paraphrase floor (*para* column; one cell ties it exactly): the same convention as nowhere else in this paper, used here because the question this table answers is specifically which cells clear it, not which model scores highest. Bottom two rows: *clear rate* is the fraction of the eleven rows bolded in that column; mean $1-P(\Delta_\infty{=}0)$ is that family's own non-collision share of the 590 (run, family) cells of §3.2. Columns sorted by clear rate, matching Table 20's ordering. †: single-seed row (DeepThinkVLA; the ECoT-bridge row is its three-seed sweep), as in Table 15.

Config	para	negation	dir. flip	verb sw.	subj. sw.	adv. pl.	grip. flip	loc. sw.
Ours no-CoT	0.193	**0.197**	**0.274**	0.163	0.165	0.180	0.067	0.118
Ours r=8	0.587	**0.673**	**0.696**	**0.643**	0.311	0.311	0.110	0.100
Ours r=16	0.547	**0.660**	**0.749**	**0.677**	0.369	0.364	0.167	0.140
Ours r=32	0.567	**0.593**	**0.652**	**0.587**	0.335	0.335	0.130	0.131
Ours r=64	0.657	**0.777**	**0.823**	**0.840**	0.403	0.369	0.173	0.217
Ours data-50A	0.537	**0.567**	**0.699**	**0.647**	0.306	0.316	0.117	0.100
Ours data-50B	0.447	**0.510**	**0.635**	**0.540**	0.272	0.257	0.147	0.100
ECoT-bridge	0.947	**0.947**	**0.963**	**0.950**	0.932	0.932	0.697	0.593
DeepThinkVLA-base†	0.960	**0.990**	**0.970**	**0.990**	**1.000**	**0.967**	0.830	0.879
DeepThinkVLA-SFT†	0.810	**0.830**	0.650	0.730	**0.883**	**0.917**	0.430	0.466
DeepThinkVLA-RL†	0.820	**0.860**	0.640	0.730	**0.850**	**0.900**	0.540	0.552
clear rate	—	1.000	0.818	0.727	0.273	0.273	0.000	0.000
mean $1-P(\Delta_\infty{=}0)$	—	0.686	0.747	0.700	0.482	0.480	0.331	0.266

Table 20: Per-family floor-clearing tracks per-family collision rate. *Clear rate*: fraction of the eleven calibrated configurations on which that family's $\mathcal{F}$ exceeds the model's own paraphrase floor (the bold cells of Table 19, by column). *Mean* $1-P(\Delta_\infty{=}0)$: that family's own share of the 590 (run, family) cells of §3.2 with a non-collided record, i.e. how often its edits move the argmax bin at all. Sorted by clear rate.

Family	below/above	clear rate	mean $1-P(\Delta_\infty{=}0)$
negation	0/11	1.000	0.686
direction_flip	2/9	0.818	0.747
verb_swap	3/8	0.727	0.700
subject_swap	8/3	0.273	0.482
adversarial_plausible	8/3	0.273	0.480
gripper_flip	11/0	0.000	0.331
location_swap	11/0	0.000	0.266

`DeepThinkVLA-RL`, are exactly the two rows where *direction_flip* alone already falls below the floor (Table 20's 2/9 column), so one low-clearing family is enough to hold the pair's mean below floor even with *negation* clearing on all eleven.

This does not come from nowhere: *negation* and *direction_flip* are also the two highest-clearing, highest-collision-rate semantic families (Table 20), which is the same mechanism the paragraph above already gives for the family split, not an independent cause. But the consequence is real and belongs on the record regardless of its cause: which comparison set a reader chooses for the semantic mean is not a free, inconsequential choice – pooling all seven versus the two tag-locality-matched families changes the sign of the headline comparison on 9 of 11 configurations. The number this paper reports throughout (§3.1 onward, Table 2) pools all seven and discloses that choice; a locality-restricted alternative was available, would have been at least as defensible *a priori*, and gives the opposite answer on most of this cohort.

Relation to the floor dependence. §3.1 already contains one sign flip that turns on which *floor* is read: $\bar{\mathcal{F}}_{\text{diff}}$ is negative against the paraphrase floor on 11 of 11 configurations but positive against the length-exact scramble floor on 10 of 11 instead (Table 2). This is a second, turning on which *families* are pooled into the semantic mean: it reverses the sign of the unmatched paraphrase-floor comparison on 9 of the same 11 configurations. In both, whichever side of the comparison changes the action more often reads as more sensitive (§3.2). §3.3 raises a related but distinct question – whether $\mathcal{F}_{\text{dir}}$'s

disagreement with $\mathcal{F}_{\text{mag}}$ at the top of the ranking is an artifact of the six displacing configurations' collapsed action space – and the answer there is no: a null built without reference to the edited action does not rescue the ranking either (§3.3, confound analysis below), so that disagreement is a separate finding, not a third instance of this one.

Movement rate and conditional size. $\mathcal{F} = P(\text{moved}) \cdot P(\text{faithful} \mid \text{moved})$, where *moved* means $\Delta_\infty \neq 0$ (the argmax bin changed at all) and *faithful* is the usual $\Delta_\infty > \tau$. This asks a sharper version of the question the family-heterogeneity check above answers at the family level: is the paraphrase floor easier to clear mainly because a meaning-preserving rewording perturbs *something* more often than a semantic edit does, or because, conditional on perturbing something, that perturbation tends to be small? Table 22 answers it, recomputed from the same per-sample records as §3.2, on all 11 calibrated configurations. *paraphrase_null* moves the argmax bin at all more often than the pooled seven semantic families on **11 of 11** configurations ($+0.018$ to $+0.164$ in raw rate) – a meaning-preserving synonym swap crosses *some* decode boundary more often than an edit designed to change what the robot does. Conditioned on having moved, whether the move clears τ is close between the two and the sign is not stable: semantic exceeds paraphrase on **4 of 11** configurations (`no-CoT`, `r=64`, `DT base`, and `r=16` by the largest margin, $+0.044$), with every residual, either sign, at most 0.091 in magnitude – smaller than the largest $|\bar{\mathcal{F}}_{\text{diff}}|$ (0.180; 0.220 once verb_swap is dropped, Table 25), though not by an order of magnitude. Most of the raw paraphrase-floor gap is therefore attributable to *how often* an edit reaches any decode boundary at all, not to semantic edits producing systematically smaller movements once they do reach one: a sharper, sample-level restatement of §3.2's diagnosis, not a competing explanation for it, and it leaves the unmatched paraphrase-floor comparison of §3.1 unchanged.

Symmetric decomposition of the gap. $\mathcal{F}_P - \mathcal{F}_S = p_P q_P - p_S q_S$ splits into a movement-rate term and a conditional-size term, but the naive split (holding one factor's baseline fixed while differencing the other) is base-dependent – which factor you hold fixed changes the answer, the standard index-number problem. We use the symmetric two-factor decomposition instead, averaging both base choices, which sums to the gap exactly with no residual: $G_{\text{move}} = (p_P - p_S)\frac{q_P+q_S}{2}$, $G_{\text{cond}} = (q_P - q_S)\frac{p_P+p_S}{2}$, $G_{\text{move}} + G_{\text{cond}} = \mathcal{F}_P - \mathcal{F}_S$ (verified to $< 10^{-16}$ residual on every row). Table 21 reports G_{move} as a percentage of the total gap. The movement-rate component is the majority explanation on **9 of 11** configurations (median 91.8%; shares above 100% occur where G_{cond} is negative, so the two terms, not the share, are the quantity to read); the two exceptions are both DeepThinkVLA (32.9% on `DT SFT`, 39.6% on `DT RL`), where the conditional-size term does most of the work instead. "Most" in the paragraph above is this: a 91.8% median, not a narrative gloss on the comparison, and one architecture family where it is not the majority term, disclosed rather than averaged away.

