NOT ALL CONCEPTS ARE CAUSAL: SPARSE AUTOENCODERS REVEAL A MOTOR LAYER IN VLAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies hide which internal features merely describe a task and which produce the action. We introduce SAE-VLA, which turns a frozen VLA's activations into a concept lookup table for reading out and steering its actions. A sparse autoencoder provides the rows, a vision-language model names them, and a blinded verifier checks every name on held-out episodes. A signed intervention records whether a row moves the action its name implies, at a dose certified by a closed-loop gate. Across five policies on two backbones, where the table is built decides whether it works. In $\pi_{0.5}$, language features receive plausible motor names that verify half as often and barely act. The action expert instead holds a *motor layer* of grasp, release, hand-height, and wrist-tilt features. On the $\pi_{0.5}$ breakfast policy, verified names identify the moving arm on 98.6% of held-out single-arm frames and contain the current subtask on 97%, by lookup alone. A name indicates where to steer but not whether, so the table's causal column defines what is steerable. On the $\pi_{0.5}$ LIBERO policy, a gated verified row makes the arm descend earlier with success unchanged. Gated handles replicate with a second dictionary and on the OpenVLA-OFT LIBERO policy. Steering by language lets the frozen $\pi_{0.5}$ breakfast policy recover bread placements beyond its demonstrations, raising task success from 17.5% to 87.5% without new data or retraining. Not every named concept is causal; the table records which ones are.

1 INTRODUCTION

Vision-language-action (VLA) models are becoming the standard recipe for generalist robot policies (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024; Physical Intelligence et al., 2025). Yet they hide the computation between an instruction and a motor command, so it is unclear which

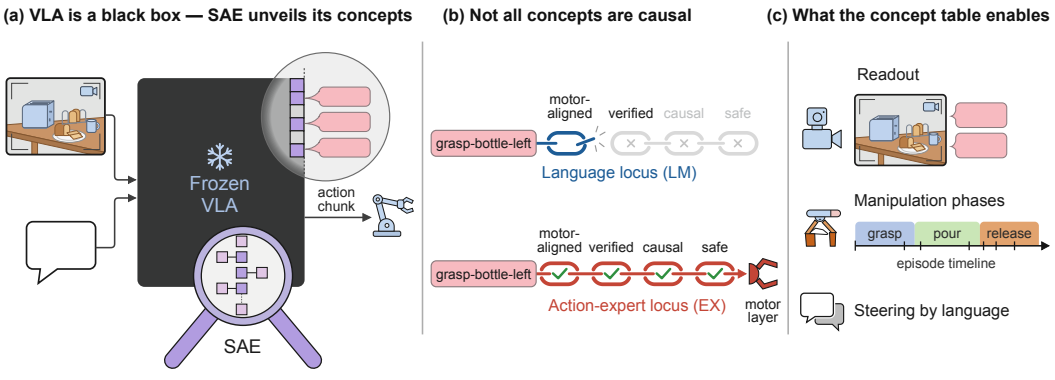

Figure 1: Not all concepts in a VLA are causal. (a) A sparse autoencoder unveils named concepts in a frozen VLA. (b) Language concepts verify half as often as action-expert ones and barely act. (c) The concept table reads out the policy, marks its phases, and steers it by language.

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.

internal features merely describe a task and which produce the action. A feature that correlates with an object or a subtask tells us what can be *decoded*, not whether the policy *uses* it to act.

The opacity has a price: retraining is the only way to fix a deployed policy. Building a VLA takes pre-training on over $10{,}000$ hours of robot data (Black et al., 2024) or $21{,}500$ A100-hours (Kim et al., 2024). Adapting it to a task takes 5 to over 100 hours of teleoperated demonstrations (Black et al., 2024), plus 10–15 GPU-hours even with low-rank adaptation (Kim et al., 2024); the 350 hours of DROID took 50 collectors 12 months (Khazatsky et al., 2024). The deployed policy is frozen, and its instruction names a subtask, not the height of a grasp. When it fails, the remedy is to collect demonstrations and fine-tune the action expert again, with no guarantee that the new data address the failure and no diagnosis when they do not. Whether demonstrations help depends on their quality and composition, and some hurt (Belkhale et al., 2023; Agia et al., 2025). Long-horizon tasks compound the cost. Our $\pi_{0.5}$ breakfast policy, fine-tuned on $1{,}850$ demonstrations, chains four subtasks: load bread from a rack into a toaster, press the lever, pour a drink, and place the toast on a plate. A systematic miss in any one fails the task, and because each subtask starts where the last one ended, fixing one step by data means re-collecting full sequences.

We ask instead whether the frozen policy already contains the knobs (Fig. 1). SAE-VLA turns its hidden activations into a *concept lookup table* (Fig. 2) whose rows are sparse autoencoder (SAE) features. Each row carries a name, a check that the name holds on episodes the namer never saw, a check that amplifying the row moves the action its name describes, and a dose the policy tolerates in closed loop. At inference, the active rows describe what the policy is doing, and a row chosen by name becomes a steering handle. A vision-language model (VLM) maps an operator's sentence onto rows to steer, so that "grasp the bread a bit lower" changes the action without touching the weights, the data, or the instruction.

The table works only if its features carry a *concept property chain*: they track a motor variable, carry a name that holds on new episodes, move the action the name implies, and do so at a tolerable dose. The links do not imply one another, so not all concepts are causal. A feature may fire on "pour the drink" because it echoes the prompt, and a feature that tracks hand height may be named after the pouring it accompanies. We hypothesize that the *representation locus* decides where the chain holds. Language residuals observe the instruction and should hold task descriptors, whereas action-expert residuals parameterize the motor plan and should concentrate features that are motor, nameable, and causal; we call this population a *motor layer*.

We trace the chain through the language and action-expert layers of five policies: the $\pi_{0.5}$ breakfast, folding, and flipping policies on a bimanual robot, and the $\pi_{0.5}$ and OpenVLA-OFT LIBERO policies. Language features leave it at the first links: at every depth few track motion, their motor-sounding names verify half as often, and even verified ones barely act. Action-expert features carry it to control: 77–97% of the most aligned rows steer their channel selectively on the $\pi_{0.5}$ folding and flipping policies, and on the $\pi_{0.5}$ LIBERO policy a gated row shifts its motor event in closed loop without costing success. Lookup in the complete table reads out the moving arm and subtask almost perfectly. Steering by language lets the frozen $\pi_{0.5}$ breakfast policy complete a sideways bread placement it misses, raising success from 5% to 95% without new data or training.

Contributions.

- **A concept lookup table for frozen VLAs** that names, verifies, and causally tests features without concept supervision, and serves readout and steering by name or by language.
- **The concept property chain**, traced across loci, depths, tasks, benchmarks, and two architectures, which places the table at the motor layer nearest to action generation.
- **Control without retraining**: readout matches held-out ground truth, the causal column rather than the name identifies handles, gated handles replicate on a second seed and backbone, and language requests recover real-robot failures of the frozen $\pi_{0.5}$ breakfast policy.

2 Related Work

Interpreting and steering VLA policies. Most work makes VLAs inspectable by inserting an explicit intermediate vocabulary (affordances, grounded masks, action chains of thought, or primitive skills) or by probing failure modes (Huang et al., 2024; Ma et al., 2025; Huang et al., 2025; Zhong

et al., 2026; Gu et al., 2026b; Lian et al., 2026; Xu et al., 2025). Concept bottlenecks fix the concepts before training (Koh et al., 2020; Tan et al., 2024), and latent-action models learn new action codes from video (Ye et al., 2025; Lin et al., 2026; Dai et al., 2026; Lee et al., 2026; Cheang et al., 2024; Gu et al., 2026a; Huang et al., 2026). Closer to us, mechanistic work steers frozen VLAs along feed-forward neurons projected onto the token vocabulary (Häon et al., 2025) and along SAE features, separating general primitives from memorized episodes (Swann et al., 2026) or ranking features by kinematic events (Jin et al., 2026). We differ in asking *where* features become causal, tracing the property chain across loci and depths against matched dense controls; in separating a verified name from a causal effect; and in using the table to fix a deployed policy from language.

Sparse autoencoders and their critics. SAEs decompose dense activations into sparse features (Makhzani & Frey, 2014; Cunningham et al., 2024; Gao et al., 2024; Rajamanoharan et al., 2024; Templeton et al., 2024) that can be described from their activating examples (Bills et al., 2023) and used as steering directions (Zhang et al., 2026). On language models, SAE features do not beat simple baselines for probing or steering (Kantamneni et al., 2025; Wu et al., 2025b), automated interpretability scores fail to separate trained from random transformers (Heap et al., 2026), and SAE concepts can be fragile (Li et al., 2026). We take these critiques as design constraints. Every unit-level claim is scored against raw neurons, principal components, and random directions, and names are grounded in motor state and verified on held-out episodes. A feature is called causal only if intervention moves the action selectively, sign-symmetrically, and in closed loop.

Correcting deployed robot policies. Language corrections steer hierarchical policies trained on low-level language motions and are then distilled by fine-tuning (Belkhale et al., 2024; Shi et al., 2024). Inference-time methods bias a generative policy's samples toward user input (Wang et al., 2024) or let a VLM choose among sampled plans (Wu et al., 2025a). SAE-VLA instead edits the frozen policy's activations along named directions with measured effects, and it needs no language-motion training, no sampling of alternatives, and no new data.

3 SAE-VLA: From Concepts to Control

SAE-VLA builds a concept table in four steps (feature discovery, naming, held-out verification, and causal testing) and uses it by lookup (Fig. 2).

3.1 Representation loci and the concept property chain

Let a frozen policy π_θ map observations o_t and an instruction ℓ to an action chunk a_t. We read, but never update, a residual $h_t^{(r,l)}$ at layer l and locus r. In $\pi_{0.5}$, the **language locus (LM)** is the residual of the last valid language token. The **action-expert locus (EX)** is the residual of the action tokens at the final denoising step, averaged over the horizon, immediately upstream of motor prediction. A feature is useful only if it is (P1) *faithful*, reconstructed without changing behavior; (P2) *motor-aligned*, its activation tracking a motor variable or event (association > 0.3); (P3) *verified*, its name predicting where it fires on held-out episodes; (P4) *causal*, its amplification moving the channel it predicts more than the others; and (P5) *safe*, shifting its motor event at a closed-loop-gated dose without costing success. Readout needs P1–P3, steering needs P1, P4, and P5, and steering by name needs all five. The locus hypothesis predicts that language features leave this chain early and that action-expert features carry it to P5.

3.2 Sparse feature discovery

For each locus and layer we train a top-k SAE on activation traces,

$$u_t = \mathrm{ReLU}(W_e(h_t - b_d) + b_e)\,, \qquad z_t = \mathrm{TopK}_k(u_t), \qquad \hat{h}_t = W_d z_t + b_d, \tag{1}$$

minimizing reconstruction error with the auxiliary dead-feature loss (Gao et al., 2024) and normalizing decoder columns after each update. Feature j is the pair of a residual direction $v_j = W_d[:, j]$ and its activation $z_{t,j}$; a *concept* is a held-out interpretation assigned to it, and a *steering handle* is one that passes the tests of Sec. 3.4. Dictionaries receive no concept labels and are trained on activations disjoint from all evaluation, calibration, and interpretation splits (Appendix A).

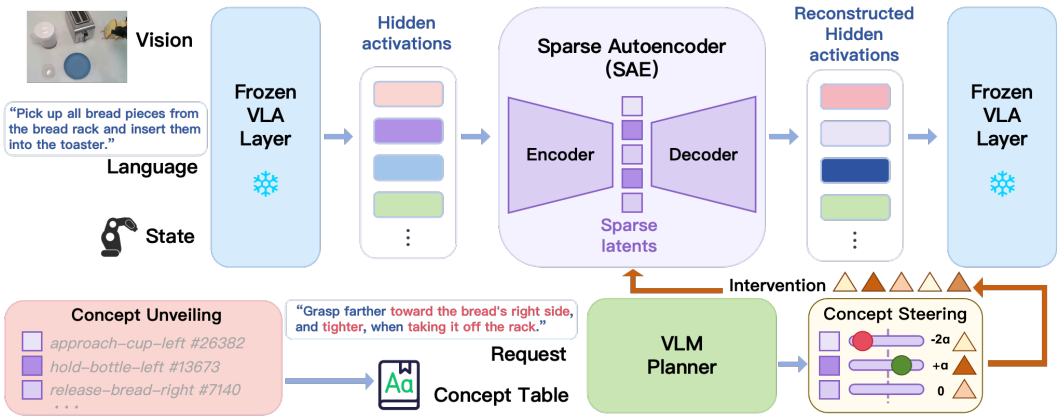

Figure 2: SAE-VLA. A sparse autoencoder between two frozen VLA layers exposes sparse latents (top). Concept unveiling names and verifies each latent, forming the concept table; a VLM planner maps an operator's request onto concepts, which concept steering amplifies or suppresses (bottom).

3.3 Evidence-grounded naming and held-out verification

For each active feature, we build an *evidence packet* from the training split: the K highest-activating windows, K matched low-activation windows from the same subtasks, the instruction, and synchronized proprioceptive and event traces. A vision-language interpreter returns a verb–object–modifier name with a one-sentence definition. A blinded verifier that sees the name but not the feature must pick the high-activation window out of held-out pairs (forced choice) and judge whether each held-out top window matches the description (top-K precision). A human rater answers the same items. Selectivity statistics use average-rank tie correction and episode splits, because sparse codes tie at zero on most frames (Appendix C).

3.4 Causal validation through intervention

To intervene on feature j we edit its sparse activation and decode,

$$\tilde{h}_t(\alpha, j) = W_d(z_t + \alpha\sigma_j e_j) + b_d = \hat{h}_t + \alpha\sigma_j v_j, \tag{2}$$

where σ_j is the standard deviation of the feature's training activations, so that positive and negative α amplify and suppress on a comparable scale. Effects are measured against *SAE-in-loop* ($\alpha = 0$), so reconstruction error cannot masquerade as intervention. For a target action group T, we report target and collateral effects and their ratio,

$$E_T = \tfrac{1}{|T|}\textstyle\sum_{i\in T}|\Delta a_i|, \qquad E_C = \tfrac{1}{d_a-|T|}\textstyle\sum_{i\notin T}|\Delta a_i|, \qquad R_{\text{spec}} = E_T/(E_C+\varepsilon). \tag{3}$$

A feature passes the *open-loop screen* if its dose response is monotone and sign-symmetric and $R_{\text{spec}} > 1$ on its target channel. The screen runs on predeclared unit sets, so dictionary features, raw residual dimensions, principal components, and random directions are scored identically. A *closed-loop gate* then fixes the dose α_{cl}, the largest grid dose whose paired success stays within a predeclared margin of SAE-in-loop on held-out rollouts. Features that pass the screen and change their motor event at α_{cl} are steering handles; rules and the freezing protocol are in Appendix A.

3.5 Using the table at inference

For every active feature, the table stores the name and verifier scores, the channel the name implies, the screen result and measured sign on that channel, and the gated dose. A name-to-channel map fixed before any screen sends grasp, hold, and release verbs to the named arm's gripper, tilt and pour verbs to its wrist, and motion verbs to its six arm joints; task-level names serve readout only. The distinct names form the table's *vocabulary*, and the SAE encodes the live residual at every policy call. *Readout* lists the verified names of the active features ranked by $z_{t,j}/\sigma_j$. *Steering by name* applies Eq. (2) at the gated dose to a verified feature whose screen passed on its named channel.

Steering by language turns a request into signed rows. A VLM planner receives the request, the current subtask, the top-camera frame, and the vocabulary with each row's verification, measured sign, and specificity, and proposes names to amplify or suppress (prompt in Appendix K). The operator ticks names and sets a strength λ, and every row $j \in S$ listed under a ticked name is applied with that name's sign $s_j \in \{\pm 1\}$ and its table dose α_j,

$$\tilde{h}_t = \hat{h}_t + \lambda \sum_{j \in S} s_j \, \alpha_j \sigma_j v_j. \tag{4}$$

The request never enters the VLA's prompt, the planner runs once per request, and no use consults the outcome.

4 Experimental Setup

Policies and domains. Fine-tuning $\pi_{0.5}$ (Physical Intelligence et al., 2025) on an Agilex Cobot Magic bimanual robot for breakfast preparation, carton folding, and package flipping (Appendix B) gives the $\pi_{0.5}$ breakfast (primary), folding, and flipping policies; fine-tuning it on the four LIBERO suites (Liu et al., 2024) gives the $\pi_{0.5}$ LIBERO policy. The single-stream OpenVLA-OFT (Kim et al., 2025) fine-tuned on LIBERO is the OpenVLA-OFT LIBERO policy. LIBERO-plus (Fei et al., 2025) and LIBERO-X (Wang et al., 2026) serve as distribution shifts. Dictionaries are trained at both loci at layers 5, 8, 11, and 14 of the $\pi_{0.5}$ breakfast policy's two 18-layer stacks (default probes: LM 11, EX 8), at the same expert layers of the $\pi_{0.5}$ LIBERO policy, and at layers 8–31 of the OpenVLA-OFT LIBERO policy, always on episodes disjoint from naming, dose calibration, and testing.

Recovering failures on the robot. We move two breakfast subtasks beyond the demonstrations (Fig. 6). In *place bread*, a slice stands in the toaster slot either upright, as demonstrated, or sideways, rotated by 90°. In *load bread*, each of the two slices on the rack leans left (L) or right (R), and right/right is the demonstrated lean. The frozen $\pi_{0.5}$ breakfast policy and SAE-VLA steered by language run in randomized, paired order from identical placements, with 20 paired trials per placement. The steering configuration (request, ticked names, and strength) is fixed per perturbation before evaluation, and success follows Appendix K.

System labels and statistics. *Frozen VLA* is the unmodified policy, *SAE-in-loop* inserts the reconstruction with $\alpha = 0$, and *SAE-VLA* adds an intervention. Main-text numbers are point estimates, with $\pm$ giving a 95% interval where a claim rests on it; paired bootstrap intervals and exact McNemar tests are in the appendix. Every feature, sign, dose, and rule was frozen before the outcome it governs (Appendix A).

5 Results

5.1 The concept property chain across loci and layers

We follow the chain link by link over four depths (Fig. 3) and five policies (Table 1).

The sparse interface preserves the policy (P1). Expert-locus dictionaries reconstruct 0.93–0.999 of held-out variance. Inserting the reconstruction leaves success within noise of the frozen policy, by $-1.5 \pm 3.0\,pp$ over 400 paired episodes of the $\pi_{0.5}$ LIBERO policy (at every denoising step) and $-2.5 \pm 6.5\,pp$ over 80 of the $\pi_{0.5}$ breakfast policy (at the final step). A 128-dimensional PCA bottleneck costs $-1.0\,pp$, and the interface is neutral on the OpenVLA-OFT LIBERO policy ($+1.0\,pp$; Appendix Table 9). A language-locus configuration that reconstructs well still loses $6\,pp$ on LIBERO-10, so locus and k are frozen before any steering experiment (Appendix F).

Language features leave the chain at its first links (P2–P3). At every depth of the $\pi_{0.5}$ breakfast policy, the action expert reconstructs better (variance explained 0.95–0.999 against 0.63–0.85) and holds 10–45% motor-aligned active features against 0.6–7% at the language locus, 6–21× more at matched depth. The contrast holds in every $\pi_{0.5}$ policy for SAE features (19–56×) and, more weakly, for raw residual dimensions (1.9–13×), and it survives matched sparsity, density, capacity, and four seeds (Appendix G). Both loci receive motor-primitive names in similar proportions (Appendix Fig. 10). Yet in each of the five policies and within every verb family, expert names pass held-out verification more often: 57 of 200 against 25 of 200 pooled (Fisher $p = 1.1 \times 10^{-4}$; Fig. 4a). Most language-feature names sound motor but fail to predict where the feature fires.

Table 1: The chain at the language (LM) and action-expert (EX) loci. P2: % of active units; P3: verified names among 40 random features; P4: % of the 20 (LM) or 30 (EX) most aligned features that pass the screen. For OpenVLA-OFT, LM and EX are its prompt and action tokens at layer 16.

	P2 motor-aligned (%)				P3 verified		P4 steers (%)	
	SAE		dims		(of 40)		aligned	
Policy / data	LM	EX	LM	EX	LM	EX	LM	EX
$\pi_{0.5}$								
Agilex breakfast	1.1	34.5	21.5	49.3	5	12	35	63
Agilex folding	0.03	1.1	2.3	30.7	3	10	65	97
Agilex flipping	0.05	0.9	22.8	43.8	6	9	85	77
LIBERO	0.24	13.8	29.7	56.0	4	12	40	33
OpenVLA-OFT								
LIBERO	0.4	0.6	19.0	20.3	7	14	10	27

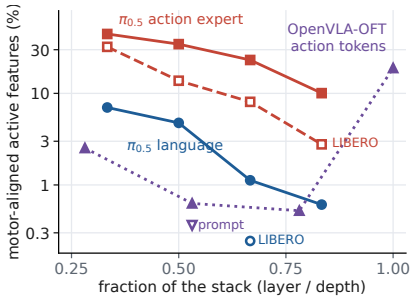

Figure 3: Motor alignment (P2) by depth; open: $\pi_{0.5}$ LIBERO policy and OpenVLA-OFT prompt tokens.

Action-expert features carry the chain to control (P4). On the $\pi_{0.5}$ breakfast policy, the expert dictionary concentrates motor structure. Its strongest features isolate single events (peak association 0.73 against 0.64 for the best neuron), and random expert features reach a top-K precision of 0.42 against 0.26 for neurons and 0.28 for principal components, an ordering a blinded human rater reproduces ($\kappa = 0.66$–0.86; Appendix Table 5). Concentration becomes control: on the $\pi_{0.5}$ folding and flipping policies, 77–97% of the most aligned expert features pass the open-loop screen (median R_{spec} 2.3–3.1; 77–100% for three re-seeded dictionaries each), whereas aligned raw dimensions, matched-energy random directions, and random active features stay near one (Appendix Table 14). Verified language concepts barely act: on the breakfast policy, the best-verified one, `release-object-right` (31063), moves its gripper eight times less than the expert row `grasp-bottle-left` (3659), without reliable sign reversal, and only one of its five passes on its named channel. The 20 most aligned language features move their channel by a median of 1.9×10^{-3}, against 3.0×10^{-2} for the 30 most aligned expert ones (Appendix D).

Where the chain bends. Naming and control come apart in both directions (Fig. 4b–f). The row `tilt-wrist-pour-left` (8176) fails verification yet tracks left-hand height and is a handle. Among the twelve cards of the $\pi_{0.5}$ folding and flipping policies, nine are handles but only three carry verified names. The causal link also depends on the task: on the single-instruction $\pi_{0.5}$ flipping policy, the most aligned language features steer the arm they track (85% pass). On the $\pi_{0.5}$ breakfast policy, aligned expert features beat random ones (63% against 47%) only in the reference dictionary. Where most aligned features target a saturated gripper (the two LIBERO policies), no unit type steers well (Appendices H and I). In OpenVLA-OFT, whose prompt positions read a proprioceptive token, depth matters more than position: action-token alignment rises to 19% at the output-proximal layer 31 (Fig. 3). The motor layer is thus the residual nearest to action generation, so the table is built at the action expert of $\pi_{0.5}$ and at layer 31 of OpenVLA-OFT.

5.2 Reading out the policy's action by lookup

We named and verified every expert feature of the $\pi_{0.5}$ breakfast policy active on held-out data. Of 513 features, 511 receive a name and 188 pass verification, and the vocabulary has 79 distinct names. Re-naming 116 features from a fresh evidence pool reproduces the verb and arm of 75% of their names. Every held-out frame has a verified row active, and verified rows carry a third of the expert activation. In Fig. 5, `grasp-bottle-left` (2371), `hold-bottle-pouring-left` (12411), and `release-bottle-left` (4357) peak in turn at the grasp, during the pour, and within 0.2 s of the release. Against held-out ground truth (Table 2, top), the arm named by the top verified row is the moving arm on 98.6% of single-arm frames (93% for the minority left arm; majority baseline 86.5%). The objects named by the three top rows contain the current subtask on 97.0% of frames, with 2.1 subtasks per set (52% for random sets). Every grasp and release event has a matching gripper name among the three top rows within 0.5 s, against 80% with names shuffled (exploratory, because the predeclared criterion saturates; Appendix E). Verified names beat all names on each measure, so verification is what makes the readout trustworthy. On the robot, the same lookup applied to the codes logged at every policy call recovers arm-motion phases, subtask identity, and gripper events in the synchronized hardware traces (Appendix K).

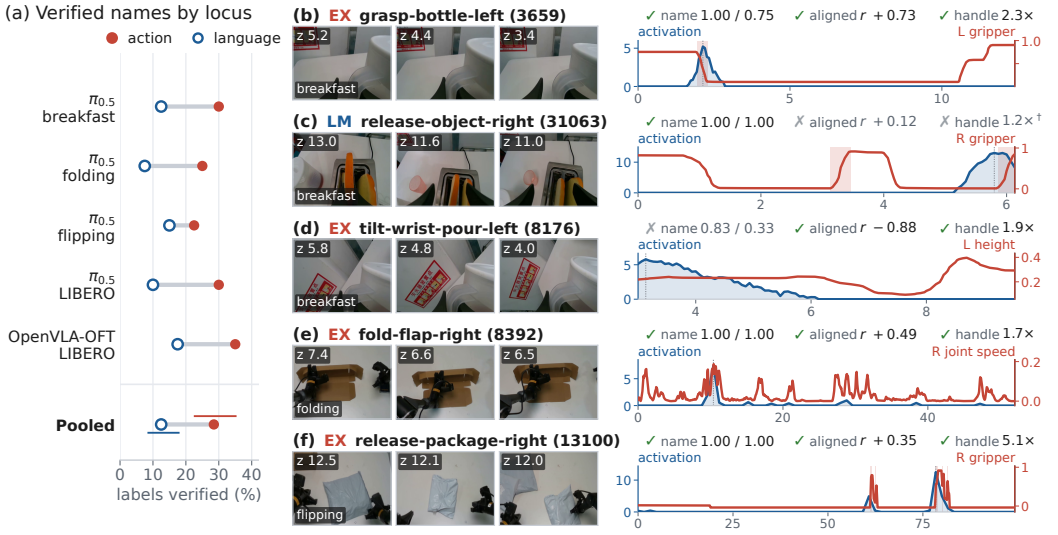

Figure 4: Naming and control come apart. (a) Names that pass held-out verification (forced choice ≥ 0.75, top-$K \geq 0.5$) among 40 random active SAE features per locus and policy (OpenVLA-OFT: layer 16); Wilson 95% intervals. (b–f) Concept cards: verified name (forced choice / top-K), motor alignment ($|r| > 0.3$), handle (R_{spec} at $\pm 16\sigma_j$; †no sign reversal), top three held-out frames, and activation (blue) against the motor variable (red) over the strongest episode. All twelve cards of the $\pi_{0.5}$ folding and flipping policies: Appendix D. Card (c) was fixed before its test.

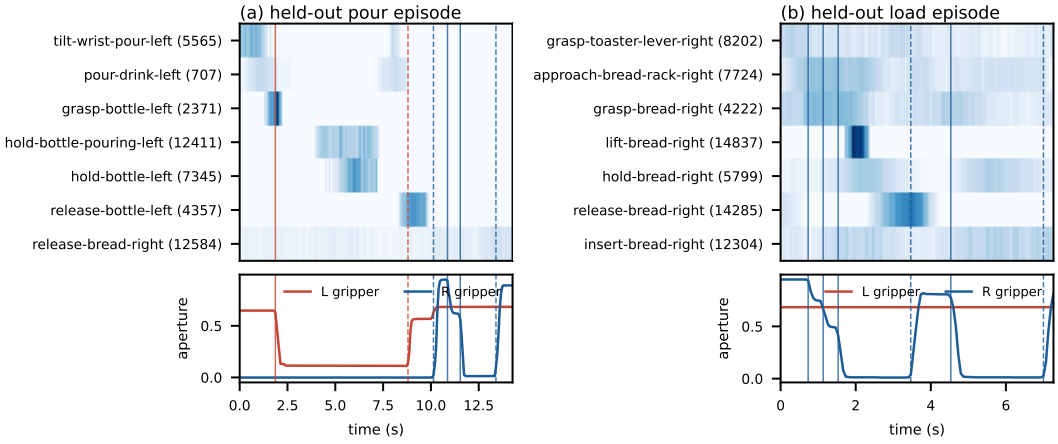

Figure 5: Readout by lookup on two held-out breakfast episodes. Top: activation z/σ_j of the verified rows that most often rank among the top three (one row per name). Bottom: gripper apertures; vertical lines mark grasp (solid) and release (dashed) events of the left (red) and right (blue) arm.

5.3 Steering by name: the causal column decides

A name tells where to steer; only the intervention tells whether. On the channel its name implies, 54% of verified rows pass the open-loop screen, against 23% when the arm in the name is swapped (Table 2, bottom). Arm- and wrist-motion names point to the right channel most often (71%, against 17% swapped) and gripper names less often (45%), and a gripper verb predicts the direction of the change in only 29% of rows. Verification leaves steerability unchanged (52% for rows whose names fail), because it certifies when a row fires, not what it does. Querying the table for the best-verified row of each verb and arm thus returns a handle in only 2 of 10 queries (Appendix E), whereas querying among the 98 verified rows that passed on their named channel returns one by construction. The causal column, not the name, defines the menu for steering by name.

Table 2: Lookup on the $\pi_{0.5}$ breakfast policy (513 active expert features; 511 named, 188 verified), in %. Top: readout on 4,771 held-out frames against size-matched baselines; the event row uses the exploratory top-3 criterion (Appendix E). Bottom: open-loop screen at $|\alpha|{=}16\sigma_j$ on 1,016 frames, scored on the channel each name implies (verified, unverified, and arm-swapped names).

Readout	Verified names	All names	Baseline
Top row names the moving arm	98.6	91.4	86.5 (majority)
Top-3 objects contain the subtask	97.0	93.9	52 (random set)
Gripper event among top-3 names	100	99.3	80 (shuffled)
Steering by name	**Verified**	**Unverified**	**Arm swapped**
Pass on the named channel	54	52	23
arm and wrist names	71	69	17
gripper names	45	40	26

Table 3: Closed-loop steering by name at the gated dose, paired against SAE-in-loop (LIBERO: $n{=}400$ at α_{cl}, 200 otherwise; robot: $n{=}12$ paired trials). Name: verified on held-out episodes; OL R_{spec}: open-loop specificity at $16\sigma_j$; target / collateral: change in the commanded channels ($\times 10^{-3}$); event: change in the named motor event (LIBERO: descend onset, steps); consistent: episodes moving as predicted. Both signs, all doses, and CIs: Appendix Table 23.

Row	Name	OL R_{spec}	Dose	Δ success (pp)	Target	Collateral	Event	Consistent (%)
$\pi_{0.5}$ LIBERO policy								
`approach-stove-top` (9168)	✓	1.5×	$+\alpha_{\text{cl}}$ ($8\sigma_j$)	−1.0	−13.5	+1.4	−1.3	74
			$+\alpha_{\text{ol}}$ ($32\sigma_j$)	−24.5	−65.5	−29.0	−5.2	87
Replications: $\pi_{0.5}$ LIBERO policy, second dictionary; OpenVLA-OFT LIBERO policy, layer 31								
`approach-bowl-open` (868)	×	1.1×	$+\alpha_{\text{cl}}$ ($16\sigma_j$)	−2.0	−12.3	−3.0	−1.0	68
`grasp-object` (7165)	×	10.4×	$+\alpha_{\text{cl}}$ ($4\sigma_j$)	−1.5	+17.3	−10.2	+4.6	88
$\pi_{0.5}$ breakfast policy (Agilex robot)								
`grasp-bottle-left` (3659)	✓	2.3×	$+\alpha_{\text{cl}}$	−8.3	+4.2	+1.8	−3.0	83
`tilt-wrist-pour-left` (8176)	×	1.9×	$+\alpha_{\text{cl}}$	−8.3	+5.0	+2.7	−3.5	75
`release-bottle-left` (8988)	×	1.1×	$+\alpha_{\text{cl}}$	0.0	+1.5	+1.4	−1.5	67
`hold-bottle-left` (10099), random control	×	0.8×	$+\alpha_{\text{cl}}$	−8.3	+0.8	+2.0	+0.2	50
`grasp-bottle-left` (7779), LM control	×	0.95×	$+\alpha_{\text{cl}}$	0.0	+0.1	+0.2	0.0	50

5.4 Closed-loop control at a gated dose

Open-loop specificity says which variable a row moves; only closed-loop rollouts say whether the policy tolerates it (P5). On the $\pi_{0.5}$ LIBERO policy, the frozen lift rule selects the verified row `approach-stove-top` (9168), and the gate sets $\alpha_{\text{cl}} = 8\sigma_j$ (Appendix Fig. 16). At that dose, success is unchanged over 400 paired episodes ($-1.0 \pm 3.2\,pp$) while the arm approaches as named (Table 3). The height command moves as predicted on 74% of episodes with the lateral command unchanged, and the end effector descends 1.3 steps and closes the gripper 1.5 steps earlier. The effect is graded and sign-symmetric, from $+2.8$ steps at $-16\sigma_j$ to -2.5 at $+16\sigma_j$ (Appendix Table 23). The gate is necessary: at the open-loop dose $\alpha_{\text{ol}} = 32\sigma_j$ the row is 1.6× specific, yet held for entire rollouts it costs $24.5\,pp$ of success because the lateral command moves twenty times more. Gated handles survive a second dictionary (descend onset -1.0 and $+1.1$ steps at $\pm 16\sigma_j$) and transfer to the OpenVLA-OFT LIBERO policy, where a layer-31 height row gated at $4\sigma_j$ delays the descent by 4.6 steps. Both rows steer although their names, `approach-bowl-open` (868) and `grasp-object` (7165), fail verification. On the robot, both signs reverse the target and the event ($n{=}12$ paired trials; Fig. 6d). At $+\alpha_{\text{cl}}$, the gripper, arm, and wrist rows move their targets by $+4.2$, $+5.0$, and $+1.5 \times 10^{-3}$ (83%, 75%, and 67% consistent), and one sign of each costs one success ($-8.3\,pp$). The language and random expert controls stay at chance (50% consistent).