Table 21: The paraphrase-floor gap $\mathcal{F}_P - \mathcal{F}_S$ (para minus pooled semantic), symmetrically decomposed into a movement-rate component G_{move} and a conditional-size component G_{cond} (definitions above); % is G_{move} as a share of the gap. $G_{\text{move}} + G_{\text{cond}}$ reproduces the gap exactly (residual $< 10^{-16}$, not shown). The gap pools records over the seven semantic families (a micro-average), whereas Table 2's $\bar{\mathcal{F}}_{\text{diff}}$ averages per-family rates (a macro-average), so the two differ in magnitude and sign convention (e.g. `r=8`: gap 0.158 against $\bar{\mathcal{F}}_{\text{diff}} = -0.180$).

Model	gap	G_{move}	G_{cond}	% move
`no-CoT`	0.026	0.029	−0.003	112%
`r=8`	0.158	0.150	0.008	95%
`r=16`	0.079	0.104	−0.026	132%
`r=32`	0.155	0.135	0.020	87%
`r=64`	0.116	0.126	−0.009	108%
`data-50A`	0.123	0.113	0.010	92%
`data-50B`	0.075	0.063	0.012	84%
`ECoT-bridge`	0.089	0.075	0.014	85%
`DT base`	0.014	0.017	−0.004	126%
`DT SFT`	0.120	0.039	0.080	33%
`DT RL`	0.104	0.041	0.063	40%

Independent cells. The 590 cells behind R^2=0.946 are not independent: seeds, retraining replicates and confound re-runs of the same checkpoints each contribute cells, and $\mathcal{F} \leq 1 - P(\Delta_\infty{=}0)$ holds

Table 22: Decomposing the floor comparison into movement probability and, conditional on movement, whether it clears τ=0.05. *p(moved)*: fraction of samples with $\Delta_\infty \neq 0$. *p(faithful|moved)*: of those, the fraction with $\Delta_\infty > \tau$. sem(7): pooled over the seven non-control semantic families.

	p(moved)		p(faithful|moved)	
Model	para	sem(7)	para	sem(7)
`no-CoT`	0.233	0.199	0.829	0.843
`r=8`	0.637	0.473	0.921	0.908
`r=16`	0.640	0.521	0.854	0.898
`r=32`	0.567	0.429	1.000	0.961
`r=64`	0.690	0.559	0.952	0.967
`data-50A`	0.560	0.441	0.958	0.938
`data-50B`	0.467	0.400	0.957	0.929
`ECoT-bridge`	0.990	0.912	0.970	0.955
`DT base`	0.990	0.972	0.970	0.973
`DT SFT`	0.910	0.863	0.890	0.800
`DT RL`	0.920	0.872	0.891	0.821

in every cell by construction. Keeping one run per (configuration, family) – seed 0 of every three-seed run and the single run otherwise, 138 cells over the 11 configurations – gives R^2=0.975, and the relation holds within each lineage separately (Table 23). The content not implied by the inequality is the conditional rate $q = \mathcal{F}/(1 - P(\Delta_\infty{=}0)) = P(\Delta_\infty{>}\tau \mid \Delta_\infty{\neq}0)$: the inequality bounds $q \leq 1$ but not its spread, and q is concentrated near 0.9 in both lineages. Two lineages are too few clusters to fit a lineage random effect (§B.4.3), so we report them separately rather than pooled.

Table 23: $\mathcal{F}$ against $1 - P(\Delta_\infty{=}0)$ on all cells and on independent cells (one run per configuration and family), and the conditional rate q (median and interquartile range over cells with any movement).

Cells	n	R^2	q median (IQR)
all (run, family) cells	590	0.946	0.929 (0.83–0.99)
independent, both lineages	138	0.975	0.905 (0.80–0.98)
independent, ECoT lineage	102	0.989	0.909 (0.82–0.98)
independent, DeepThinkVLA	36	0.928	0.898 (0.78–0.96)

Table 24: $P(\Delta_\infty{=}0)$ by lineage and competence, on the independent cells of Table 23 (identity control excluded): median and interquartile range over (configuration, family) cells. Exact no-change is concentrated in the weak LoRA policies; ECoT-bridge and DeepThinkVLA rarely produce it, yet show the same floor dependence.

Group	cells	median	IQR
six LoRA variants, CoT target	72	0.51	0.32–0.76
`no-CoT` LoRA	12	0.83	0.80–0.85
ECoT-bridge	10	0.02	0.01–0.06
`DT base` (pretrain only)	11	0.01	0.00–0.02
DT SFT/RL (competent)	22	0.07	0.02–0.23

B.4.6 JUDGE VALIDATION OF THE EDIT GENERATORS

Every result in this paper depends on generator claims we asserted from source code but never measured: that *paraphrase_null* preserves meaning, that *bbox_jitter_null* perturbs a number without touching semantics, that *adversarial_plausible* substitutes an object that stays visually plausible. These are not decorative: $\mathcal{F}_{\text{diff}}$ is a floor-corrected statistic *only* if the floor family really is meaning-preserving, so the paraphrase-floor comparison of §3.1 ($\bar{\mathcal{F}}$ below the floor on 11 of 11 configurations) rested on an unmeasured premise. So we measured it.

Protocol. An independent LLM judge (`Qwen2.5-7B-Instruct`, neither of the architecture families under test) reads each (original, edited) reasoning pair and answers four binary questions (meaning preserved, same referent, same direction, still plausible) plus a 1–5 fluency rating. It is blind to which family produced the pair and never sees the model's action. 480 pairs were constructed over 40 samples $\times$ 11 families plus an identity control; 437 were judged and all 43 skips are itemized,

every one of them a case where the *generator* found no applicable edit (16 *adversarial_plausible*, 16 *subject_swap*, 10 *location_swap*) plus one identical render.

Gates. A judge that answers "preserved" unconditionally would validate every null in the paper, so the identity check alone is not enough. (a) *Identity control*: on 40 pairs where the edit is $X \to X$, preserved rate $= \mathbf{1.000}$. (b) *Negative control*: pooled over the families designed to destroy meaning, preserved rate $= \mathbf{0.175}$, which is what rules out the degenerate judge the first gate would pass. (c) *Order agreement*: presenting the two traces in the opposite order changes the verdict on 77 of 437 pairs, so agreement is $\mathbf{0.824}$, a real presentation bias, and the noise floor to attach to every rate below. All three pass and the run is marked valid.

A second judge. `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023) fails its negative-control gate (it rates 0.943 of meaning-destroying edits as preserved, against Qwen's 0.175), so it is invalid by its own gate and not used as corroboration.

Premises. *paraphrase_null* preserves meaning at 0.975 (Wilson 95% CI $[0.871, 0.996]$, n=40; fluency 4.92) and *bbox_jitter_null* at 1.000 ($[0.912, 1.000]$, n=40; fluency 5.00): the two nulls $\mathcal{F}_{\text{diff}}$ and the numeric-floor argument depend on are validated. *syntactic_scramble* also comes back at 1.000 ($[0.912, 1.000]$). These intervals are wide at n=40 (paraphrase_null's lower bound sits 10 pp below its point estimate), the cost of a single-judge, n=40-per-family design. *adversarial_plausible* changes the referent on 0.958 of pairs but is judged plausible on only $\mathbf{0.125}$, so it is an unconstrained object swap; no claim turns on its plausibility.

A fourth null, added later and validated separately (the length check of §3.2): *paraphrase_null_lenexact* preserves meaning at 1.000 with fluency 4.92, on its own 40-pair run against a *direction_flip* negative control (n_{pairs}=120, 119 judged, identity/negative/order gates all pass). This validates the generator once, on the one checkpoint (`r=32`) §3.1 floors it on; unlike *paraphrase_null* and *bbox_jitter_null*, which the 437-pair sweep (§2) validates across the full judged cohort, we do not have a judge-gated meaning-preservation or fluency score for this null on any other checkpoint.