5.5 Steering by language recovers failures of the frozen policy

Finally, we test whether the handles fix failures without retraining (Fig. 6; Appendix Table 22). Over 20 paired trials, the frozen $\pi_{0.5}$ breakfast policy completes the demonstrated upright *place bread* on 80% of trials but a sideways slice on only 5%. The request "grasp a bit lower and deeper when taking the bread out of the toaster" amplifies the rows under the three names ticked in Fig. 6a and raises both placements to 95%. For loading, no row is named for a sideways offset, so the request

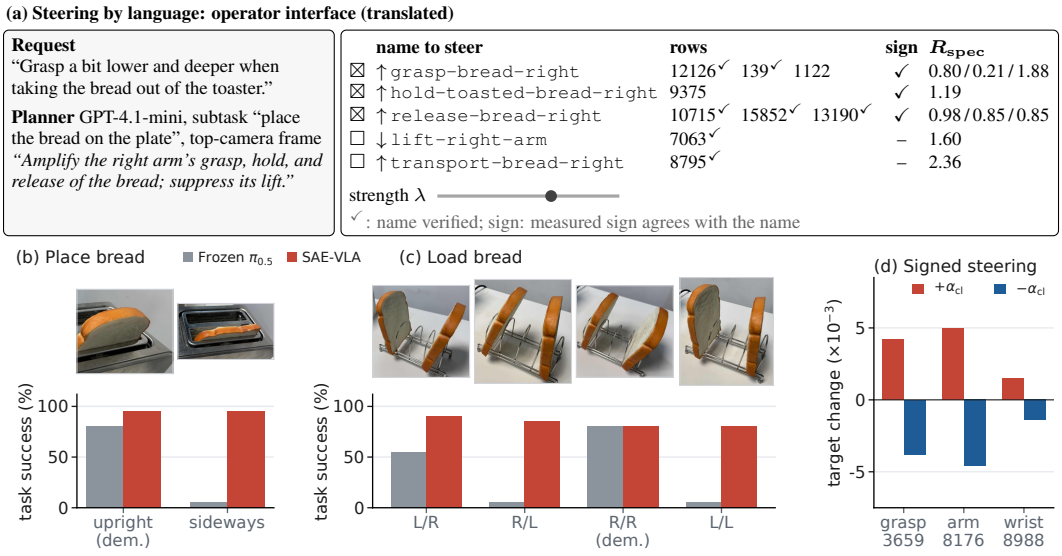

Figure 6: Steering on the robot. (a) The place-bread request and the names the planner proposes from the vocabulary (✓: name verified; –: arm verb without a sign). (b) Place bread and (c) load bread: initial placements and task success over 20 paired trials (dem.: demonstrated; Appendix Table 22). (d) Signed target change of the three steered rows at $\pm\alpha_{\text{cl}}$ (Appendix Table 23).

"grasp toward the bread's right side, and tighter, off the rack" amplifies three verified steerable rows: `approach-bread-rack-right` (7724) (2.2×), `grasp-bread-right` (10565) (2.6×), and `hold-bread-right` (12396) (2.2×). This leaves the demonstrated lean at 80% and raises the other three from 5–55% to 80–90%. Pooled over the four placements beyond the demonstrations, task success rises from 17.5% to 87.5% (n=80). Both requests use λ=4 ($32\sigma_j$ per row), with the task instruction and the weights unchanged.

6 Discussion and Conclusion

Not all concepts are causal. Held-out evidence certifies when a row fires, which readout needs, and only intervention certifies what it does, which steering needs; the two come apart in both directions. Locus explains where the chain holds: language residuals observe the prompt, so task identity is decodable there but rarely acts, whereas expert residuals parameterize the motor plan at every depth. The contrast exists without any dictionary; the SAE factors it into rows that can be named, looked up, and tested one by one. On the $\pi_{0.5}$ folding and flipping policies they steer more selectively than any dense direction, unlike SAE features on language models, which do not beat simple baselines (Kantamneni et al., 2025; Wu et al., 2025b). Steering by language builds on these records: instead of a prompt, the planner picks rows whose verification and measured effects it can see.

Scope. The table was completed only for the $\pi_{0.5}$ breakfast policy and checked by one blinded human rater. Where the gripper is a saturated switch, dense directions steer as well as the dictionary. Constant biases do not improve success under distribution shift (Appendix J), so deciding when to engage a handle calls for a state-dependent controller. Steering by language also needs an operator and a vocabulary that covers the correction. Within this scope, the table reads the action out and lets an operator correct the policy with a sentence.

References

Christopher Agia, Rohan Sinha, Jingyun Yang, Rika Antonova, Marco Pavone, Haruki Nishimura, Masha Itkina, and Jeannette Bohg. CUPID: Curating data your robot loves with influence functions. In *Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, 2025.

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Data quality in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quan Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. RT-H: Action hierarchies using language. In *Robotics: Science and Systems (RSS)*, 2024.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeffrey Wu, and William Saunders. Language models can explain neurons in language models. In *OpenAI Technical Report*, 2023.

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

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Weisheng Dai, Kai Lan, Jianyi Zhou, Bo Zhao, Xiu Su, Junwen Tong, Weili Guan, and Shuo Yang. ConLA: Contrastive latent action learning from human videos for robotic manipulation. *arXiv preprint arXiv:2602.00557*, 2026.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Songen Gu, Yunuo Cai, Tianyu Wang, Simo Wu, and Yanwei Fu. Say, dream, and act: Learning video world models for instruction-driven robot manipulation. *arXiv preprint arXiv:2602.10717*, 2026a.

Zhihao Gu, Ming Yang, Difan Zou, and Dong Xu. SDP: Learning diffusion policy from primitive skills for robot manipulation. *arXiv preprint arXiv:2601.01948*, 2026b.

Bear Häon, Kaylene Caswell Stocking, Ian Chuang, and Claire Tomlin. Mechanistic interpretability for steering vision-language-action models. In *Conference on Robot Learning (CoRL)*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2743–2762, 2025.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2501.17727.

Haifeng Huang, Xinyi Chen, Hao Li, Xiaoshen Han, Yilun Chen, Tai Wang, Zehan Wang, Jiangmiao Pang, and Zhou Zhao. RoboGround: Robotic manipulation with grounded vision-language priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. CoPa: General robotic manipulation through spatial constraints of parts with vision-language models. *arXiv preprint arXiv:2403.08248*, 2024.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. PointWorld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Xinchen Jin, Aditya Chatterjee, Pranav Kumar, and Rohan Paleja. Event-grounded sparse autoencoders for vision-language-action policies. *arXiv preprint arXiv:2605.17204*, 2026.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049, 2025. arXiv:2502.16681.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

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

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020.

Jung Min Lee, Dohyeok Lee, Seokhun Ju, Taehyun Cho, Jin Woo Koo, Li Zhao, Sangwoo Hong, and Jungwoo Lee. MVP-LAM: Learning action-centric latent action via cross-viewpoint reconstruction. *arXiv preprint arXiv:2602.03668*, 2026.

Aaron J. Li, Suraj Srinivas, Usha Bhalla, and Himabindu Lakkaraju. Evaluating adversarial robustness of concept representations in sparse autoencoders. *arXiv preprint arXiv:2505.16004*, 2026.

Shijie Lian, Bin Yu, Xiaopeng Lin, Laurence T. Yang, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Cong Huang, and Kai Chen. LangForce: Bayesian decomposition of vision language action models via latent action queries. *arXiv preprint arXiv:2601.15197*, 2026.

Shan Lin, Lirui Wang, et al. CLAP: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv preprint arXiv:2601.04061*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Teli Ma, Jia Zheng, Zifan Wang, Ziyao Gao, Jiaming Zhou, and Junwei Liang. GLOVER++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv preprint arXiv:2505.11865*, 2025.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2014.

Physical Intelligence, Kevin Black, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z. Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. In *Robotics: Science and Systems (RSS)*, 2024.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy III, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in VLA models. *arXiv preprint arXiv:2603.19183*, 2026.

Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision (ECCV)*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jermyn, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Guodong Wang, Chenkai Zhang, Qingjie Liu, Jinjin Zhang, Jiancheng Cai, Junjie Liu, and Xinmin Liu. LIBERO-X: Robustness litmus for vision-language-action models. *arXiv preprint arXiv:2602.06556*, 2026.

Yanwei Wang, Lirui Wang, Yilun Du, Balakumar Sundaralingam, Xuning Yang, Yu-Wei Chao, Claudia Pérez-D'Arpino, Dieter Fox, and Julie Shah. Inference-time policy steering through human interactions. *arXiv preprint arXiv:2411.16627*, 2024.

Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: VLM-in-the-loop policy steering via latent alignment. In *Robotics: Science and Systems (RSS)*, 2025a.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025b. arXiv:2501.17148.

Siyu Xu, Zijian Wang, Yunke Wang, Chenghao Xia, Tao Huang, and Chang Xu. Affordance field intervention: Enabling VLAs to escape memory traps in robotic manipulation. *arXiv preprint arXiv:2512.07472*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Ruikang Zhang, Shuo Wang, and Qi Su. Mechanistic knobs in LLMs: Retrieving and steering high-order semantic features via sparse autoencoders. *arXiv preprint arXiv:2601.02978*, 2026.

Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Maoqing Yao, Si Liu, and Guanghui Ren. ACoT-VLA: Action chain-of-thought for vision-language-action models. *arXiv preprint arXiv:2601.11404*, 2026.

AI USE STATEMENT

Generative AI tools were used for code-authoring assistance and for manuscript drafting and editing. The authors verified all quantitative claims against experiment logs.

ETHICS STATEMENT

Agilex experiments use a dual-arm manipulator, room-temperature colored water, and an unpowered toaster body. An e-stop and a second spotter are present, and every collision is logged as a safety event and a failed trial. No human-subject data are collected beyond blinded annotation of robot videos by a member of the research team.

REPRODUCIBILITY STATEMENT

We will release SAE configurations, activation-harvest scripts, interpreter and verifier prompts, the tie-corrected evaluation, intervention hooks, matched-control code, the human-annotation package, the predeclared rule files with their timestamps, and anonymized robot logs. Released feature indices refer only to the accompanying dictionaries.

A Implementation and Data Splits

The default LM probe is layer 11 with $d = 2048$ and dictionary size $m = 32768$; the default EX probe is layer 8 with $d = 1024$ and $m = 16384$. We train for 30k Adam steps with batch size 2048, learning rate 3×10^{-4}, and $k \in \{64, 128, 256\}$ at both loci (EX k=256 enters only the reconstruction sweep of Table 10). The auxiliary dead-feature loss follows Gao et al. (2024). Agilex activations are harvested at stride 2; LIBERO activations at stride 1.

Seeds, data, and splits. All dictionaries use training seed 0 unless stated; the cross-seed analysis adds EX L=8, k=128 seeds 1–3 with identical hyperparameters. Closed-loop evaluation uses environment seed 0 and fixed per-frame flow-matching noise (seeded by episode and step) for all paired conditions. $\pi_{0.5}$ *breakfast policy (Agilex):* 1,424 subtask episodes (374,722 frames at 30 Hz) yield 187,695 harvested frames; the expert dictionary is trained on 182,924 frames from 1,360 episodes and evaluated on 4,771 frames from 63 disjoint episodes, the language dictionary on 173,924 frames and evaluated on 13,771 frames from 134 disjoint episodes. Interpretation evidence comes from a pool of 240 training episodes (60 per subtask); verification uses only the held-out episodes. Open-loop intervention uses 1,016 frames from 24 stratified episodes (6 per subtask, stride 6). $\pi_{0.5}$ *LIBERO policy:* the LeRobot release of the LIBERO demonstrations (`physical-intelligence/libero`, 256×256, 10 Hz, 1,693 episodes over 40 tasks); we use 16 episodes per task for dictionary training (640 episodes, 107,294 frames) and 3 disjoint episodes per task for held-out annotation and dose calibration (120 episodes, 19,925 frames). Its dictionaries: EX $L \in \{5, 8, 11, 14\}$ with k=128, EX L=8 with k=64, and LM L=11 with k=256 ($d = 2048$, $m = 32768$); held-out variance explained is 0.999, 0.995, 0.967, 0.95 (EX by depth), 0.992 (EX k=64), and 0.652 (LM). Closed-loop LIBERO: 4 suites $\times$ 10 tasks $\times$ 10 initial states per condition; LIBERO-X: 50 tasks per group with the official initial states (Appendix J). The LIBERO-X calibration split is 30 LEVEL1 tasks disjoint from the 50 test tasks (two initial states each).

Interpreter and compute. The interpreter and blinded verifier are the same locally hosted vision-language model, Qwen3.5-27B (bf16, greedy decoding, reasoning mode disabled, transformers 5.15), so naming and verification use no external API. Each SAE trains for 30k steps in $\approx$ 20 min on one A800-80GB; harvesting 107,294 LIBERO frames (prefix pass plus 10 denoising steps) takes $\approx$ 1.5 GPU-hours; one LIBERO closed-loop suite (100 episodes) takes 35–75 min per process, and a LIBERO-X group (50 tasks, 600-step budget) 40–90 min. The simulation results in this paper used four A800-80GB GPUs for roughly one day.

The $\pi_{0.5}$ breakfast policy was converted from JAX to PyTorch with an adaRMS-aware procedure. On held-out real frames, the open-loop action MAE is 0.0129 (left-arm dimensions 0.002–0.006; right-arm dimensions 0.015–0.028). This is a conversion check, not a success rate.

Qualification rules. A feature passes the *open-loop screen* when, on held-out frames and relative to SAE-in-loop, (i) its dose response on the predeclared target channel is monotone over $\alpha \in \{\pm4, \pm8, \pm16\}\sigma_j$ (Spearman $\rho \geq 0.8$), (ii) the response reverses sign between $-\alpha$ and $+\alpha$ where the motor variable is signed, and (iii) $R_{\text{spec}} > 1$ (Eq. equation 3) with both bootstrap intervals excluding zero. The target channel of each unit is the action channel that its own held-out association predicts, so that dictionary features, raw residual dimensions, principal components, and random directions are scored identically. The *open-loop dose* α_{ol} is the largest grid dose at which every screened feature remains monotone, sign-symmetric, and near-linear ($|\Delta a(2\alpha)| \leq 3|\Delta a(\alpha)|$). The *closed-loop gate* selects α_{cl} as the largest grid dose whose paired success rate on held-out in-distribution rollouts stays within 3 *pp* of SAE-in-loop; a feature is a *steering handle* when it passes the screen and changes its named motor event at α_{cl} with the predicted sign. On the robot, closed-loop qualification additionally requires limited collateral motion and no increase in safety violations beyond one additional event per 12 paired trials. A locus is a *motor layer* when its dictionary is enriched in behavior-aligned features relative to the language locus, the aligned population spans several motor families, and representative members pass these tests against matched controls.

Predeclaration. Every feature, sign, dose, and rule was frozen in a time-stamped rule file before the outcome it governs, and the rule files are released with the code.