Scramble and verb swap. *syntactic_scramble* is a second within-CoT floor independent of *paraphrase_null*'s lexical route. *verb_swap* preserves meaning on 0.575 of pairs, so it is a weaker semantic edit than its name suggests; the families with no declared premise behave as intended (*negation* preserved on 0.100, *direction_flip* on 0.200).

Dropping *verb_swap*. Recomputing $\bar{\mathcal{F}}_{\text{diff}}$ over the remaining six families (Table 25), the sign does not change on any configuration, and the gap widens on 10 of 11, because *verb_swap* has the third-highest clear rate of the seven semantic families (§B.4.5).

Table 25: Sensitivity of the unmatched paraphrase-floor comparison to dropping *verb_swap*. $\bar{\mathcal{F}}_{7,\text{diff}}$: the published statistic (seven semantic families, Table 2). $\bar{\mathcal{F}}_{6,\text{diff}}$: the same statistic over the remaining six. Both against each model's own paraphrase floor; both negative on all 11 rows.

Model	$\bar{\mathcal{F}}_{7,\text{diff}}$	$\bar{\mathcal{F}}_{6,\text{diff}}$	sign unchanged
`no-CoT`	−0.027	−0.027	yes
`r=8`	−0.180	−0.220	yes
`r=16`	−0.100	−0.138	yes
`r=32`	−0.172	−0.204	yes
`r=64`	−0.142	−0.196	yes
`data-50A`	−0.144	−0.186	yes
`data-50B`	−0.095	−0.127	yes
`ECoT-bridge`	−0.088	−0.103	yes
`DT base`	−0.013	−0.021	yes
`DT SFT`	−0.109	−0.114	yes
`DT RL`	−0.095	−0.096	yes

Likelihood and fluency. A length-exact paraphrase, *paraphrase_null_lenexact*, changes 8.7 words and floors at 0.560 on `r=32` (n=300), close to *paraphrase_null*; *paraphrase_null* also adds 0.97 words on average. The policy's own log-likelihood of the edited CoT has partial $R^2 \leq 0.037$ across

Table 26: InstructVLA (continuous flow head): semantic mean minus each meaning-preserving floor at τ=0.05 with bootstrap 95% intervals; the largest $P(\Delta_\infty{=}0)$ over edit families; $\mathcal{F}_{\text{dir}}$ on *direction_flip*; $\mathcal{F}$ when only the flow noise is resampled; and $\mathcal{F}$ of the identity edit under shared noise.

Format, reasoning	$\bar{\mathcal{F}}-$ floor			max $P(\Delta{=}0)$	$\mathcal{F}_{\text{dir}}$	noise only	identity
primitive, gold	*verb_paraphrase*: +0.636 [+0.589, +0.681]	*reorder*: +0.647 [+0.579, +0.714]	*punct*: +0.920 [+0.898, +0.941]	0.00	0.405	0.999	0.000
primitive, self-gen.	*verb_paraphrase*: +0.635 [+0.589, +0.681]	*reorder*: +0.684 [+0.604, +0.759]	*punct*: +0.875 [+0.851, +0.899]	0.00	0.373	0.993	0.000
free text, gold	*verb_paraphrase*: +0.357 [+0.298, +0.407]	*contraction*: +0.523 [+0.487, +0.561]	*punct*: +0.190 [+0.150, +0.232]	0.00	—	0.991	0.000
free text, self-gen.	*verb_paraphrase*: +0.385 [+0.326, +0.439]	*contraction*: +0.530 [+0.493, +0.566]	*punct*: +0.190 [+0.147, +0.235]	0.00	—	0.991	0.000

the 6 checkpoints tested, and judged fluency (4.00 for *syntactic_scramble*, 4.92 for *paraphrase_null*) has partial $R^2(\text{fluency}) = 0.056$, with the wrong sign ($n$=236).

Scope. The judge validates the *generators*, not the specific pairs that produced any published number: demo/step identity is not recoverable without re-iterating the 17 GB tfds shards, so the judge re-runs the same generators over the same corpus, restricted to the demo files the scored run drew from. It is also one judge model at n=40 per family, with a measured 17.6% order-flip rate, so these rates should be read to two decimal places at most.

A programmatic check. A lexical-overlap check without an LLM does not corroborate the judge on *paraphrase_null*: its reasoning-span TF-IDF cosine (0.831) is lower than for *direction_flip* (0.955) and *negation* (0.960), with anchors behaving as expected (the identity control 1.000, *cross_task_swap* 0.228), because the paraphrase changes more words (20.85). Lexical overlap tracks how much text changed, not meaning, which is the edit-size point of §3.2.

B.4.7 Two continuous-head VLAs

InstructVLA. This is a flow-matching action head on an Eagle2-2B VLM (checkpoint `instructvla_finetune_v2_xlora_freeze_head_instruction`). We score it in distribution on Bridge V2 with its own normalization, on 300 samples, in both of its reasoning formats: a short *primitive* ("move up, open gripper") and a *free-text* answer. We edit either the gold reasoning or the model's own self-generated reasoning.

- Noise. The flow noise is held fixed across the two decodes of a pair (common random numbers, 16 shared seeds per sample), so the identity edit scores at most 0.0002.
- Floors. The meaning-preserving floors are a verb paraphrase, a conjunct reorder and a punctuation change for the primitive format, and a verb paraphrase, a contraction and a punctuation change for the free-text format.
- Competence. Open-loop, the model beats a constant-input baseline on 84% of samples (L1 gain 0.0079, 95% CI $[0.0068, 0.0090]$).

MolmoAct2-Think-LIBERO. This model reasons in a 10×10 grid of depth tokens and carries two action heads on the same weights: a continuous flow head and a discrete FAST head. We edit its self-generated depth grid on 200 LIBERO samples.

- Edits. Seven depth edits (mirror, flip, invert, roll, and inversion of 5 cells) and two depth nulls (a ± 1 jitter of all cells or of 5 cells). A uniform-random grid and a cross-sample swap serve as ceilings.
- Competence. It beats a constant-input baseline on 99.5% of samples (L1 gain 0.151).

Both runs are open loop. Both checkpoints declare no license upstream; they are used for research evaluation only and not redistributed.

Reading.

- *All floors agree in sign.* On InstructVLA no edit leaves the action exactly unchanged, and the semantic mean exceeds every meaning-preserving floor, by +0.64 to +0.92 in its primitive format. Its floors change as many words as its semantic edits (1.0 and 3.3 against 1.0–2.8; in free text *role_swap* changes 3.6 words against 1.0–2.0), and *direction_flip* reverses the

Table 27: MolmoAct2-Think-LIBERO, the same weights under two action heads: $\mathcal{F}$ range, largest mean Δ_∞ and $P(\Delta_\infty{=}0)$ range over the nine depth-reasoning edits, nulls and ceilings; $\mathcal{F}$ of a five-token instruction substitution; and the identity edit.

Head	$\mathcal{F}$ (depth edits)	max Δ_∞	$P(\Delta{=}0)$	instr. sub.	identity
continuous (flow)	0.000–0.000	0.0010	0.00–0.00	0.279	0.000
discrete (FAST)	0.005–0.025	0.0062	0.84–0.97	0.455	0.000

translation on 0.40 (gold) and 0.37 (self-generated) of samples, against a family ceiling of at most 0.012. This is consistent with the token-decoded scope of the main claims; it does not show that continuous heads are calibrated.

- *Fixed noise is required.* Resampling only the flow noise gives $\mathcal{F} = 0.99$, and an instruction substitution gives 0.91–0.99. A magnitude score at $\tau{=}0.05$ on a sampled head is saturated by the sampler unless the noise is shared across the pair, which we add to the reporting checklist.
- *A magnitude score can report a null CoT correctly.* On MolmoAct2's continuous head, every depth edit, including a random grid, moves the action by at most 0.001 ($\mathcal{F} = 0.000$), while an instruction substitution moves it past τ on 0.28 of samples. Its depth reasoning is not load-bearing for its continuous actions. Its discrete head leaves the action bin-identical on 84–97% of the same edits.

This is two models, open loop, 300 and 200 samples; MolmoAct2's depth edits move its continuous action by at most 0.001 and its discrete action past τ on at most 0.025 of samples. Because InstructVLA's floors are roughly size-matched to its semantic edits, the sign agreement does not isolate head type; it shows neither that the floor dependence is specific to token-decoded heads nor that continuous-head magnitude scores are calibrated.

B.4.8 EDIT-SIZE-MATCHED NULLS

Construction. Two size-matched nulls plus an in-vocabulary reorder reported as a caveat, all built and judged before any action was decoded with them. *paraphrase_null_matched_dir* edits the tags *direction_flip* edits (`MOVE`, `MOVE REASONING`) and swaps exactly as many word positions as *direction_flip* changes on that sample, using single-word synonyms placed nearest to the positions *direction_flip* edits; the synonym list excludes every direction, location, gripper and *verb_swap* word. Changed-word counts are equal on 890 of 890 corpus pairs; 11% of samples have too few swappable words and are inapplicable. *paraphrase_null_matched_neg* makes *negation*'s two-word prefix insertion in the same tags (`SUBTASK`, `MOVE`), "you should X" for "do not X" (equal on 1000 of 1000). *paraphrase_null_invocab* replaces `MOVE` with the most frequent verbatim LIBERO-ECoT training `MOVE` string that has the same (verb, direction) primitive multiset and speed modifier, drawn from 2,318 training strings; in practice it reorders conjuncts ("move back and left" → "move left and back") and applies only to compound `MOVE`s (83 of 300 samples per configuration). Edit size is the number of changed word positions over the full rendered CoT, counting a replaced, inserted or deleted block as its longer side (Table 28). This differs from the reasoning-span token count of §B.4.6 ($20.85/8.6/4.0$), which gives the same ordering.

Judge. Same judge and protocol as §B.4.6, 866 of 960 pairs judged, all three gates passing (identity 1.00, pooled negative control 0.071, order agreement 0.886). Meaning is preserved on every judged pair of all three nulls: *matched_dir* $n{=}74$ (same direction 0.932, fluency 4.80), *matched_neg* $n{=}120$ (same direction 1.000, fluency 4.25), *invocab* $n{=}75$ (same direction $\mathbf{0.747}$, fluency 4.68); *paraphrase_null* reads 1.000 in the same run (fluency 4.88).

Sweep and statistics. First step, teacher-forced, $\tau{=}0.05$, the same observations and sampling seeds as the canonical runs (three seeds of 100; `DT base` one seed; DeepThinkVLA SFT/RL at three seeds here against one in Table 2). Every family is re-scored within the same job, and the sweep reproduces the published paraphrase and scramble floors exactly on every configuration that has three seeds in both runs. Every semantic family also reproduces except *location_swap*: its generator has since been corrected for idiom false positives, leaving 194 rather than 221 applicable samples, which moves $\bar{\mathcal{F}}_{\text{diff}}$ by at most 0.008 against Table 2. The contrasts are paired on the samples where both

families apply; intervals are observation-clustered BCa bootstraps (20,000 draws); Holm correction runs across the 11 configurations within each test. Table 30 counts every test run.

Table 28: Changed words over the full rendered CoT, 1000 step-0 LIBERO-90 samples. n: samples where the edit applies.

Family	n	mean	median	p90
direction_flip	996	3.76	3	6
gripper_flip	1000	2.54	2	4
verb_swap	1000	8.37	8	10
negation	1000	3.99	4	4
subject_swap	722	16.87	16	24
location_swap	589	4.68	2	6
adversarial_plausible	722	15.26	14	21
paraphrase_null	1000	10.60	10	12
paraphrase_null_lenexact	1000	8.72	8	10
syntactic_scramble	991	6.20	6	10
paraphrase_null_matched_dir	890	3.51	3	6
paraphrase_null_matched_neg	1000	3.99	4	4
paraphrase_null_invocab	271	4.06	4	4

Reading. The matched contrasts give one sign on every configuration for *direction_flip*, and *negation* fails to clear only where both rates are near ceiling (ECoT-bridge, DeepThinkVLA) or the policy has no CoT target (`no-CoT`). The pooled semantic mean against a matched null is mixed (5 of 11 positive against *matched_dir*), as it should be: that pool mixes 2–4-word edits with 15–17-word ones (Table 28), so it is not itself size-matched to anything. In this re-scored sweep the unmatched counts are 0 of 11 positive against *paraphrase_null* and 9 of 11 against *syntactic_scramble*; the second differs from Table 2's 10 of 11 because DeepThinkVLA SFT/RL gained two seeds. The movement-rate term is the larger share of the *direction_flip* gap on 9 of 11 configurations; the exceptions are `DT base` and `DT SFT`, where the conditional term is as large or larger, as it was for the paraphrase gap (Table 21).

The in-vocabulary reorder. On the six CoT-trained LoRA variants this meaning-preserving reorder, which changes about as many words as the semantic edits, moves the action past τ on 0.940–0.988 of applicable samples, and the semantic mean falls significantly below it on all six. Even a size-comparable meaning-preserving null is not unique, so one control is not enough, which is why closed-loop success is the arbiter. We do not use it as a floor because the judge rates it direction-preserving on only 0.747 of pairs.

B.4.9 Rule-based oracle positive control

Every $\mathcal{F}_{\text{dir}}$ elsewhere in this paper is measured on a trained neural policy, so a low score is always compatible with an explanation this paper has not been arguing for: the instrument itself cannot detect faithfulness even when it is genuinely present. We rule this out with a "policy" that has no network in it at all. The oracle reads the MOVE phrase text and nothing else (no image, no instruction, no learned weights) and returns a fixed 7-DoF vector: each of the eight direction words *direction_flip* itself flips (the same word-pair table that edit family uses) is mapped by a hand-written table to ± 1 on one of three translation axes (`left`/`right` $\to y$, `forward`/`back` $\to x$, `up`/`above`/`down`/`below` $\to z$), multiple words in one phrase compose, and the same string always returns the same vector; rotation and gripper are fixed at 0. This is a claim about what a MOVE-conditioned action generator, built to be one, does – not about what a real policy does.