B SELF-COLLECTED AGILEX DATASETS

We collected the demonstrations of all three Agilex tasks on the Agilex Cobot Magic platform of Appendix K, which records the views of a top camera and two wrist cameras together with the 14-D action (six joints and one gripper per arm). Figures 7 and 8 show top-camera frames from one demonstration of each task.

Breakfast. The scene holds a rack with two bread slices, an unpowered toaster, a bottle of room-temperature colored water, a cup, and a plate. The task chains four subtasks, each with its own instruction (Fig. 7): *load bread* from the rack into the toaster slots, *activate toaster* by pushing its lever down, *pour drink* from the bottle into the cup, and *place bread* from the toaster onto the plate. The right arm handles the bread and the lever, and the left arm pours. The $\pi_{0.5}$ breakfast policy is fine-tuned on $1{,}850$ demonstrations: 350 of the full task and 600, 300, 400, and 200 of the four subtasks in the order above. The dictionaries and all analyses use $1{,}424$ subtask episodes ($374{,}722$ frames at 30 Hz), split by episode as in Appendix A.

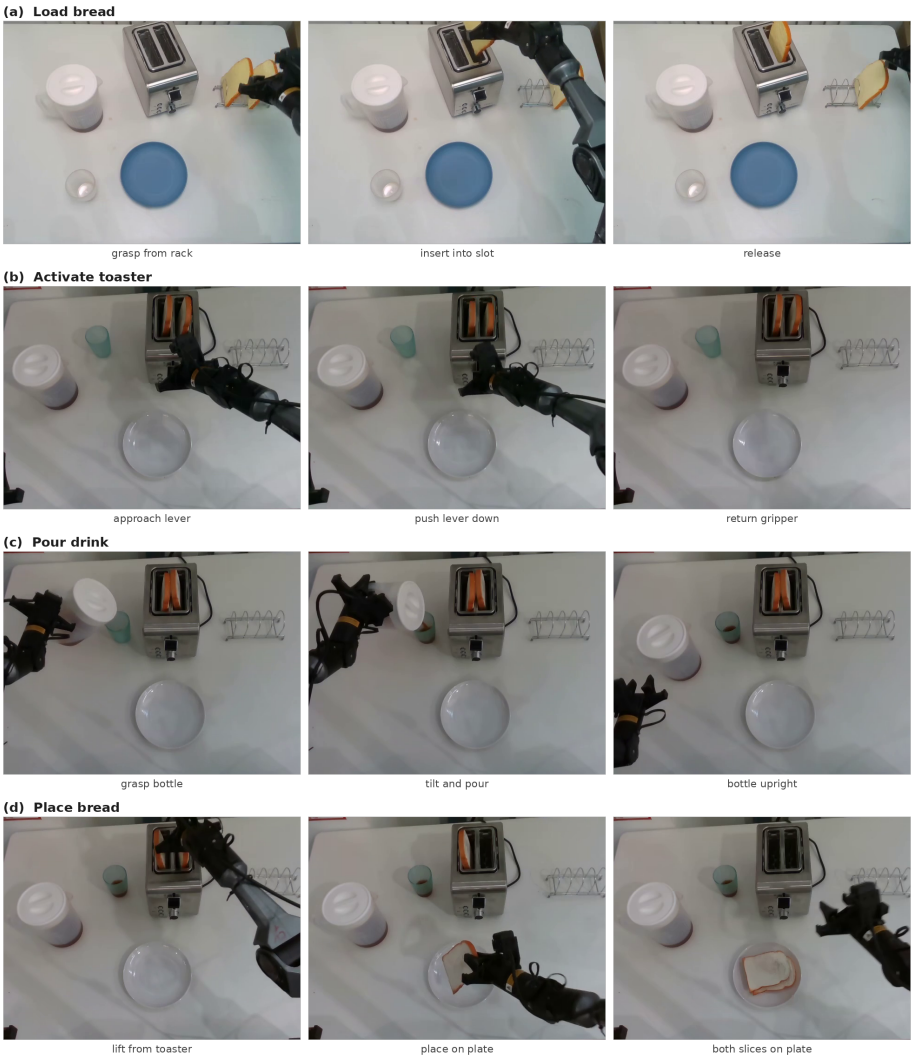

Figure 7: Breakfast dataset of the $\pi_{0.5}$ breakfast policy: top-camera frames from one demonstration of each subtask, (a) load bread, (b) activate toaster, (c) pour drink, and (d) place bread. The full task chains the four subtasks in this order.

Carton folding and package flipping. The $\pi_{0.5}$ folding policy is fine-tuned on 169 demonstrations in which both arms fold a flat carton blank into a box (Fig. 8a), with a relative-action parameterization. The $\pi_{0.5}$ flipping policy is fine-tuned on 160 demonstrations under the single instruction "turn the package so that the label faces up": starting from a stack, the packages are turned one after another so that each label faces up (Fig. 8b). For each task, one set of demonstrations trains the dictionaries (13,073 and 13,681 frames) and a disjoint set serves for analysis (5,550 and 10,221 frames; Appendix H).

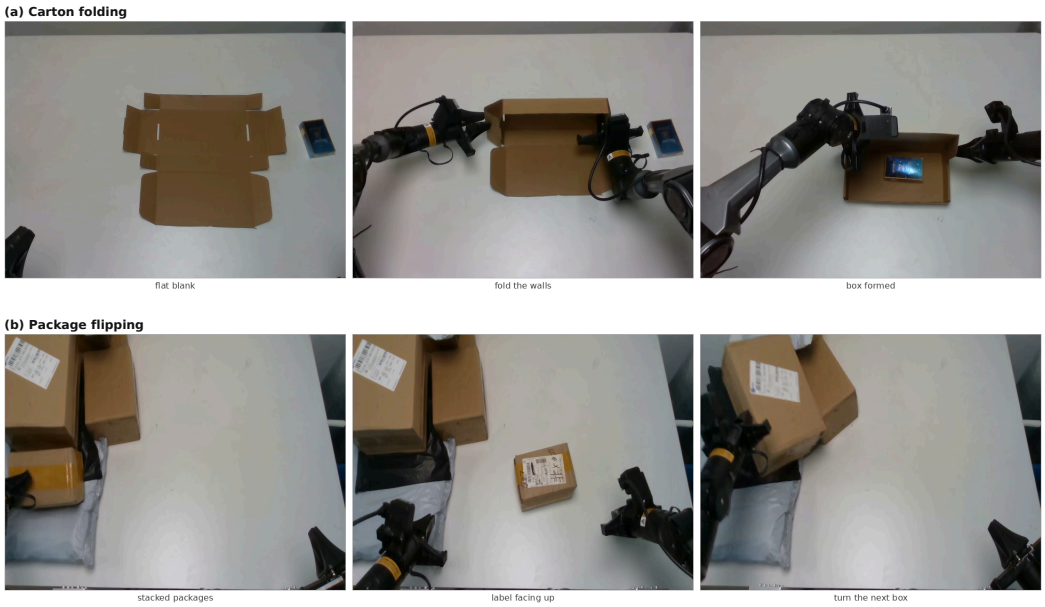

Figure 8: Carton-folding and package-flipping datasets: top-camera frames from one demonstration of (a) carton folding, used by the $\pi_{0.5}$ folding policy, and (b) package flipping, used by the $\pi_{0.5}$ flipping policy.

C Auto-Interpretation Protocol

For each feature, the interpreter receives $K = 6$ positive windows and the same number of matched negatives. Positive windows are the highest-activating peaks in the evidence pool with at least 40 frames between peaks and at most two peaks per episode; negatives are zero-activation frames drawn from the same subtasks as the peaks. Each window spans ± 8 frames around its center (≈ 0.5 s at 30 Hz) rendered as five top-camera frames at stride 4 plus both wrist views at the center frame, and is accompanied by the instruction, subtask, episode progress, and a textual summary of the synchronized proprioceptive/event traces (gripper aperture and velocity, arm speed, hand-height and wrist-pitch change, grasp/release events for each arm). The label-stability re-run draws a new evidence pool and uses its next-ranked peaks (rank offset 3) and a different negative sample; PCA components keep the basis fitted for the first run, so that both runs describe the same direction. Evidence is shuffled across features for the negative-control condition. The interpreter receives the prompt below and returns a verb–object–modifier label, a definition, visual, linguistic, and motor evidence, a confidence, and an abstention flag. Here `{k}` is K, `{image_note}` describes the attached images, and `{evidence_block}` gives, for each window, the instruction, subtask, episode progress, and proprioceptive/event summary; the braces inside the JSON are literal.

```
Evidence for one feature. It consists of {k} HIGH-activation windows
(where the feature fires strongly) and {k} LOW-activation windows
(matched control windows from the same subtasks where the feature is
silent).
{image_note}
{evidence_block}

Rules:
- Describe what is common to the HIGH windows but absent from the LOW
  windows.
- Do NOT infer the concept from the instruction text alone; the label
  must be supported by what the robot is doing (motion, gripper, wrist,
  object contact) across several HIGH windows.
- If the HIGH windows are inconsistent, abstain.

Answer in JSON only with keys:
"label": short verb-object-modifier label (e.g. "grasp-bottle-left",
  "tilt-wrist-pour", "approach-toaster-right"),
"definition": one sentence saying when the feature is active,
"evidence": {"visual": "...", "linguistic": "...", "motor": "..."}
  (each at most 25 words),
"confidence": number in [0,1],
"abstain": true or false
```

The held-out verifier sees no feature ID and no training examples: it receives the generated label and definition together with (i) K held-out pairs, each consisting of one high-activation and one zero-activation window in random order, and must name the matching window (forced choice), and (ii) the K highest-activating held-out windows, each to be marked as consistent or inconsistent with the description (top-K precision). We report two verifier variants: image-only windows, and windows accompanied by the same proprioceptive/event summary shown to the interpreter. Exact prompts, the model identifier (Qwen3.5-27B), and decoding parameters (greedy, reasoning disabled) are released with the code. Figure 9 shows six named features with their held-out evidence, and Table 4 reports the evidence ablations and unit-level baselines.

Human verification. One human rater, blinded to unit type, locus, feature identity, and the verifier's answers, answered exactly the items of the image+proprioception verifier: every held-out forced-choice pair and top-K window of the first frozen sample of 40 units per unit type. This covers the $\pi_{0.5}$ breakfast policy (SAE EX, SAE LM, EX neurons, EX principal components; $1{,}898$ decisions) and the $\pi_{0.5}$ folding and flipping policies of Appendix H (SAE EX, EX neurons, EX principal components; $1{,}443$ decisions each). Items were interleaved across unit types, shuffled with a fixed seed, and shown with the generated label and definition, the frame crops, and the verifier's proprioceptive summary; the rater answered every pair and could flag it as uncertain (14% of SAE EX pairs and 33–40% of the other pairs on the breakfast policy; 5–10% on the other two). Agreement with the verifier (Table 5) is substantial on window decisions ($\kappa = 0.66$–0.86 on the breakfast policy, 0.33–0.70 on the other two) and fair to moderate on pairs (0.32–0.55), where the human is also less accurate against ground truth. The human reproduces the verifier's unit-type ranking on the breakfast and flipping policies, but rates neuron and PCA labels at least as well as SAE labels on the folding policy, so the SAE margin is policy-dependent (pooled over the three policies: top-K 0.53 against 0.47 for neurons and 0.45 for PCA, differences $+0.07$ $[-0.02, +0.15]$ and $+0.08$ $[-0.00, +0.16]$). On the folding and flipping policies the human accepts more top-K windows than the verifier, so the verifier is the more conservative judge. A single rater bounds the verifier's reliability but cannot provide inter-rater agreement.

D Concept Naming Across Loci

Figures 4a and 10 summarize the labels of the five paired samples (40 random active SAE features per locus and policy, frozen before labeling), produced and verified by the unchanged pipeline of Appendix C. Verb families are assigned from the leading verb of each label with a fixed table: *grasp/hold* (grasp, grip, pick, hold, contact), *release/place* (release, place, insert, put, drop, open), *reach/transport* (approach, reach, transport, carry, return, lift, lower, raise, hover, retract), *reorient/tool* (tilt, pour, turn, rotate, flip, fold, push, press, pull, slide), and *task-level* (complete,

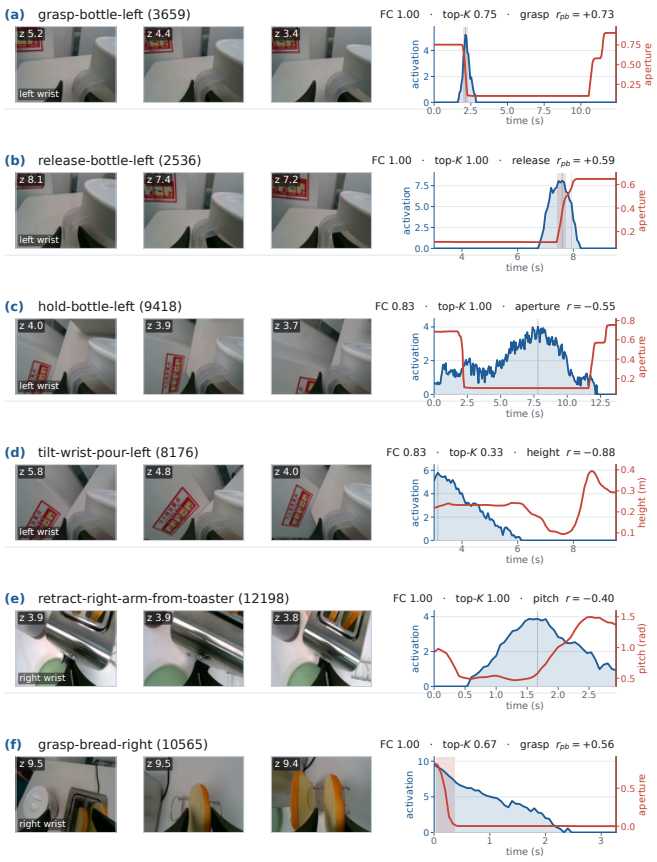

Figure 9: Held-out atlas of six action-expert features of the $\pi_{0.5}$ breakfast policy (EX L=8). Each row gives the feature's name in the concept table with its ID, held-out verifier scores (FC, top-K), and the feature's association with a gripper event or proprioceptive variable; frames are the three highest-activating moments from distinct held-out episodes, and the trace overlays activation (blue) on that motor variable (red) around the strongest frame, with the labeled grasp or release event shaded.

manipulate, idle, move, perform, execute, interact, handle, operate, finish); each of the 400 labels matched exactly one entry. Task-level names are rare at both loci (3% of action-locus and 7% of language-locus labels). On the $\pi_{0.5}$ breakfast policy they occur only at the language locus, 8 of 40 labels such as `complete-pouring-task` (32228) and `idle-toaster-active` (13685), whereas on the $\pi_{0.5}$ flipping policy six features at each locus carry a manipulate-package label, e.g. `manipulate-package-right` (5012) and `manipulate-package-left` (15939). The loci therefore differ less in what their names say than in whether the names hold on new episodes. Table 6 lists the pass counts and the three best-verified names of every policy and locus (ranked by top-K precision, then forced choice).

Verified language concepts under intervention. To test whether a verified name suffices to make a language feature a handle, every language-locus feature of the $\pi_{0.5}$ breakfast policy in the random sample whose label passes verification (5 of 40) was screened at $\pm 16\sigma_j$ with the rule of Appendix A, on the 1,016 frames of the population screen. The target is the channel group its label names (grasp, release, or hold: that arm's gripper; approach: that arm's six joints; tilt-wrist-pour: joints 3–5 of the pouring left arm). Targets, pass rule, and the card example of Fig. 4c were fixed in a rule file before the screen ran. One of the five passes on its named channel, `grasp-toaster-lever-right` (17726) at $R_{\text{spec}} = 1.18$, and it opens rather than closes the right gripper when amplified; the median is 0.93 (Table 7). Scored instead on the channel that each feature's own held-out association predicts, two pass, both at 1.09 on the right-arm joints. The effects are also small in absolute terms: the five target changes (5–7 $\times$ 10^{-4} in normalized action

Table 4: Auto-interpretation ablations and unit-level baselines, scored by the image+proprioception verifier on held-out windows (K=6 pairs and K=6 top windows per unit). Pass rate: forced-choice accuracy ≥ 0.75 and top-K precision ≥ 0.5. Full evidence includes visual frames, instructions, events, actions, and proprioception. Label stability is the fraction of units whose labels from two runs on largely different evidence (a re-drawn pool sharing 7–31% of its episodes, peak ranks 4–9 instead of 1–6, and a fresh negative sample) describe the same behavior; it measures consistency, not validity, which is why instruction-only labels are the most stable and the least often verified. Samples were frozen before labeling (n=80 rows pool two disjoint samples of 40; ablations use the first). The upper block is the $\pi_{0.5}$ breakfast policy; the lower block applies the unchanged pipeline to 40 random units of other loci and policies; across the nine matched-depth comparisons, the motor-proximal locus has the higher pass rate every time (105/360 against 36/360 pooled; Fisher $p < 10^{-10}$). Means with bootstrap 95% CIs.