Applied to the judge-validation pairs (the same (a_head, b_head) text pairs §B.4.6 scores, not a new capture): *paraphrase_null* (N=40) substitutes `move`→`shift` and leaves every direction word untouched, so the oracle's action is byte-identical on all 40, $\Delta_\infty = 0$ exactly, $\mathcal{F}_{\text{mag}} = \mathcal{F}_{\text{dir}} = 0.000$. *direction_flip* ($N$=40) inverts them: $\mathcal{F}_{\text{mag}} = 1.000$ (every sample moves), $\mathcal{F}_{\text{dir}} = 0.925$ (37 of 40 reverse, mean cosine -0.925). The three non-reversing samples are a named limit of this simple parser, not noise: each mixes a translation direction word with a same-named *rotation* direction word in one phrase (e.g. "move down and forward and left and *rotate up*"), and a word-only parser lets the later match overwrite the earlier one on the shared z axis. A clause-scoped parser would likely

Table 29: Size-matched contrasts per configuration. *dir*, *m-dir*: $\mathcal{F}$ of *direction_flip* and its matched null; Δ: paired difference on shared samples with its BCa 95% interval, * Holm-significant across the 11 configurations; *move*: share of the pooled *direction_flip* gap assigned to movement rate by the symmetric decomposition of §B.4.5. The same columns follow for *negation*. *invocab*: $\mathcal{F}$ of the in-vocabulary reorder on its 83 applicable samples (30 for `DT base`). Paired differences can differ from the difference of the two rates because the matched null is inapplicable on 11% of samples.

	direction_flip					*negation*				
Model	dir	m-dir	Δ	BCa 95%	move	neg	m-neg	Δ	BCa 95%	invocab
`no-CoT`	0.274	0.138	$+0.142^*$	$[+0.094, +0.195]$	87%	0.197	0.207	-0.010	$[-0.057, +0.030]$	0.325
`r=8`	0.696	0.558	$+0.092^*$	$[+0.059, +0.134]$	60%	0.673	0.453	$+0.220^*$	$[+0.170, +0.267]$	0.976
`r=16`	0.749	0.423	$+0.288^*$	$[+0.235, +0.345]$	61%	0.660	0.453	$+0.207^*$	$[+0.147, +0.263]$	0.964
`r=32`	0.652	0.475	$+0.124^*$	$[+0.088, +0.169]$	87%	0.593	0.350	$+0.243^*$	$[+0.183, +0.300]$	0.988
`r=64`	0.823	0.646	$+0.150^*$	$[+0.103, +0.202]$	81%	0.777	0.483	$+0.293^*$	$[+0.223, +0.357]$	0.940
`data-50A`	0.699	0.431	$+0.223^*$	$[+0.175, +0.276]$	90%	0.567	0.383	$+0.183^*$	$[+0.130, +0.233]$	0.976
`data-50B`	0.635	0.412	$+0.177^*$	$[+0.134, +0.227]$	98%	0.510	0.367	$+0.143^*$	$[+0.080, +0.203]$	0.988
`ECoT-bridge`	0.963	0.907	$+0.054^*$	$[+0.015, +0.097]$	63%	0.947	0.900	$+0.047$	$[+0.010, +0.080]$	0.928
`DT base`	0.970	0.828	$+0.149^*$	$[+0.082, +0.236]$	25%	0.990	0.970	$+0.020$	$[-0.020, +0.050]$	0.900
`DT SFT`	0.643	0.438	$+0.190^*$	$[+0.118, +0.260]$	49%	0.817	0.813	$+0.003$	$[-0.050, +0.050]$	0.614
`DT RL`	0.616	0.422	$+0.182^*$	$[+0.108, +0.258]$	66%	0.830	0.810	$+0.020$	$[-0.030, +0.063]$	0.566

Table 30: Every test in the size-matched sweep, counted across the 11 configurations: point estimate positive, and Holm-significant positive or negative.

Test	positive	Holm +	Holm −
direction_flip − matched null (primary)	11/11	11	0
negation − matched null (primary)	10/11	6	0
$\bar{\mathcal{F}}$ − *paraphrase_null*	0/11	0	9
$\bar{\mathcal{F}}$ − *syntactic_scramble*	9/11	2	0
$\bar{\mathcal{F}}$ − *matched_dir*	5/11	3	5
$\bar{\mathcal{F}}$ − *matched_neg*	3/11	0	3
$\bar{\mathcal{F}}$ − *invocab* (its subset)	3/11	1	6

recover all three; we did not build one, since 0.925 is already above every real checkpoint's $\mathcal{F}_{\text{dir}}$ (at most 0.779, `r=64`).

$\mathcal{F}_{\text{dir}} = 0.925$ against `ECoT-bridge`'s 0.117 (Table 5) is an $8\times$ gap, the collapsed LoRA variants reaching at most 0.779, and $\mathcal{F}_{\text{dir}} = 0.000$ on *paraphrase_null* is the same floor-respecting behaviour §3.1 asks of every real model. Both numbers are consequences of the construction, checkable by reading the oracle function, not measurements with sampling uncertainty – which is why n=40 per family is enough here though it would not be for a comparative claim between two real checkpoints. The construction is deliberately narrow: it says the diagnostic can register a difference this large when one is really there, not that any real checkpoint's low score shares the oracle's mechanism, and not that the oracle's own $\mathcal{F}_{\text{mag}} = 1.000$ says anything about a family other than the two tested here.

Mixing-weight sweep. We build synthetic policies on the same *direction_flip* pairs: $a(\alpha) = \alpha \cdot f(\text{CoT}_{\text{edited}}) + (1-\alpha) \cdot g(\text{CoT}_{\text{ignored}})$, where f is the oracle above applied to the edited MOVE phrase and g the same oracle applied to the *original* phrase (a policy that ignores the CoT edit). At α=1 this is the oracle above ($\mathcal{F}_{\text{dir}} = 0.925$); at α=0 the "edited" action is byte-identical to the original, so $\mathcal{F}_{\text{dir}} = 0$ by construction, the same way *paraphrase_null* is. Sweeping $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$: $\mathcal{F}_{\text{dir}}(\alpha)$ is 0.000 for $\alpha \le 0.4$ and 0.925 for $\alpha \ge 0.6$, monotone non-decreasing across the full grid with no reversal, mean cosine moving from $+1.000$ (α=0) to -0.925 (α=1) through $+0.987$ at α=0.4 and -0.885 at α=0.6. This is not a smooth ramp: *direction_flip*'s two endpoints are near-antipodal, so linearly mixing them scales the vector to (numerically) zero at α=0.5 – only 3 of 40 pairs have a defined cosine there, the same 3 exceptions above – and flips sign on either side, so $\mathcal{F}_{\text{dir}}$ steps rather than ramps. The sweep confirms monotonicity but on a two-point-mixture construction, not a smoothly graded one, so it does not by itself establish that $\mathcal{F}_{\text{dir}}$ resolves finer degrees of partial CoT-dependence. *This is a sensitivity validation, not a calibration one*: it shows $\mathcal{F}_{\text{dir}}$ responds monotonically to a constructed, varying degree of CoT-dependence, not that its absolute value or

its $\cos < -0.5$ threshold is calibrated on any real, learned policy – that is what the per-model null ceiling in Table 5 is for (§3.3).

B.5 VERIFICATION

The primary per-sample edit files hold 52,338 per-sample edit records (45,989 scored); the rest are skipped because the edit target is absent from the CoT. Counting the replicate and confound runs stored alongside them, the collision analysis of §3.2 covers 52,922 scored records from 76 runs. Our audit script asserts 2,057 claims against these and every other artifact and exits non-zero on any mismatch. Δ_∞ is a within-policy quantity and is not invalidated by weak policies, but six of the eight three-seed rows come from a policy whose single-step open-loop prediction does not beat a constant (§B.6); the behavioural checks are an eight-scene trajectory probe on `r=32` (Stage 6a) and closed-loop rollouts of the two competent DeepThinkVLA policies (Stage 6b, §B.6.3).

B.6 LIMITATIONS

Each limitation the body lists is stated here once, with its scope and what mitigates it. The central one is already the paper's finding: on token-decoded heads the unmatched-floor verdict is not identified.

(1) The token-decoded cohort is mostly weak. For the LoRA variants the in-suite replay gives $0/20$ for every arm and the paired rollout returns $0/40$ in both arms, a checkpoint-competence problem rather than a compute or suite-availability one, so an edit-induced ΔSR is undefined rather than zero; $8\times$ longer training (120,000 steps) still gives $0/10$ (Wilson 95% CI $[0, 0.28]$). ECoT-bridge cannot be rolled out on LIBERO: its released norm-stats carry only `bridge_orig`, so any action would score 0 independent of any edit, a corpus mismatch rather than a competence measurement; its own paper reports $66 \pm 3.8\%$ real-robot success on Bridge V2 (Zawalski et al., 2024). The rollout harness reproduces 0.857–$1.018\times$ the published success of four public LIBERO checkpoints, and the trajectory-level readout we substituted for SR on `r=32` and the closed-loop rollouts of §B.6.3 are reported above and below. The weak cohort also confounds $\mathcal{F}_{\text{dir}}$. The six configurations that clear its null most clearly have collapsed action spaces: at most one translation axis is off its zero bin on 72–84% of samples, so reversal is close to a sign test on one axis. Training `r=32` $8\times$ longer leaves both diagnostics near where they started ($\mathcal{F}_{\text{dir}}$ $0.582 \to 0.569$, clearance $7.9\times \to 10.0\times$; collapse $76.0\% \to 72.7\%$), and $\mathcal{F}_{\text{dir}}$ does not transfer to the second architecture family: all three DeepThinkVLA checkpoints fail it. *Mitigation:* the floor contradiction is shown first on the strongest models (Table 3): the two competent DeepThinkVLA policies at three seeds, and ECoT-bridge in distribution on Bridge V2 (§B.4.1). The closed-loop test uses only competent policies. OpenVLA-OFT (Kim et al., 2025) is excluded after failing to load at its authors' declared pins.

(2) Open-loop scores are first-step and the primary sweep is teacher-forced. A teacher-forced edit cannot rule out a memorized pairing between an annotation string and an action. *Mitigation:* self-generated CoT, one random-step Bridge V2 run (§B.4.1), and closed-loop rollouts with the edit applied at every re-plan where its target is present (§B.6.3). On self-generated CoT the floor comparison is mixed (§B.6.1): it replicates on `ECoT-bridge` and `DT-RL`, points the same way on `DT-SFT` without significance, reverses on `r=32`, depends on the statistic on `r=64`, and is inconclusive on `no-CoT`. Four generators (*subject_swap*, *adversarial_plausible*, *selfsplice_control*, *bbox_jitter_null*) yield no usable self-generated samples, because they look for substrings free-decoded text rarely contains.

(3) The matched contrast was adopted as primary after the sweep, and cannot separate reading from lexical priming. The size-matched nulls were built and judged before any action was decoded with them; choosing their contrast as primary came after. The action-only `no-CoT` control clears the direction null ($+0.142$). It was fine-tuned from the CoT-trained ECoT-bridge and may inherit direction-word sensitivity. Under self-generated CoT the same control also clears its own $\mathcal{F}_{\text{dir}}$ ceiling, at $1.94\times$. *Mitigation:* the paper reads the matched contrast as necessary, not sufficient, and adds the direction-aware diagnostic and the closed-loop test. A null controlling token identity is future work.

(4) The Bridge V2 identity control is not zero. On that corpus *selfsplice_control* scores 0.074–0.227, the same order as the matched contrasts there. *Mitigation:* the paired contrasts share each original decode. The in-distribution $\mathcal{F}_{\text{dir}}$ result is a cosine, not a difference of rates. The LIBERO identity control is 0.00 everywhere.

(5) Closed-loop coverage and headroom. *direction_flip* and *gripper_flip* apply at only 8% and 38% of re-plans, and clean success near 0.96 leaves little room for improvement. A same-task, other-episode CoT was not run, so the cross-task drop may reflect distraction by an out-of-distribution trace. *Mitigation:* the cross-task CoT is a positive control, showing the rollouts detect an effect. The empty CoT applies at every re-plan, and negation and object swaps at $\geq 91\%$ of re-plans; none of these lowers success.

(6) Continuous heads are only probed. There are two models, open loop (§B.4.7). InstructVLA's floors are roughly size-matched to its semantic edits, so sign agreement there does not isolate head type. MolmoAct2's depth reasoning is not load-bearing. *Mitigation:* none beyond scoping the claims to token-decoded heads.

(7) Direct tests not yet run. These are a within-generator dose–response over edit sizes $k{=}1, 2, 4, 8$, a resolution sweep of the floor verdict itself rather than of the collision rate, and validation on checkpoints trained with the CoT coupled to the action or shuffled against it.

Retraining variance. Both runs of each same-config pair share the model/LoRA-init and optimizer seed. The training code pins execution to one GPU, so batch size, learning rate and step count are identical across every pair. The bar therefore measures data-order and GPU-nondeterminism noise only. Measured against the widest Wilson half-width, per-family $\mathcal{F}$ moves by up to 0.317 and $\bar{\mathcal{F}}$ by up to 0.106, which is $1.6\times$ it. It likely understates full seed-to-seed variance, and it is not measured for ECoT-bridge or DeepThinkVLA. Per-family N varies from 58 to 100 with a visibility gate, so each model is compared only against its own floor and ceiling.

B.6.1 SELF-GENERATED COT

On each checkpoint's own free-decoded CoT, which removes any memorized pairing between an annotation string and an action, the floor comparison replicates on `ECoT-bridge` and `DT-RL`, points the same way on `DT-SFT` without significance, reverses on `r=32`, depends on the statistic on `r=64`, and is inconclusive on `no-CoT` (Table 31). The reversal is on `r=32`, one of the weak LoRA variants. For $\mathcal{F}_{\text{dir}}$ on *direction_flip*, `r=32` clears at 0.761 ($n{=}71$) against 0.582 teacher-forced, `ECoT-bridge` stays low at 0.151 ($n{=}53$) against 0.117, and `no-CoT`'s raw rate rises to 0.211 ($n{=}38$) against 0.087. Against its self-generated ceiling of 0.109 (*negation*), `no-CoT` *clears* at $1.94\times$: on this control, under self-generated CoT, $\mathcal{F}_{\text{dir}}$ fails its own calibration criterion.

Table 31: Self-generated-CoT paraphrase-floor comparison on the four checkpoints with a single-statistic comparison, against the teacher-forced (TF) number it replaces (Table 12); the two DeepThinkVLA checkpoints use a differently-shaped paired-rate comparison instead and are reported in text below. $\overline{\text{TV}}_{\text{diff}}$ is self-generated *direction_flip* vs. *paraphrase_null*. `r=64`'s two entries are two different statistics that disagree in sign (5-family mean vs. *direction_flip* alone), computed separately rather than merged. `no-CoT` has no defined comparison.

Checkpoint	self-generated result	verdict
`ECoT-bridge`	$\overline{\text{TV}}_{\text{diff}} = -0.072$, sig. (TF: -0.072)	replicates closely
`r=32`	$\overline{\text{TV}}_{\text{diff}} = +0.021$, sig. (TF: -0.031)	reverses sign
`r=64`	5-fam. mean $+0.006$ n.s. / $\overline{\text{TV}} = 0.371$ against 0.182, $+0.189$, bootstrap 95% CI $[0.169, 0.208]$	statistic-dependent
`no-CoT`	—	inconclusive

On DeepThinkVLA ($n{=}1000$ self-generated samples), DT-RL replicates the teacher-forced comparison (paired bootstrap 95% CI $[-0.419, -0.047]$) and DT-SFT points the same way without significance ($[-0.390, 0.0]$); both give $\mathcal{F}_{\text{dir}}{=}0.000$ on *direction_flip*.