Unit	Evidence / sample	n	Top-K precision ↑	Forced-choice accuracy ↑	Pass rate ↑	Label stability ↑
SAE feature (EX)	Full evidence, random active	80	0.42 [0.36, 0.49]	0.72 [0.67, 0.77]	32.5%	0.43 [0.28, 0.58]
SAE feature (EX)	Full evidence, behavior-aligned stratum	40	0.53 [0.43, 0.62]	0.80 [0.73, 0.86]	55.0%	0.58 [0.43, 0.73]
SAE feature (EX)	Instruction only	40	0.15 [0.08, 0.22]	0.48 [0.40, 0.57]	5.0%	0.85 [0.73, 0.95]
SAE feature (EX)	Without visual frames	40	0.38 [0.29, 0.47]	0.70 [0.62, 0.77]	37.5%	0.43 [0.28, 0.58]
SAE feature (EX)	Without proprioception / events	40	0.18 [0.11, 0.27]	0.58 [0.48, 0.67]	10.0%	0.68 [0.53, 0.80]
SAE feature (EX)	Shuffled evidence packet	40	0.09 [0.03, 0.15]	0.48 [0.39, 0.57]	7.5%	0.45 [0.30, 0.60]
SAE feature (LM)	Full evidence, random active	40	0.24 [0.16, 0.33]	0.56 [0.50, 0.63]	12.5%	0.18 [0.08, 0.31]
Neuron (EX)	Full evidence, random	80	0.26 [0.20, 0.32]	0.63 [0.57, 0.69]	17.5%	0.33 [0.18, 0.48]
PCA component (EX)	Full evidence, leading components	40	0.28 [0.19, 0.37]	0.61 [0.53, 0.69]	15.0%	0.20 [0.08, 0.33]
Other loci and policies (full evidence, random units; verifier only; EX SAE and EX neurons of the $\pi_{0.5}$ folding and flipping policies: Table 5)						
Neuron (LM)	$\pi_{0.5}$ breakfast policy	40	0.25 [0.17, 0.33]	0.53 [0.46, 0.60]	7.5%	0.05 [0.00, 0.13]
SAE feature (EX)	$\pi_{0.5}$ LIBERO policy	40	0.35 [0.26, 0.45]	0.67 [0.59, 0.74]	30.0%	0.17 [0.06, 0.31]
Neuron (EX)	$\pi_{0.5}$ LIBERO policy	40	0.42 [0.33, 0.52]	0.70 [0.60, 0.78]	32.5%	0.28 [0.15, 0.43]
SAE feature (LM)	$\pi_{0.5}$ LIBERO policy	40	0.31 [0.23, 0.39]	0.53 [0.45, 0.60]	10.0%	0.15 [0.05, 0.28]
Neuron (LM)	$\pi_{0.5}$ LIBERO policy	40	0.30 [0.22, 0.38]	0.55 [0.49, 0.62]	10.0%	0.08 [0.00, 0.18]
SAE feature (LM)	$\pi_{0.5}$ folding policy	40	0.26 [0.18, 0.35]	0.52 [0.45, 0.58]	7.5%	0.25 [0.13, 0.40]
Neuron (LM)	$\pi_{0.5}$ folding policy	40	0.15 [0.09, 0.20]	0.52 [0.43, 0.60]	2.5%	0.35 [0.20, 0.50]
SAE feature (LM)	$\pi_{0.5}$ flipping policy	40	0.36 [0.26, 0.47]	0.57 [0.49, 0.64]	15.0%	0.13 [0.03, 0.23]
Neuron (LM)	$\pi_{0.5}$ flipping policy	40	0.31 [0.22, 0.40]	0.51 [0.45, 0.58]	7.5%	0.15 [0.05, 0.28]
SAE feature (action L=16)	OpenVLA-OFT LIBERO policy	40	0.43 [0.34, 0.53]	0.70 [0.63, 0.78]	35.0%	0.20 [0.08, 0.33]
SAE feature (action L=31)	OpenVLA-OFT LIBERO policy	40	0.47 [0.36, 0.58]	0.73 [0.65, 0.80]	37.5%	0.30 [0.18, 0.45]
SAE feature (prompt L=16)	OpenVLA-OFT LIBERO policy	40	0.44 [0.35, 0.53]	0.60 [0.51, 0.69]	17.5%	0.20 [0.08, 0.33]

Table 5: Human verification of automatic labels on the identical held-out items answered by the VLM verifier (first frozen sample of 40 units per row; K=6 pairs and K=6 top windows per unit; one blinded rater). Human columns are unit-level means with bootstrap 95% CIs; VLM columns are point estimates on the same units. κ is Cohen's kappa between the human and the VLM on the per-pair (A/B) and per-window (consistent / not) decisions. Pass rate is the fraction of units with forced-choice accuracy ≥ 0.75 and top-K precision ≥ 0.5 under each judge. The pooled rows concatenate the three policies (120 units per unit type).

Policy	Unit	n	FC human ↑	FC VLM	Top-K human ↑	Top-K VLM	κ pairs	κ windows	Pass human / VLM
$\pi_{0.5}$ breakfast	SAE feature (EX)	40	0.72 [0.62, 0.81]	0.68	0.43 [0.33, 0.53]	0.39	0.55	0.76	37.5% / 30.0%
$\pi_{0.5}$ breakfast	Neuron (EX)	40	0.65 [0.57, 0.73]	0.68	0.32 [0.23, 0.41]	0.30	0.32	0.86	17.5% / 22.5%
$\pi_{0.5}$ breakfast	PCA component (EX)	40	0.62 [0.54, 0.71]	0.61	0.30 [0.20, 0.40]	0.28	0.35	0.66	25.0% / 15.0%
$\pi_{0.5}$ breakfast	SAE feature (LM)	40	0.57 [0.51, 0.64]	0.56	0.23 [0.15, 0.32]	0.24	0.32	0.77	10.0% / 12.5%
$\pi_{0.5}$ folding	SAE feature (EX)	40	0.64 [0.56, 0.72]	0.62	0.61 [0.50, 0.71]	0.40	0.34	0.47	25.0% / 25.0%
$\pi_{0.5}$ folding	Neuron (EX)	40	0.71 [0.62, 0.80]	0.76	0.57 [0.49, 0.66]	0.46	0.49	0.70	45.0% / 47.5%
$\pi_{0.5}$ folding	PCA component (EX)	40	0.72 [0.65, 0.80]	0.71	0.55 [0.46, 0.64]	0.40	0.52	0.54	45.0% / 30.0%
$\pi_{0.5}$ flipping	SAE feature (EX)	40	0.68 [0.60, 0.75]	0.67	0.56 [0.46, 0.66]	0.44	0.42	0.39	32.5% / 22.5%
$\pi_{0.5}$ flipping	Neuron (EX)	40	0.57 [0.49, 0.65]	0.62	0.51 [0.42, 0.61]	0.32	0.45	0.47	22.5% / 17.5%
$\pi_{0.5}$ flipping	PCA component (EX)	40	0.58 [0.50, 0.66]	0.61	0.51 [0.42, 0.60]	0.31	0.35	0.33	27.5% / 20.0%
All three policies	SAE feature (EX)	120	0.68 [0.63, 0.73]	0.66	0.53 [0.47, 0.60]	0.41	0.44	0.53	31.7% / 25.8%
All three policies	Neuron (EX)	120	0.64 [0.59, 0.69]	0.69	0.47 [0.41, 0.53]	0.36	0.42	0.67	28.3% / 29.2%
All three policies	PCA component (EX)	120	0.64 [0.60, 0.69]	0.64	0.45 [0.40, 0.51]	0.33	0.40	0.51	32.5% / 21.7%

units) are an order of magnitude below that of the expert row `grasp-bottle-left` (3659) on its gripper (4.9×10^{-3}) and at the level of the language-locus control `grasp-bottle-left` (7779) (5.0×10^{-4}). At the population level, the 20 most motion-aligned language features of this policy change their predicted channel by a median of 1.9×10^{-3}, against 3.0×10^{-2} for the 30 most aligned expert features.

Concept cards on the $\pi_{0.5}$ folding and flipping policies. Figures 11 and 12 apply the three checks of Fig. 4 to six behavior-aligned expert features per policy, frozen before labeling: the most aligned feature of every motor family among the 30 features of the population screen, completed with the next most aligned ones. Names come from the unchanged pipeline, and the handle check is the policy's population screen on each feature's predicted channel (538 and 625 held-out frames).

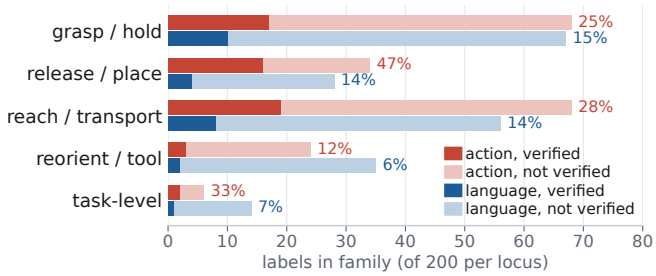

Figure 10: Label families pooled over the five policies (200 labels per locus): bar length counts the labels in a family, the dark segment those that pass held-out verification.

Table 6: Verified concept names per policy and locus. Pass: names that pass held-out verification among 40 random active SAE features. Names: the three best-verified names (top-K precision, then forced choice); for the EX locus of the $\pi_{0.5}$ breakfast policy, the best-verified names that also appear in the concept table of Appendix E.

Policy	Locus	Pass	Verified names
$\pi_{0.5}$ breakfast	EX	12/40	`hold-bottle-left` (13673); `push-toaster-lever-right` (7343); `release-bread-right` (7140)
	LM	5/40	`release-object-right` (31063); `hold-bottle-left` (29781); `approach-cup-left` (26382)
$\pi_{0.5}$ folding	EX	10/40	`grasp-carton-right` (6612); `hold-folded-flaps-still` (9354); `grasp-flap-left` (1251)
	LM	3/40	`hold-folded-flaps` (11065); `hold-box-structure` (16462); `grasp-flap-right` (12372)
$\pi_{0.5}$ flipping	EX	9/40	`hold-package-stationary` (12844); `rotate-package-right-arm` (5208); `approach-package-left` (7809)
	LM	6/40	`hold-package-left` (11031); `manipulate-package-right` (13831); `turn-package-label-up` (23346)
$\pi_{0.5}$ LIBERO	EX	12/40	`approach-object-initial` (808); `approach-table-fast-open` (7508); `approach-object-table` (15802)
	LM	4/40	`grasp-object` (3527); `approach-object-open-gripper` (6506); `grasp-object` (11538)
OpenVLA-OFT LIBERO	action	14/40	`release-object` (14673); `release-object` (3254); `grasp-bowl` (10061)
	prompt	7/40	`transport-object-fast` (5716); `release-object-near-target` (12416); `release-object-stationary` (15449)

Nine of the twelve features are handles, but only three carry verified names. On the $\pi_{0.5}$ folding policy, five of the six are named as flap manipulations, such as `grasp-flap-left` (4311), `fold-flap-left` (6589), and `fold-flap-right` (8392), although their strongest association is joint speed; only the last survives verification, while four features whose names fail are among the most selective handles we measured (2.6–6.9×). On the $\pi_{0.5}$ flipping policy, the two gripper features carry verified names that match their events, `grasp-package-right` (11963) and `release-package-right` (13100), but only the latter is a handle (5.1×, sign-reversing); gripper-targeted units are the hard cases for open-loop steering on this policy (Appendix H).

Table 7: Open-loop screen of the five language concepts of the $\pi_{0.5}$ breakfast policy whose names pass verification ($|\alpha|{=}16\sigma_j$, 1,016 held-out frames). Pass: $R_{\text{spec}} > 1$ and, for a single gripper channel, sign reversal with both intervals excluding zero.

	Channel named by label		Channel predicted by association		
Row	R_{spec}	Pass	Channel (r)	R_{spec}	Pass
`release-object-right` (31063)	1.16	no	R release (+0.12)	1.16	no
`hold-bottle-left` (29781)	0.88	no	R arm motion (−0.12)	1.09	yes
`approach-cup-left` (26382)	0.93	no	L joint speed (+0.14)	0.93	no
`tilt-wrist-pour` (12301)	0.89	no	L arm motion (+0.10)	0.92	no
`grasp-toaster-lever-right` (17726)	1.18	yes	R arm motion (+0.10)	1.09	yes

Two left-arm features are named after the right arm, `push-package-right` (5781) and `hold-package-right` (10017); both fail verification and move the left arm 2.7–3.7× more than the other channels. The main-text cards (Fig. 4e–f) are the best-verified feature of each policy under the frozen rule.

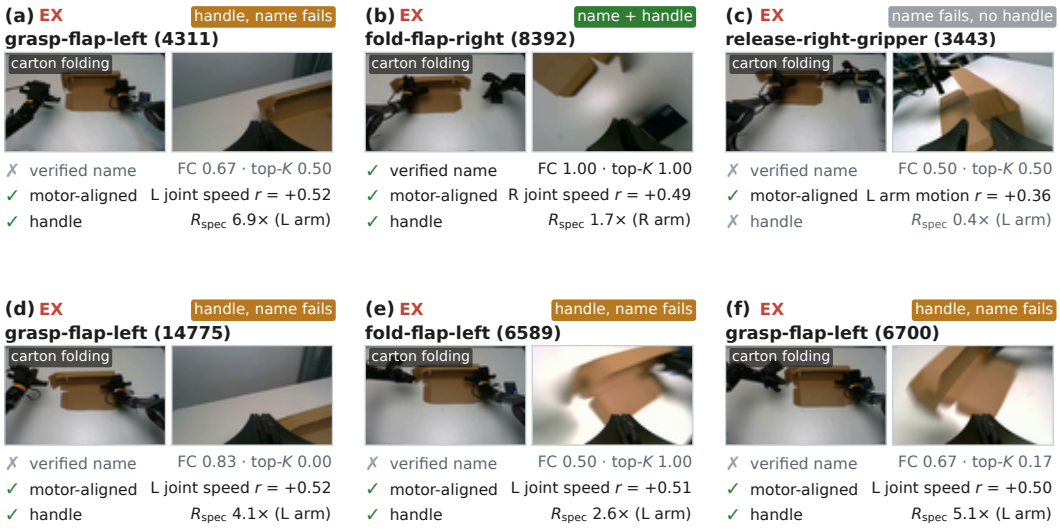

Figure 11: Concept cards of the $\pi_{0.5}$ folding policy for the six frozen behavior-aligned expert features (checks as in Fig. 4; handle check on the policy's population screen, 538 held-out frames).

E THE CONCEPT TABLE OF THE $\pi_{0.5}$ BREAKFAST POLICY

The table protocol was frozen in a rule file before any table label or screen existed: the scope (every EX $L{=}8$ feature with held-out fire rate above 0.5%, 513 features), labels from one fresh run of the unchanged pipeline (seed-1 training pool, $K{=}6$), the name-to-channel map, the readout metrics, and the steering criterion. The verb map assigns grasp, grip, pick, hold, contact, and close to the named arm's gripper with an expected closing response to $+\alpha$; release, open, and drop to the gripper with an opening response; tilt, pour, rotate, turn, flip, and twist to joints 3–5 of that arm; and approach, reach, transport, carry, return, lift, lower, raise, hover, retract, move, push, press, pull, slide, place, insert, put, fold, extend, position, align, and withdraw to its six arm joints. A name without an arm token maps to both arms; task-level verbs (complete, manipulate, idle, interact, monitor, . . .) and the five labels whose leading token is not a verb, such as `toaster-lever-pressed-right` (1748), are readout-only. Of the 511 named features, 175 carry closing gripper verbs, 111 opening verbs, 116 arm verbs, 74 wrist verbs, and 35 are readout-only.

Readout metrics. The predeclared event criterion (a matching gripper name active anywhere within ±15 frames of an event) saturates: with 188 verified names it is met for every event also when names are shuffled across rows, so it cannot distinguish a correct table from a scrambled one.