B.6.2 DIRECTION-AWARE SCORING: CRITERIA, NULLS AND CHECKS

Family ceiling. For each checkpoint, the family ceiling is the largest $\mathcal{F}_{\text{dir}}$ over the eight families with no direction to reverse, computed on the checkpoint's own de-quantization grid (Table 5); a signed edit's $\mathcal{F}_{\text{dir}}$ must exceed it.

direction_flip uses the translation cosine of Eq. 3. *gripper_flip* counts a sign flip of the gripper channel, defined only when both signs are nonzero. *negation* counts a reduction of $\|a\|_2$. A sample with a zero translation vector has an undefined cosine and is excluded; none occurs on *direction_flip*. ECoT-bridge moves on 288 of 299 samples (96%), so its low $\mathcal{F}_{\text{dir}}$ is not a lack of movement. An isotropic action distribution scores 0.25 at our $\cos < -0.5$ criterion, but that reference is the wrong one for a collapsed action space, hence the nulls below. The unedited scene geometry does not explain ECoT-bridge's non-reversal (Table 32). *direction_flip_no_geom* blanks the contradicting bounding boxes and gripper position: $\mathcal{F}_{\text{dir}} = 0.090 \pm 0.028$ against 0.117 ± 0.032, so removing the contradicting anchor makes the action more similar to the unedited one, the opposite of what geometric grounding predicts; the same check extended to the six LoRA/data variants falls on every row. *direction_flip_geom_consistent* mirrors the anchor to agree with the flip (n=96 paired): $\mathcal{F}_{\text{dir}} = 0.083$ against 0.062, mean cosine $+0.612$ against $+0.640$, a small move. Across the three geometry conditions, $\mathcal{F}_{\text{dir}}$ stays within 0.062–0.117 and the mean cosine within $+0.418$ to $+0.640$. The unedited-pair null gives a chance rate of 0.06–0.10 on the six collapsed configurations and 0.016 on `ECoT-bridge`; against it the six clear by 6.4–9.8× and `ECoT-bridge` by 7.3×, so correcting for the collapsed action spaces does *not* restore `ECoT-bridge` to first (Table 33).

Table 32: The two geometry checks on $\mathcal{F}_{\text{dir}}$ (*direction_flip*), by checkpoint. Both baselines are *direction_flip* restricted to the identical samples paired with that column's check, not the full sample Table 5 reports *direction_flip* on, so they differ slightly from Table 5 (e.g. `r=64`: 0.820 here vs. 0.779 there). *no-geom*: the contradicting `VISIBLE OBJECTS/GRIPPER POSITION` anchor blanked (ECoT-bridge n=298 pooled over 3 seeds; six LoRA/data rows, 3 seeds, N=100 each, paired subset). *geom-consistent*: the anchor mirrored to agree with the flipped direction instead, left/right and up/down flips only, a smaller paired subset (n=96 each), with its own paired baseline.

	no-geom		geom-consistent	
Checkpoint	contradicting	removed	contradicting (restr.)	agreeing
`ECoT-bridge`	0.117	0.090	0.062	0.083
`r=8`	0.649	0.639	0.396	0.406
`r=16`	0.653	0.471	0.521	0.531
`r=32`	0.586	0.439	0.354	0.344
`r=64`	0.820	0.500	0.406	0.417
`data-50A`	0.659	0.542	0.396	0.406
`data-50B`	0.526	0.305	0.250	0.260

Table 33: The same-observation null for $\mathcal{F}_{\text{dir}}$ on *direction_flip*, tabulated (the calibration §3.3 and the paragraph above describe in prose): for each configuration, real_rate is the matched-pair reversal rate (reproduces $\mathcal{F}_{\text{dir}}$ exactly); orig. null is $\cos(a_{\text{orig}}[i], a_{\text{orig}}[j])$, $i \neq j$, exhaustive over all ordered pairs of *unedited* actions (299×298 each, ecot-bridge and the seven `Ours` configurations only – the three single-seed DeepThinkVLA checkpoints are not resampled this way, but their raw $\mathcal{F}_{\text{dir}} \leq 0.010$ already sits below every null in this table). This null never reads a_{edit}, so it cannot be inflated by how well a model reverses, unlike the edited-permutation null $\cos(a_{\text{orig}}[i], a_{\text{edit}}[j])$ discussed above. `no-CoT` is the one configuration whose ratio is < 1: it does not clear its own null, as a control should not. `ECoT-bridge`'s ratio (7.35×) sits inside the six collapsed configurations' own range (6.36–9.81×), not above it.

Model	$\mathcal{F}_{\text{dir}}$	orig. null	ratio	clears
`ECoT-bridge`	0.117	0.016	7.35	yes
`r=64`	0.779	0.079	9.81	yes
`r=16`	0.659	0.085	7.73	yes
`data-50A`	0.642	0.101	6.36	yes
`r=8`	0.649	0.081	7.99	yes
`r=32`	0.582	0.091	6.42	yes
`data-50B`	0.532	0.060	8.89	yes
`no-CoT`	0.087	0.098	0.88	no

Ranking comparison and the three nulls. Over the eight three-seed configurations the $\mathcal{F}_{\text{mag}}$ and raw $\mathcal{F}_{\text{dir}}$ rankings correlate at Spearman $\rho = 0.476$ (Kendall $\tau_b = 0.500$; 21 concordant against 7 discordant pairs; 0.400–0.500 on the three sampling seeds; Fisher-z 95% interval of $[-0.34, 0.88]$ at n=8). The family ceiling is cleared by the six LoRA variants at 6.6–$11.1\times$ and by ECoT-bridge at $2.5\times$. The edited-permutation null is rejected because the treatment inflates it (Pearson $r = -0.94$ with the reversal rate; a perfectly reversing oracle scores $2.5\times$, below ECoT-bridge's $3.29\times$). Under the unedited-pair null ECoT-bridge ranks fifth, where $\rho \approx 0.333$; across the five forms below, ρ ranges from 0.071 to 0.476.

Calibration form and near-zero translations. Two re-analyses of the same per-sample *direction_flip* records, with no new inference, all actions on each checkpoint's own de-quantization grid. *Form.* Table 34 ranks the eight three-seed configurations five ways – raw $\mathcal{F}_{\text{dir}}$, and the ratio and the difference of $\mathcal{F}_{\text{dir}}$ over the family ceiling and over the unedited-pair null – and correlates each with the $\mathcal{F}_{\text{mag}}$ rank. ECoT-bridge is seventh under four forms and fifth under one; it is never above fifth. *Norm.* No sample on any configuration has an undefined cosine (zero translation vector). Requiring both the original and the edited translation to have norm ≥ 0.05 (on the $[-1, 1]$ scale) removes none of ECoT-bridge's samples but 29–53% of the six CoT-trained LoRA variants' and 94% of `no-CoT`'s (Table 35); on the remaining samples the LoRA variants reverse almost always (0.983–1.000) while their family ceilings stay at 0.043–0.133. The collapsed action spaces therefore depress, rather than inflate, the LoRA variants' raw $\mathcal{F}_{\text{dir}}$, and ECoT-bridge's non-reversal is not a near-zero-vector artifact. At a 0.10 threshold the ECoT-lineage numbers move by at most 0.027; on the two DeepThinkVLA policies 8–10% of samples are then excluded and $\mathcal{F}_{\text{dir}}$ stays ≤ 0.011.

Table 34: Spearman ρ between the $\mathcal{F}_{\text{mag}}$ rank and five $\mathcal{F}_{\text{dir}}$ rankings of the eight three-seed configurations on *direction_flip*, and ECoT-bridge's rank (of 8) under each, on all samples and on samples with both translation norms ≥ 0.05.

	all samples		norm ≥ 0.05	
$\mathcal{F}_{\text{dir}}$ ranking	ρ	ECoT-bridge	ρ	ECoT-bridge
raw	0.476	7	−0.048	8
ratio over family ceiling	0.071	7	0.333	7
difference over family ceiling	0.429	7	0.333	7
ratio over unedited-pair null	0.333	5	0.310	7
difference over unedited-pair null	0.476	7	0.071	8

Table 35: *direction_flip* samples excluded when both translation norms must be ≥ 0.05, and $\mathcal{F}_{\text{dir}}$ and the family ceiling on the samples that remain.

Model	excluded	n kept	$\mathcal{F}_{\text{dir}}$	family ceiling
`ECoT-bridge`	0.000	299	0.117	0.047
`r=64`	0.291	212	0.995	0.058
`r=16`	0.411	176	0.983	0.043
`data-50A`	0.472	158	0.994	0.133
`r=8`	0.378	186	1.000	0.074
`r=32`	0.468	159	0.994	0.073
`data-50B`	0.532	140	0.986	0.092
`no-CoT`	0.940	18	0.222	0.333
`DT base`	0.000	100	0.000	0.000
`DT SFT`	0.000	100	0.010	0.050
`DT RL`	0.010	99	0.010	0.021

Figure 8 plots $\mathcal{F}_{\text{dir}}$ against action-space richness.

B.6.3 Closed-loop rollouts of the two competent LIBERO policies

DeepThinkVLA-SFT and -RL are rolled out on all four LIBERO suites: 10 tasks per suite and 10 episodes per task, 400 episodes per arm per model. Every arm of an episode shares the same environment and settled initial state, so the arms are paired. The CoT is regenerated at every re-plan, and the arm's edit is applied to it, where its target is present, before the action is decoded. Clean success is 0.965 (SFT) and 0.960 (RL), with 0.93–0.99 per suite. ΔSR is the arm's success minus the clean arm's on the same episodes, with a cluster bootstrap over (suite, task) (10,000 draws) and

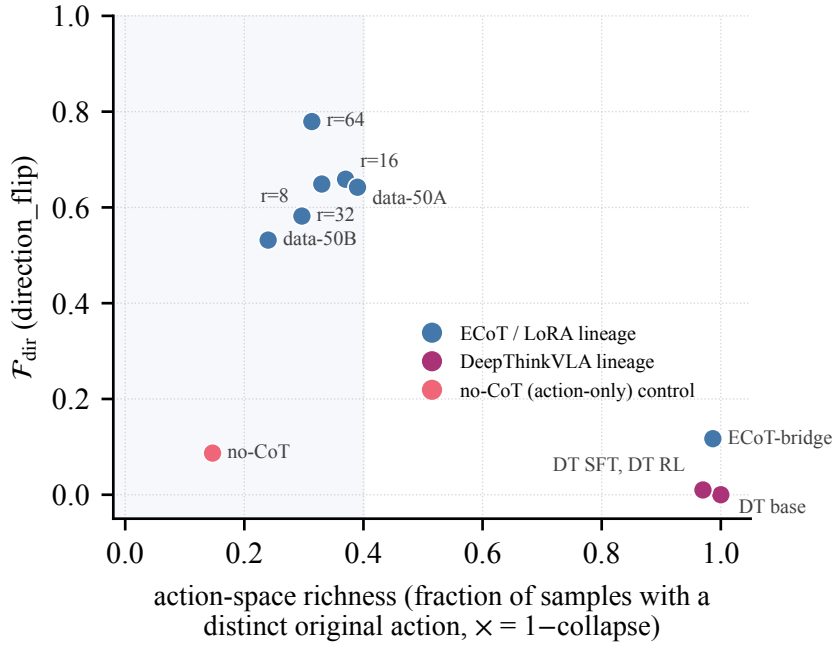

Figure 8: $\mathcal{F}_{\text{dir}}$ against action-space richness (x, higher = less collapsed), the confound noted in §B.6.2; action-space richness is not a competence measurement. Shaded band: the six collapsed-action-space LoRA/data variants, 6 of which clear the null (`no-CoT`, also collapsed, does not); unshaded: the four rich-action-space configurations, of which only `ECoT-bridge` clears (all three DeepThinkVLA rows do not). Action-space richness alone does not cleanly separate clearing from non-clearing configurations, so we report the diagnostic beside every score rather than reading $\mathcal{F}_{\text{dir}}$'s absolute value as comparable across the cohort.

Table 36: Closed-loop ΔSR against the clean arm (400 paired episodes per arm, cluster-bootstrap 95% interval) and the on-policy open-loop $\mathcal{F}$ at the same re-plans. *cov.*: edit coverage (RL; SFT within 0.01). †: applied at every re-plan by construction (the empty CoT is decoded with no reasoning, the identity arm re-tokenizes the generated CoT).

		DT-SFT			DT-RL		
Arm	cov.	ΔSR	95%	$\mathcal{F}$	ΔSR	95%	$\mathcal{F}$
cross_task_cot	1.00	−0.158	[−0.253, −0.075]	—	−0.158	[−0.250, −0.075]	—
negation	0.94	+0.005	[−0.013, +0.025]	0.544	+0.005	[−0.015, +0.025]	0.542
gripper_flip	0.38	−0.003	[−0.025, +0.020]	0.341	+0.005	[−0.018, +0.028]	0.324
direction_flip	0.08	+0.000	[−0.022, +0.022]	0.254	−0.003	[−0.020, +0.015]	0.233
object_swap	0.91	+0.005	[−0.013, +0.022]	0.384	+0.022	[+0.003, +0.045]	0.402
object_swap_all	0.91	−0.007	[−0.030, +0.015]	0.567	+0.003	[−0.018, +0.022]	0.579
paraphrase_null	0.93	−0.015	[−0.040, +0.010]	0.296	−0.003	[−0.025, +0.020]	0.300
paraphrase_1tok	0.93	−0.013	[−0.035, +0.007]	0.286	+0.003	[−0.020, +0.030]	0.287
syntactic_scramble	0.94	+0.003	[−0.020, +0.028]	0.324	−0.007	[−0.028, +0.013]	0.333
empty_cot	1.00†	+0.003	[−0.020, +0.025]	—	+0.000	[−0.015, +0.018]	—
identity_retok	1.00†	−0.018	[−0.037, +0.003]	—	+0.005	[−0.010, +0.020]	—

an exact McNemar test. Trajectory divergence is the mean end-effector distance from the clean arm's path. The open-loop $\mathcal{F}$ is scored on the clean arm's own re-plans, so it is on-policy. Two arms have no counterpart in the LIBERO protocol: *object_swap_all* swaps the object name everywhere in the CoT, and *paraphrase_1tok* is a one-token synonym swap. Edit coverage is the fraction of re-plans at which the edit applied; it is low for *direction_flip* and *gripper_flip*, which need their tag to be present.

Reading. Only the wholesale cross-task CoT changes success: b=72, c=9 discordant episodes on RL, McNemar $p = 2.5\times10^{-13}$; 5.8×10^{-13} on SFT. Every content edit, every null and the empty

CoT stay within 0.022 of clean. The one interval that excludes zero is RL *object_swap*, where success goes up ($+0.022$). The open-loop comparison on the same re-plans gives the opposite impression. The semantic mean exceeds all three floors on both models, by $+0.083$ to $+0.132$ with intervals excluding zero, whereas the closed-loop difference $\Delta\text{SR}_{\text{sem}} - \Delta\text{SR}_{\text{floor}}$ includes zero for all three. Across the eight text families, $\mathcal{F}$ and $\overline{\text{TV}}$ correlate with trajectory divergence at Spearman 0.98 and 0.93 (RL) and at 0.98 and 0.98 (SFT). Against the loss of success they correlate at -0.42 and -0.56 (RL) and at -0.37 and -0.37 (SFT).

These policies act on what the reasoning names, well enough that a wrong task's reasoning costs 16 points. Within-task edits to its content move the action at the steps they apply to, and success is unchanged, which is consistent with closed-loop correction but does not identify it. The open-loop score therefore measures trajectory perturbation, not outcome dependence. This is two policies from one lineage on one benchmark, and the direction and gripper edits reach few re-plans.