Figure 12: Concept cards of the $\pi_{0.5}$ flipping policy for the six frozen behavior-aligned expert features (checks as in Fig. 4; handle check on the policy's population screen, 625 held-out frames).

We therefore report, as an exploratory criterion not in the rule, whether a matching name ranks among the three top verified rows within the window (100% against 80% shuffled, permutation $p = 0.001$; all names 99.3% against 82%). The arm readout is correct on 93.3% of the 494 frames in which only the left arm moves and on 99.4% of the 3,168 in which only the right arm moves (balanced 96.4%; all names 92.1%).

Steering by name. All 513 features were screened once on the frames of the population screen and scored afterwards on the channel their name implies, so the screen could not depend on the names. The arm-swapped baseline (exploratory) scores each named row on the same channel type of the other arm. Table 8 lists the best-verified row per verb type and arm; the two handles among them are opening-gripper rows.

Table 8: Querying the table by name: the best-verified row (top-K, then forced choice, then fire rate) for each verb type and arm, with its specificity on the channel its name implies ($|\alpha|{=}16\sigma_j$) and whether it passes the screen.

Verb type	Arm	Row	FC / top-K	R_{spec}	Handle
arm motion	left	`return-bottle-left` (5846)	1.00/0.75	0.55	no
arm motion	right	`retract-right-arm-from-toaster` (12198)	1.00/1.00	0.65	no
gripper close	both	`hold-bottle-pouring` (4494)	1.00/1.00	0.73	no
gripper close	left	`hold-bottle-left` (13673)	1.00/1.00	0.45	no
gripper close	right	`grasp-toaster-lever-right` (8202)	1.00/1.00	0.87	no
gripper open	both	`release-object-grippers` (6617)	1.00/1.00	10.74	yes
gripper open	left	`release-bottle-left` (2536)	1.00/1.00	3.09	yes
gripper open	right	`release-bread-right` (398)	1.00/1.00	0.81	no
wrist	both	`tilt-wrist-pour` (16135)	1.00/1.00	0.56	no
wrist	left	`tilt-wrist-pour-left` (3900)	1.00/1.00	0.96	no

Names of the closed-loop rows. Labeled after their closed-loop tests with the same pipeline: on the $\pi_{0.5}$ LIBERO policy, the main closed-loop row `approach-stove-top` (9168) verifies (forced choice 1.00, top-K 0.50), as do `grasp-object-closing` (4270) and `open-gripper-command` (3701) (1.00/1.00 each), whereas the second-seed row `approach-bowl-open` (868) (0.67/1.00) and the height row `grasp-object` (7165) of the OpenVLA-OFT LIBERO policy (0.50/0.17) fail. On the $\pi_{0.5}$ breakfast policy, the hardware rows are `grasp-bottle-left` (3659), `tilt-wrist-pour-left` (8176), `release-bottle-left` (8988), and the random control `hold-bottle-left` (10099); only

the first verifies, and all but the control pass the screen on their named channel. The language-locus control `grasp-bottle-left` (7779) fails verification (0.50/0.00), and the controls and population features named elsewhere in the appendix were labeled in the same way after their tests.

F Full Faithfulness Results

Table 9 gives the closed-loop faithfulness of each interface used for steering, and Table 10 the variance-explained sweep of the $\pi_{0.5}$ breakfast policy. Table 11 gives the language-locus sweep of the $\pi_{0.5}$ LIBERO policy behind the $6\,pp$ loss on LIBERO-10 quoted in Sec. 5.1; with 50 episodes per cell, single cells vary by several points, so we read it only as a warning against steering at the language locus.

Table 9: Faithfulness: SAE-in-loop minus Frozen VLA on paired episodes (95% CI, McNemar p), and mean open-loop action discrepancy on held-out frames (per-channel s.d. for $\pi_{0.5}$; fraction of the $[-1, 1]$ range for OpenVLA-OFT).

Policy (locus)	Interface	n	Δ success (pp)	p	Discrepancy
$\pi_{0.5}$ LIBERO (EX L=8)	SAE, k=128	400	-1.5 $[-4.5, +1.5]$	0.41	0.8%
	PCA, 128	400	-1.0 $[-4.0, +2.0]$	0.61	0.7%
OpenVLA-OFT LIBERO (action L=16)	SAE, k=128	200	$+1.0$ $[-2.0, +4.0]$	0.75	0.7%
	SAE, k=512	200	$+1.5$ $[-1.0, +4.0]$	0.45	0.6%
$\pi_{0.5}$ breakfast (EX L=8)	SAE, k=128	80	-2.5 $[-9.0, +4.0]$	0.7	2.6%

Table 10: Held-out variance explained for the SAE sweep of the $\pi_{0.5}$ breakfast policy. Action expert: $d = 1024$, $m = 16384$; language: $d = 2048$, $m = 32768$. Higher is better.

Representation locus / k	64	128	256
LM L=5	0.797	0.824	0.852
LM L=8	0.748	0.768	0.800
LM L=11	0.617	0.647	0.684
LM L=14	0.554	0.583	0.629
EX L=5	0.998	0.999	0.999
EX L=8	0.990	0.993	0.996
EX L=11	0.973	0.983	0.989
EX L=14	0.929	0.949	0.968

Table 11: Reconstruction effect Δ_{recon} (percentage-point success-rate change from Frozen VLA to SAE-in-loop) for the language-locus sweep of the $\pi_{0.5}$ LIBERO policy. Pooled reconstructs all attended language tokens; last-valid reconstructs the final non-padding token. Each cell uses 5 episodes for each of 10 tasks.

Suite	P k=64	P k=128	P k=256	LV k=64	LV k=128	LV k=256
Spatial	$+2$	$+4$	$+2$	0	$+2$	$+2$
Object	0	-4	-2	-10	0	-2
Goal	0	0	-6	-2	-2	-2
LIBERO-10	-6	-10	0	-4	0	-4

G Representation-Locus Robustness

Figure 13 summarizes the motor layer of the $\pi_{0.5}$ breakfast policy. The motor-layer claim must also survive probe selection and SAE randomness. Table 12 repeats the LM–EX comparison at matched relative depth (layer 8 in both 18-layer stacks), matched sparsity (k=128 for both dictionaries), matched activation density (the LM dictionary restricted to its 513 most frequently firing active features, the number of active EX features), and across four EX SAE seeds (0–3; the default row uses seed 0). Enrichment intervals are feature-level bootstraps (10,000 resamples of the active-feature sets). Across seeds the EX dictionary retains 480.8 ± 35.2 active features with $37.1 \pm 4.9\%$

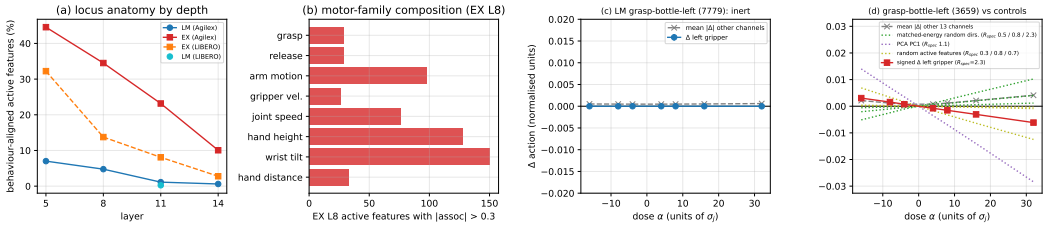

Figure 13: Motor-layer evidence on the $\pi_{0.5}$ breakfast policy. (a) Behavior-aligned fraction of active features by layer, LM vs. EX (legend: Agilex), with the dictionaries of the $\pi_{0.5}$ LIBERO policy (legend: LIBERO) for comparison. (b) Motor-family composition of the 513 active EX features (held-out). (c) Signed open-loop response of the language-locus control `grasp-bottle-left` (7779) on the left-gripper channel, $\alpha \in [-16, 32]\sigma_j$ (1,016 frames, bootstrap 95% CIs). (d) Signed left-gripper response of the expert row `grasp-bottle-left` (3659) and mean collateral change over the other 13 channels; dotted lines: the three matched-energy random directions, PCA PC1, and the three random active features of Table 15 (legend: per-unit R_{spec} at $16\sigma_j$). One random direction is as selective as this row, which is why selectivity is judged on unit populations (Table 16).

behavior-aligned and $17.2 \pm 2.2\%$ task-selective (mean $\pm$ s.d.), and every one of the ten motor families used for screening (grasp, release, arm motion, gripper velocity, joint speed, each per arm) is recovered in all four seeds. The prevalence contrast also decreases monotonically with depth at the expert locus (44.9%, 34.5%, 23.2%, 10.1% for EX $L \in \{5, 8, 11, 14\}$ versus 4.8%, 4.8%, 1.1%, 0.6% for LM at the same layers). The same anatomy appears in the $\pi_{0.5}$ LIBERO policy under identical definitions: 13.8% of active EX L=8 features versus 0.24% of active LM L=11 features are behavior-aligned ($56\times$; raw residual dimensions 56.0% vs. 29.7%, $1.9\times$ $[1.7, 2.1]$), and naive AUROC would mark 99.99% of active LM features as task-selective versus 0.36% after tie correction.

Reconstruction quality does not explain the gap. The two loci cannot be brought to the same held-out variance explained with the same dictionary architecture: raising the LM dictionary from k=256 to k=512 and k=1024 moves its VarExp from 0.65 to 0.69 and 0.73 on the $\pi_{0.5}$ LIBERO policy ($0.66 \to 0.74 \to 0.82$ on the $\pi_{0.5}$ breakfast policy), while a deliberately small EX dictionary (k=16) still explains 0.97 ($\pi_{0.5}$ LIBERO) and 0.95 ($\pi_{0.5}$ breakfast). The aligned fraction is flat along the LM capacity sweep ($0.24\% \to 0.21\% \to 0.25\%$ for $\pi_{0.5}$ LIBERO; $1.1\% \to 0.9\% \to 0.8\%$ for $\pi_{0.5}$ breakfast) and the k=16 EX dictionary remains enriched over every LM dictionary (7.9% aligned for $\pi_{0.5}$ LIBERO, 31–36×; 16.1% for $\pi_{0.5}$ breakfast, 14–19×; Table 12). The dense-unit rows of Table 1 make the same point without any dictionary. Each analysis uses the same held-out behavioral labels and tie-corrected task-selectivity estimator as Table 1.

Table 12: Robustness of the representation-locus contrast. Behavior-aligned and task-selective entries are percentages of active features. Enrichment is the EX-to-LM ratio in behavior-aligned prevalence with its 95% confidence interval; the default rows repeat the main comparison.

Analysis	LM behavior-aligned	EX behavior-aligned	EX/LM enrichment ↑	LM task-selective	EX task-selective	Enrichment 95% CI
$\pi_{0.5}$ breakfast policy						
Default probes (LM L=11, k=256 vs EX L=8, k=128)	1.1%	34.5%	30.7×	0.2%	16.8%	[23.8, 40.7]
Matched relative depth (LM L=8 vs EX L=8)	4.8%	34.5%	7.2×	1.3%	16.8%	[5.8, 9.2]
Matched sparsity (LM k=128 vs EX k=128)	1.9%	34.5%	18.6×	0.2%	16.8%	[14.3, 24.9]
Matched activation density (top-513 LM features)	8.0%	34.5%	4.3×	1.9%	16.8%	[3.2, 6.1]
Cross-seed aggregate (EX seeds 0–3, pooled)	1.1%	$37.1 \pm 4.9\%$	32.8×	0.2%	$17.2 \pm 2.2\%$	[26.3, 42.6]
Higher-capacity LM (LM k=1024, VarExp 0.82, vs EX k=128)	0.8%	34.5%	41.5×	0.2%	16.8%	[33.1, 53.5]
Low-capacity EX (EX k=16, VarExp 0.95, vs LM k=256)	1.1%	16.1%	14.3×	0.2%	1.6%	[9.7, 20.6]
Raw residual dimensions (LM L=11 vs EX L=8, no dictionary)	21.5%	49.3%	2.3×	8.5%	43.6%	[2.1, 2.5]
$\pi_{0.5}$ LIBERO policy						
Default probes (LM L=11, k=256 vs EX L=8, k=128)	0.24%	13.8%	56.3×	0.36%	8.1%	[39.9, 85.4]
Higher-capacity LM (LM k=1024, VarExp 0.73, vs EX k=128)	0.25%	13.8%	54.2×	0.59%	8.1%	[40.9, 73.4]
Low-capacity EX (EX k=16, VarExp 0.97, vs LM k=256)	0.24%	7.9%	32.1×	0.36%	2.4%	[18.4, 54.1]

Under the stricter core rule (association > 0.3 with a gripper event or a core proprioceptive variable), 90 of 1,024 EX residual dimensions and none of the 2,048 LM dimensions are aligned on the $\pi_{0.5}$ LIBERO policy, so the locus contrast holds for dense units at both thresholds.

Second backbone: the OpenVLA-OFT LIBERO policy. To test whether the anatomy depends on the two-stream design of $\pi_{0.5}$, we repeated the pipeline on the OpenVLA-OFT LIBERO policy, the public LIBERO checkpoint of OpenVLA-OFT (Kim et al., 2025), a Llama-2-7B single stream with 56 action-placeholder tokens and an L1 regression head; it succeeds on $19/20$ LIBERO-Spatial episodes on our machine, as reported. Without an action expert, the two loci are the residuals at the *prompt* positions (last prompt token) and at the *action-token* positions (mean over the 56 placeholders) of the same layer, harvested and split as for the $\pi_{0.5}$ LIBERO policy (107,294 training and 19,925 held-out frames). Dictionaries (d=4096, m=16,384, k=128, 10k steps) were trained at layers 8, 16, 24, and 31 of the 32-layer stack. Figure 14 summarizes three findings. (i) *Faithfulness replicates*: the action-token SAE at layer 16 changes LIBERO success from 96.5% to 97.5% (k=128) or 98.0% (k=512; n=200; Table 9). (ii) *Motor alignment concentrates at the output-proximal layer, not at the prompt/action split*: the aligned fraction of active action-token features is 2.5%, 0.6%, 0.5%, and 18.9% at layers 8, 16, 24, and 31 (held-out variance explained 0.90, 0.73, 0.65, 0.93), against 0.4% for the prompt-token dictionary at layer 16. At matched layer the action-over-prompt enrichment is only $1.7\times$ [1.1, 2.6], it vanishes at matched activation density, and raw residual dimensions agree (20.3% against 19.0% at layer 16 under the rule of Table 1). The reason is architectural: OpenVLA-OFT feeds proprioception as a token before the prompt, so its prompt positions read the robot state, whereas the language prefix of $\pi_{0.5}$ never does. The motor layer therefore transfers as the residual nearest to action generation (layer 31 here, the action expert in $\pi_{0.5}$), not as the claim that language positions are motor-inert. (iii) *Handles behave as on the $\pi_{0.5}$ LIBERO policy*: at layer 16, 23 of the 30 most aligned action-token features predict the saturated gripper channel and fail like every unit type there ($1/23$), while all six height-targeted aligned features pass (median R_{spec} 1.6). At layer 31 (145 held-out frames), 26 of the 30 aligned features again predict the gripper (27% passing), and both unit types steer the height channel with high specificity (median $4.2\times$ for the four height-targeted features, $3.4\times$ for the 29 height-targeted residual dimensions). The four height features were then taken into closed loop (Table 20): the layer-31 dictionary is faithful (98.0% against 96.5% frozen), the most specific feature passes the gate at $4\sigma_j$ and delays the descent by 4.6 steps, and all four shift the descend onset with the predicted sign. On this 7-D single-arm benchmark the dictionary thus offers discovery, naming, and gated closed-loop handles, but no per-unit open-loop advantage over the best raw dimensions, as for the $\pi_{0.5}$ LIBERO policy; the open-loop advantage is a property of the 14-D bimanual $\pi_{0.5}$ policies, whose action space leaves room for selectivity.

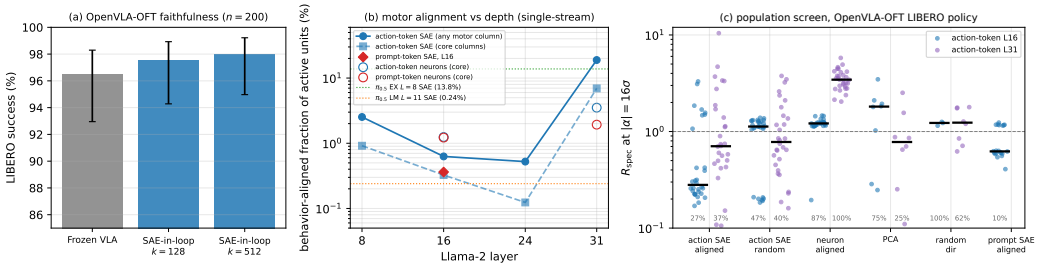

Figure 14: Second backbone: the OpenVLA-OFT LIBERO policy. (a) Closed-loop faithfulness of the action-token SAE at layer 16 (n=200 paired episodes, Wilson 95% CIs). (b) Behavior-aligned fraction of active units versus depth for the action-token dictionaries (filled), the prompt-token dictionary at layer 16 (diamond), and raw residual dimensions of both positions (hollow); dotted lines are the EX and LM dictionaries of the $\pi_{0.5}$ LIBERO policy. (c) Population open-loop screen at the action-token locus at layers 16 (blue) and 31 (purple); unit-set rules as in Fig. 15, pass rates below each group.

H THE $\pi_{0.5}$ FOLDING AND FLIPPING POLICIES

To test whether the motor-layer anatomy and the handle population are specific to the $\pi_{0.5}$ breakfast policy, we repeated the locus analysis and the population screen on the $\pi_{0.5}$ folding and flipping policies, fine-tuned on the same Agilex Cobot Magic platform for bimanual carton-box folding (169 demonstrations, relative-action parameterization) and single-arm package flipping ("turn the

package so that the label faces up"; 160 demonstrations; Appendix B). Each policy was converted from the JAX release, its EX L=8 final-step activations were harvested on one demonstration set (SAE training; 13,073 and 13,681 frames) and on a disjoint set of the same task (analysis; 5,550 and 10,221 frames), an EX dictionary (k=128, m=16,384, 15k steps; held-out VarExp 0.955 and 0.972) was trained per policy, and the frozen rules of Table 1 were applied to SAE features, raw EX and LM residual dimensions, and the leading principal components. The population screen used the same unit-set rules as Fig. 15 on 538 and 625 held-out frames (16 episodes each; matched-energy random directions 1.4× and 1.0×). Because these dictionaries were trained on only ∼ 13k frames they retain many more active features than the dictionary of the $\pi_{0.5}$ breakfast policy (5,472 and 5,627 versus 513), so aligned *counts* rather than fractions are the comparable quantity in Table 13.

Table 13: The $\pi_{0.5}$ folding and flipping policies. *Anatomy*: behavior-aligned units under the identical ten-column rule (core $|\mathrm{assoc}| > 0.3$) with the SAE-free EX/LM residual-dimension ratio (feature-level bootstrap 95% CI). *Handles*: population open-loop screen at $|\alpha|$=16σ (pass rate; median R_{spec} with bootstrap CI).

Policy	Quantity	EX SAE features	EX residual dims	LM residual dims	EX PCA (128)
$\pi_{0.5}$ folding	aligned units (of active)	58 (of 5,472)	314 (of 1,024; 30.7%)	47 (of 2,048; 2.3%)	3
$\pi_{0.5}$ folding	peak association	0.52	0.59	0.46	0.32
$\pi_{0.5}$ folding	EX/LM dense ratio		13.4× [10.1, 18.5]		
$\pi_{0.5}$ folding	screen: pass rate	**96.7**% (aligned 30); 50% (random 30)	20% (aligned 30)	65% (LM SAE, aligned 20)	37.5% (8)
$\pi_{0.5}$ folding	screen: median R_{spec}	**3.10** [2.60, 3.86]; random 1.01	0.93 [0.86, 0.95]	1.16 [0.98, 1.89] (LM SAE)	0.87
$\pi_{0.5}$ flipping	aligned units (of active)	49 (of 5,627)	448 (of 1,024; 43.8%)	467 (of 2,048; 22.8%)	3
$\pi_{0.5}$ flipping	peak association	0.70	0.73	0.67	0.58
$\pi_{0.5}$ flipping	EX/LM dense ratio		1.9× [1.7, 2.1]		
$\pi_{0.5}$ flipping	screen: pass rate	76.7% (aligned 30); 63% (random 30)	70% (aligned 30)	**85**% (LM SAE, aligned 20)	75% (8)
$\pi_{0.5}$ flipping	screen: median R_{spec}	2.29 [1.37, 2.61]; random 1.06	1.12 [1.06, 1.24]	**3.91** [3.07, 5.89] (LM SAE)	1.36

The LM columns come from a language-locus dictionary trained per policy (k=256, 15k steps) and screened with the same rule on the prompt-side hook (20 most motion-aligned LM features). These dictionaries reproduce the anatomy contrast at the dictionary level as well: 5 of 16,892 active LM features are behavior-aligned on the $\pi_{0.5}$ folding policy against 58 of 5,472 EX features (35.8×, 95% CI [17.0, 170]), and 8 of 17,700 against 49 of 5,627 on the $\pi_{0.5}$ flipping policy (19.3×, [10.1, 53.5]). Both policies reproduce the locus effect for raw residual dimensions (13.4× and 1.9× more behavior-aligned EX than LM dimensions), and both dictionaries recover arm-motion and joint-speed families; the folding dictionary additionally has a left-grasp family and the flipping dictionary right-grasp and right-release families, matching the arm each task uses. On the $\pi_{0.5}$ folding policy the population screen gives the cleanest handle result of the paper: 29 of the 30 most behavior-aligned features move the arm they are associated with 3.1× more than the other arm and grippers (matched-energy random directions 1.4×, random active features 1.0×), whereas the 30 most aligned raw residual dimensions do not lateralize at all (0.93×, 20% passing; pass-rate ratio 4.8 [2.7, 14.5]); three dictionaries re-trained from other seeds reproduce it with 30 of 30 aligned features passing each (Table 17). Random dense directions with a matched energy budget pass more often than aligned neurons here because they are scored on their best channel, an upper bound that favors the control. On the $\pi_{0.5}$ flipping policy the aligned expert features again beat the dense controls (2.3× median; 84% of the 25 arm-motion features pass, and `release-package-right` (13100) reaches 5.1×), but the aligned residual dimensions also steer their arm on this single-arm task (1.1×, 70% passing), so the pass-rate gap is not significant there; as on the $\pi_{0.5}$ LIBERO policy, gripper-targeted units are the hard cases (2 of 5 aligned features, 0 of 3 random ones). Across the $\pi_{0.5}$ breakfast, folding, and flipping policies, the median specificity of the aligned expert features is 2.0, 3.1, and 2.3, against 1.0, 0.9, and 1.1 for the most aligned raw dimensions.

The language locus is not universally inert. On the $\pi_{0.5}$ breakfast policy the 20 most aligned LM features sit at 0.9× (35% passing) and on the $\pi_{0.5}$ folding policy at 1.2× (65%; expert features 1.5× [1.1, 2.2] more likely to pass), but on the $\pi_{0.5}$ flipping policy (single instruction; the left arm moves in 43% of frames while the right arm manipulates the package) the 20 most motion-aligned LM features move the arm they predict 3.9× more than the other channels and pass at 85%, exceeding the expert features. The language prefix of $\pi_{0.5}$ never sees proprioception, but it does see the cameras, and in a task with a single fixed instruction its last attended token evidently carries the visually observable phase of the left arm. Together with the OpenVLA-OFT result (Appendix G), this bounds the locus claim: the action expert is where motor-aligned, steerable features concentrate in every policy we tested, while the language locus is inert when its content is dominated by task description (multi-

subtask prompts) and can carry handles when it is not. The causal status of a feature is decided by the intervention test, not by its locus.

Human verification on the folding and flipping policies. The interpretation pipeline of Appendix C was run unchanged on both policies (40 random active SAE features, 40 EX neurons, and the 40 leading EX principal components per policy), and the same blinded rater answered every held-out item (1,443 decisions per policy; Table 5). On the $\pi_{0.5}$ flipping policy the human favors SAE names, as on the $\pi_{0.5}$ breakfast policy (top-K 0.56 against 0.51 for neurons and PCA; pass rate 32.5% against 22.5% and 27.5%). On the $\pi_{0.5}$ folding policy the ordering reverses for forced choice (0.64 for SAE against 0.71 and 0.72) and is flat for top-K precision (0.61, 0.57, 0.55), and neuron and PCA names pass more often (45% each against 25%). The folding policy's dictionary was trained on 13k frames and keeps 5,472 active features, so a random active feature is far less likely to be one of its 58 aligned features than on the $\pi_{0.5}$ breakfast policy. The advantage of the dictionaries therefore lies in the aligned population and the handle screen, not in the nameability of an arbitrary active feature.

I Additional Steering Results

Table 14: Which units are handles. Open-loop screen at $|\alpha|{=}16\sigma$ on held-out frames. Pass: share of the 30 most aligned EX features with $R_{\text{spec}} > 1$ and sign reversal on their own predicted channel, for the reference dictionary and (range) for three dictionaries re-trained from other seeds on the same data and screened on the same frames. Median R_{spec} per predeclared unit set (Sec. 3.4); bold: highest per policy. †Prompt-token dictionary at layer 16: layer 31 is the last layer, where prompt positions no longer reach the action tokens, and a screen there leaves every action unchanged. Full statistics: Appendices G, H, I.

	Aligned EX SAE			Median R_{spec} of the other unit sets				
Policy	Pass (%)	Pass, seeds 1–3 (%)	Median R_{spec}	Random EX	Aligned dims	Random dirs	PCA	Aligned LM
Agilex bimanual robot (14-D actions)								
$\pi_{0.5}$ breakfast	63	17–53	**2.0**	0.9	1.0	1.1	1.3	0.9
$\pi_{0.5}$ folding	97	100	**3.1**	1.0	0.9	1.4	0.9	1.2
$\pi_{0.5}$ flipping	77	77–83	2.3	1.1	1.1	1.0	1.4	**3.9**
LIBERO single arm (7-D actions)								
$\pi_{0.5}$ LIBERO	33	30–43	0.8	0.8	1.1	1.2	**1.9**	1.1
OpenVLA-OFT LIBERO (L=31)	37	50–70	0.7	0.8	**3.4**	1.2	0.8	0.6†

Table 14 summarizes the population screens below. On the $\pi_{0.5}$ LIBERO policy, 24 of the 30 most aligned expert features predict the near-binary gripper channel, which no unit type steers in open loop.

Feature selection uses only held-out interpretation and open-loop calibration data. For the $\pi_{0.5}$ breakfast policy, the frozen intervention set contains `grasp-bottle-left` (3659), `tilt-wrist-pour-left` (8176), and `release-bottle-left` (8988), selected for left grasp, left hand height, and left wrist tilt (EX L=8, k=128 dictionary, SHA-256 prefix `240c237dba01bddd`), with the inert language-locus control `grasp-bottle-left` (7779) (LM L=11, k=256, `1b0f822918e608ef`). For the $\pi_{0.5}$ LIBERO policy, it contains `grasp-object-closing` (4270), `approach-stove-top` (9168), and `open-gripper-command` (3701), selected for grasp, descent, and release (EX L=8, k=128, `639a333b2779f20b`), with three random active controls. Open-loop dose responses on both policies are monotone and sign-symmetric over $\alpha \in \{\pm 4, \pm 8, \pm 16, 32\}\sigma_j$ (Spearman $\rho = 1.0$ for every motor feature). On the $\pi_{0.5}$ LIBERO policy only the descent row is selective in open loop ($R_{\text{spec}} = 1.54$ on the height command), whereas the grasp and release rows and all controls stay below one on the near-binary gripper channel. The open-loop dose was frozen from this grid before any closed-loop outcome ($\alpha_{\text{ol}} = 32\sigma_j$ for the $\pi_{0.5}$ LIBERO policy, the largest dose at which every screened feature stays monotone, sign-symmetric, and near-linear), and the closed-loop gate then selected $\alpha_{\text{cl}} = 8\sigma_j$ (Table 18). For every concept we report all tested doses, target and collateral channels, sign consistency, and closed-loop trials, not only the best dose.

Table 15: Open-loop intervention on the $\pi_{0.5}$ breakfast policy at $|\alpha|{=}16\sigma_j$ on the same 1,016 frames. Targets: left-gripper channel for `grasp-bottle-left` (3659) and all controls, the six left-arm joints for `tilt-wrist-pour-left` (8176), and left joints 3–5 for `release-bottle-left` (8988). SAE features are compared with SAE-in-the-loop; dense directions are compared with their matched no-intervention baseline. The sign of a random control's effect on the gripper carries no meaning; pooled over the three units of each kind, R_{spec} is $0.33\times$ for random active features and $1.1\times$ for matched random directions. One of the three matched random directions moves the left gripper as selectively as `grasp-bottle-left` (3659), which is why selectivity is assessed on predeclared unit populations (Table 16) rather than on single directions. The signed columns give the mean target change at $-16\sigma_j$ and $+16\sigma_j$. Arrows indicate the preferred direction.

Row or control	Target effect $\lvert\Delta a\rvert\uparrow$	Collateral effect $\lvert\Delta a\rvert\downarrow$	Specificity $R_{\text{spec}}\uparrow$	Signed target at -16σ	Signed target at $+16\sigma$
`grasp-bottle-left` (3659)	0.0049	0.0021	2.3×	+0.0031	−0.0031
`tilt-wrist-pour-left` (8176)	0.0162	0.0087	1.9×	+0.0058	−0.0054
`release-bottle-left` (8988)	0.0016	0.0015	1.05×	+0.0001	−0.0001
`retract-right-arm-up` (7493), random	0.0067	0.0236	0.28×	+0.0069	−0.0064
`hold-bottle-left` (10099), random	0.0008	0.0010	0.82×	+0.0002	−0.0002
`grasp-bottle-left` (13220), random	0.0010	0.0014	0.73×	+0.0004	−0.0004
Matched random direction 1	0.0013	0.0028	0.48×	−0.0006	+0.0006
Matched random direction 2	0.0024	0.0029	0.82×	−0.0021	+0.0020
Matched random direction 3	0.0052	0.0023	2.3×	−0.0051	+0.0051
PCA PC1	0.014	0.013	1.1×	+0.0140	−0.0140
`grasp-bottle-left` (7779), LM	0.00050	0.00052	0.95×	0.0000	0.0000

Table 16: Population open-loop causal screen (Fig. 15): pass rate (specificity > 1 and, for scalar targets, sign reversal between $\pm16\sigma$ with both bootstrap CIs excluding zero), median specificity (bootstrap 95% CI), and the fraction of scalar-target units whose $+\alpha$ response has the sign predicted by their association (n/a: no scalar-target unit in the set, as for the aligned neurons of the $\pi_{0.5}$ breakfast policy, which all target a multi-joint arm group, or no predicted sign, as for random directions). Gripper-targeted subsets are listed where they contain at least four units.

Policy	Unit set	n	Pass rate ↑	Median $R_{\text{spec}}\uparrow$	Sign-consistent ↑
$\pi_{0.5}$ breakfast	EX SAE, behavior-aligned	30	63%	1.99 [0.86, 2.61]	80%
$\pi_{0.5}$ breakfast	gripper-targeted subset	5	100%	3.07	80%
$\pi_{0.5}$ breakfast	EX SAE, random active (unaligned)	30	47%	0.91 [0.69, 1.21]	75%
$\pi_{0.5}$ breakfast	Residual neuron, most aligned	30	50%	1.01 [0.93, 1.17]	n/a
$\pi_{0.5}$ breakfast	PCA component (1–8)	8	75%	1.25 [0.85, 2.70]	50%
$\pi_{0.5}$ breakfast	Random direction (best scalar target)	8	50%	1.07 [0.67, 1.82]	n/a
$\pi_{0.5}$ breakfast	LM SAE, most aligned	20	35%	0.89 [0.73, 1.35]	75%
$\pi_{0.5}$ breakfast	gripper-targeted subset	4	0%	0.43	75%
$\pi_{0.5}$ LIBERO	EX SAE, behavior-aligned	30	33%	0.79 [0.71, 0.97]	17%
$\pi_{0.5}$ LIBERO	gripper-targeted / height-targeted	24 / 5	29% / 60%	0.74 / 1.01	17% / 20%
$\pi_{0.5}$ LIBERO	EX SAE, random active (unaligned)	30	40%	0.79 [0.62, 1.08]	44%
$\pi_{0.5}$ LIBERO	gripper-targeted / height-targeted	14 / 3	7% / 100%	0.48 / 1.40	43% / 67%
$\pi_{0.5}$ LIBERO	Residual neuron, most aligned	30	73%	1.11 [1.10, 1.25]	23%
$\pi_{0.5}$ LIBERO	gripper-targeted / height-targeted	6 / 19	0% / 89%	0.34 / 1.25	33% / 16%
$\pi_{0.5}$ LIBERO	PCA component (1–8)	8	62%	1.94 [0.36, 3.23]	75%
$\pi_{0.5}$ LIBERO	Random direction (best scalar target)	8	62%	1.24 [0.91, 1.45]	n/a
$\pi_{0.5}$ LIBERO	LM SAE, most aligned	20	40%	1.06 [0.45, 1.25]	35%
$\pi_{0.5}$ LIBERO	gripper-targeted / continuous-channel	8 / 12	0% / 67%	0.42 / 1.25	63% / 11%

Seed replication of the population screen. A population result obtained with one dictionary could reflect that dictionary rather than the locus. For each of the $\pi_{0.5}$ breakfast, folding, and flipping policies we therefore trained three further EX dictionaries from different seeds with the identical recipe and data (held-out variance explained within 0.002 of the reference dictionary), re-derived the 30 most behavior-aligned and 30 random active features of each with the same frozen rule, and screened them on exactly the frames of Fig. 15; the seed-independent unit sets (raw dimensions, principal components, random directions) were not re-run. Unit sets were frozen before any screening outcome. Table 17 gives the result. On the $\pi_{0.5}$ folding policy every aligned feature of every re-seeded dictionary passes, with median specificity 3.6–4.0 against 1.0–1.1 for random active features, and the pass-rate ratio excludes one for all four dictionaries. On the $\pi_{0.5}$ flipping policy the aligned pass rate is 77–83% against 53–67% and the median specificity 1.8–2.3 against 1.0–1.1 for all four dictionaries; the interval of the aligned median specificity excludes the random median for all four, while the pass-rate ratio, a coarser statistic, excludes one for one of them. On the $\pi_{0.5}$

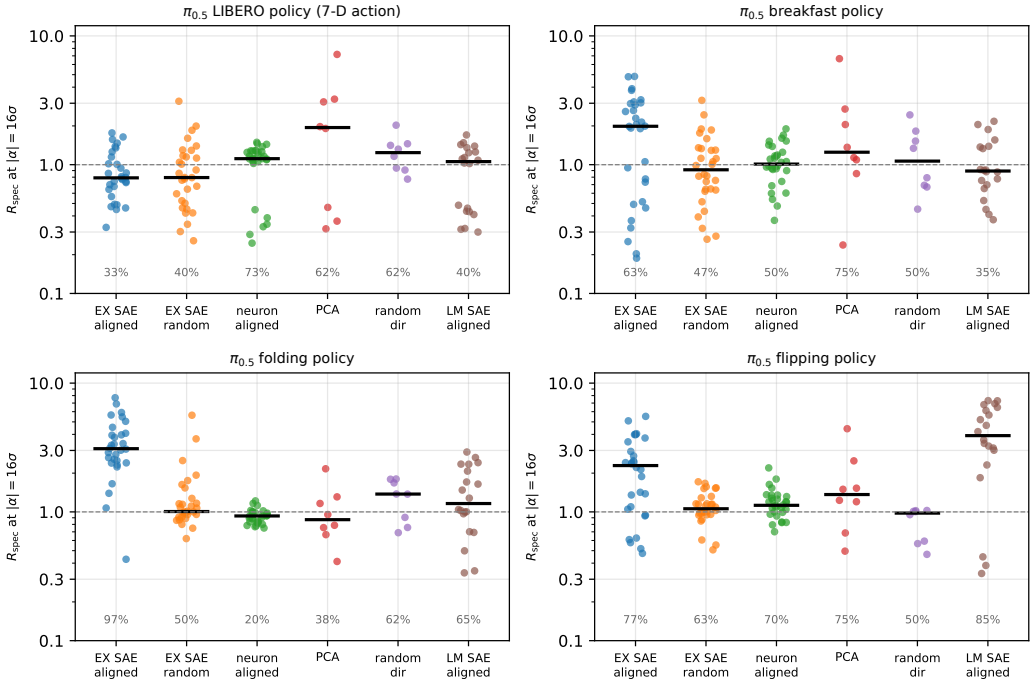

Figure 15: Population open-loop causal screen. Each point is one unit's specificity R_{spec} at $|\alpha|{=}16\sigma$ (mean of $\pm\alpha$), scored on the action channel predicted by that unit's own held-out association; bars mark group medians and the dashed line $R_{\text{spec}}{=}1$. Unit sets were frozen before any steering outcome was observed: the 30 most behavior-aligned EX SAE features, 30 random active but unaligned EX features, the 30 most aligned raw residual dimensions, the 8 leading principal components, 8 matched-energy random directions (scored on their best scalar channel), and the 20 most aligned LM features. Panels: $\pi_{0.5}$ LIBERO policy (384 held-out frames), $\pi_{0.5}$ breakfast policy (1,016), $\pi_{0.5}$ folding policy (538; LM features 306), and $\pi_{0.5}$ flipping policy (625; LM features 351; Appendix H).

breakfast policy the ordering of the reference dictionary (63% against 47%, ratio 1.36 [0.86, 2.3]) is not reproduced: the three re-seeded dictionaries pass 17–53% of aligned features against 37–60% of random ones, and pooled over the four dictionaries the two sets are indistinguishable ($52/120$ against $57/120$). Its dictionaries are also the sparsest of the three (≈ 0.5k active features on held-out frames against 5.5k and 5.6k), so their 30 most aligned features are drawn from a much smaller and more polysemantic pool. We therefore report the pass rate of the breakfast policy in Table 14 as a single-dictionary estimate and state the population claim for the $\pi_{0.5}$ folding and flipping policies, where it holds for every dictionary we trained. The same replication on the two 7-D LIBERO policies reproduces the absence of an aligned-over-random advantage there: aligned features pass at 30–43% against 27–37% for random ones on the $\pi_{0.5}$ LIBERO policy, and at 50–70% against 43–57% on the OpenVLA-OFT LIBERO policy, with every pass-rate interval including one. The absolute pass rate follows how many aligned features predict the saturated gripper channel: 26 of 30 for the reference OpenVLA-OFT dictionary (37% passing) against 14–19 for its re-seeded dictionaries, whose other aligned features mostly target the steerable height channel.

Table 19 gives the behavioral events. Under SAE-in-loop, the descend onset is at step ≈ 23, the first gripper closure at step ≈ 50, the minimum height ≈ 661 mm, and the lateral path ≈ 0.8 m per episode. At α_{ol}, `approach-stove-top` (9168) descends 5.2 steps early with a 97 mm shorter lateral path and a 24.5 *pp* success loss, so the open-loop dose buys a larger target effect at the price of collateral motion and task failure. The schedule variants trace the same dose–effect curve. Applying the intervention only at the last three denoising steps, or only when the feature is already active, reduces the commanded-height effect roughly in proportion to the exposure (-0.0069 and -0.0057 at $+8$, against -0.0135 at every step) while leaving success within noise, and $+16$ on late steps

Table 17: Seed replication of the population open-loop screen (same frames and pass rule as Table 16; 30 aligned and 30 random active EX features per dictionary). Median R_{spec} with bootstrap 95% CI; ratio: aligned over random pass rate with bootstrap 95% CI. Seed 0 is the reference dictionary of the main text. OpenVLA-OFT LIBERO policy: action-token dictionaries at layer 31, screened on the 145 frames of its reference screen.

		Aligned EX SAE		Random active EX SAE		
Policy	Dictionary	Pass	Median R_{spec}	Pass	Median R_{spec}	Pass-rate ratio
$\pi_{0.5}$ breakfast	seed 0	63%	1.99 [0.86, 2.61]	47%	0.91 [0.69, 1.21]	1.36 [0.86, 2.3]
$\pi_{0.5}$ breakfast	seed 1	40%	0.61 [0.46, 1.20]	37%	0.89 [0.68, 1.06]	1.09 [0.56, 2.2]
$\pi_{0.5}$ breakfast	seed 2	17%	0.62 [0.52, 0.82]	47%	0.96 [0.79, 1.27]	0.36 [0.09, 0.78]
$\pi_{0.5}$ breakfast	seed 3	53%	1.16 [0.69, 1.68]	60%	1.14 [0.94, 1.32]	0.89 [0.56, 1.4]
$\pi_{0.5}$ folding	seed 0	97%	3.10 [2.60, 3.86]	50%	1.01 [0.96, 1.15]	1.93 [1.4, 3.0]
$\pi_{0.5}$ folding	seed 1	100%	4.00 [2.69, 4.61]	60%	1.08 [0.95, 1.19]	1.67 [1.3, 2.3]
$\pi_{0.5}$ folding	seed 2	100%	3.65 [2.76, 4.14]	43%	0.96 [0.86, 1.08]	2.31 [1.7, 3.8]
$\pi_{0.5}$ folding	seed 3	100%	3.81 [3.22, 4.41]	50%	1.00 [0.89, 1.08]	2.00 [1.5, 3.0]
$\pi_{0.5}$ flipping	seed 0	77%	2.29 [1.37, 2.61]	63%	1.06 [0.99, 1.15]	1.21 [0.87, 1.75]
$\pi_{0.5}$ flipping	seed 1	77%	1.85 [1.36, 2.19]	67%	1.07 [1.02, 1.17]	1.15 [0.83, 1.62]
$\pi_{0.5}$ flipping	seed 2	83%	1.83 [1.59, 2.96]	63%	1.07 [0.98, 1.17]	1.32 [0.96, 1.87]
$\pi_{0.5}$ flipping	seed 3	80%	2.14 [1.83, 3.06]	53%	1.02 [0.96, 1.15]	1.50 [1.05, 2.33]
$\pi_{0.5}$ LIBERO	seed 0	33%	0.79 [0.71, 0.97]	40%	0.79 [0.62, 1.08]	0.83 [0.39, 1.67]
$\pi_{0.5}$ LIBERO	seed 1	43%	0.98 [0.79, 1.30]	37%	0.95 [0.82, 1.06]	1.18 [0.63, 2.38]
$\pi_{0.5}$ LIBERO	seed 2	30%	0.85 [0.73, 1.00]	37%	0.92 [0.56, 1.07]	0.82 [0.36, 1.75]
$\pi_{0.5}$ LIBERO	seed 3	43%	0.92 [0.73, 1.24]	27%	0.83 [0.71, 0.91]	1.62 [0.80, 4.00]
OpenVLA-OFT LIBERO (L=31)	seed 0	37%	0.70 [0.50, 1.12]	40%	0.78 [0.60, 1.18]	0.92 [0.46, 1.83]
OpenVLA-OFT LIBERO (L=31)	seed 1	50%	1.36 [0.73, 3.00]	43%	0.86 [0.70, 1.24]	1.15 [0.67, 2.11]
OpenVLA-OFT LIBERO (L=31)	seed 2	70%	2.54 [1.04, 3.89]	53%	1.08 [0.82, 1.69]	1.31 [0.87, 2.09]
OpenVLA-OFT LIBERO (L=31)	seed 3	57%	1.58 [0.77, 2.53]	57%	1.17 [0.73, 1.81]	1.00 [0.62, 1.58]

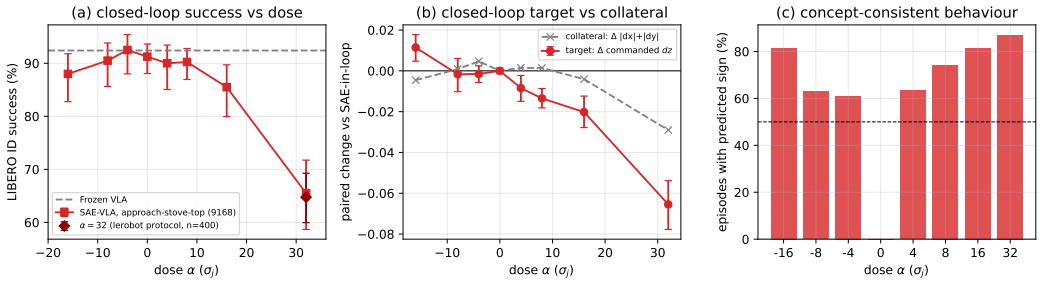

Figure 16: Closed-loop dose response of `approach-stove-top` (9168) on the $\pi_{0.5}$ LIBERO policy. (a) Success vs. dose on the paired harness; diamond: official-protocol point at $\alpha_{\text{ol}} = 32\sigma_j$. (b) Paired change in commanded height vs. collateral lateral command (bootstrap 95% CIs). (c) Fraction of episodes whose commanded-height change has the predicted sign.

reproduces the full-schedule $+8$ effect (-0.0141) at similar success (89.0% against 90.2%). The negative branch is sign-symmetric at $|\alpha|$=16 ($+0.0114$ against -0.0202) but not at $|\alpha| \leq 8$, where a rarely active feature has little to suppress. No schedule improves faithfulness: SAE-in-loop on late steps only matches the every-step configuration (90.0% on the same 200 episodes).

The descent row is the strongest handle of the $\pi_{0.5}$ LIBERO policy. A closed-loop test of the eight other features with a continuous target that passed the open-loop screen (three aligned, five random; n=200 each) gives weaker effects: at α_{cl} one of eight shifts its commanded channel significantly and six move it with the predicted sign; at $16\sigma_j$ three are significant and seven have the predicted sign, and one feature costs 6 *pp* of success.

Replication with a second dictionary and a second backbone. Table 20 reports every closed-loop dose of the two replications summarized in Table 3, all predeclared and run on the paired harness (4 suites, 5 initial states per task, n=200 per cell; the harness of the OpenVLA-OFT LIBERO policy uses the same suites and initial states with the policy's own action-chunk controller).

Table 18: Closed-loop dose sweep for `approach-stove-top` (9168) of the $\pi_{0.5}$ LIBERO policy on the paired BDDL harness (4 suites; 10 init states per task, n=400, for the 0 and +8 rows, 5 init states, n=200, otherwise). Δ values are paired against SAE-in-loop (α=0) on the shared episodes. Commanded Δz is the mean first action-chunk z command; concept-consistent behavior is the fraction of episodes whose Δ commanded z has the sign predicted by the open-loop screen (+ dose $\Rightarrow$ more negative dz). The lower block varies *when* the intervention is applied at the same doses: only at the last three of ten denoising steps (late), or only on policy calls where the feature is already active (gated). For the two rows without steering, the concept-consistent column gives the share of episodes with negative Δz, the value the metric takes without an intervention (its complement for negative doses). On the official protocol (n=400, no action telemetry), $+\alpha_{\mathrm{ol}}$ gives 64.8% ($-26.8\,pp$ vs. Frozen VLA, $[-31.7, -21.8]$, $p < 0.001$).

Dose α (σ_j)	SR (%)	ΔSR vs. SAE-in-loop	Δ commanded z	$\Delta\lvert xy\rvert$ collateral	Concept-consistent (%)	McNemar p
Frozen VLA (n=395)	92.4	+1.3 [−1.5, +4.1]	−0.0044 [−0.009, +0.001]	+0.0024 [−0.001, +0.006]	54.2	0.47
0 (SAE-in-loop, n=400)	91.2		reference condition of the paired columns			
+4	90.0	0.0 [−4.5, +4.5]	−0.0085 [−0.015, −0.002]	+0.0014 [−0.003, +0.006]	63.5	1.00
+8 (α_{cl}, n=400)	90.2	−1.0 [−4.2, +2.2]	−0.0135 [−0.018, −0.009]	+0.0014 [−0.002, +0.005]	74.3	0.66
+16	85.5	−4.5 [−10.0, +1.0]	−0.0202 [−0.028, −0.012]	−0.0041 [−0.010, +0.002]	81.5	0.15
−4	92.5	+2.5 [−1.0, +6.5]	−0.0015 [−0.006, +0.002]	+0.0047 [+0.001, +0.009]	61.0	0.30
−8	90.5	+0.5 [−3.5, +4.5]	−0.0017 [−0.010, +0.006]	+0.0009 [−0.004, +0.006]	63.0	1.00
−16	88.0	−2.0 [−7.5, +3.5]	+0.0114 [+0.005, +0.018]	−0.0046 [−0.010, +0.001]	81.5	0.58
+32 (α_{ol}, harness, n=200)	65.5	−24.5 [−31.5, −17.5]	−0.0655 [−0.078, −0.054]	−0.0290 [−0.037, −0.022]	87.0	< 0.001
0, late steps only	90.0	0.0 [−4.0, +4.0]	−0.0009 [−0.008, +0.006]	+0.0014 [−0.003, +0.006]	50.0	1.00
+8, late steps only	91.5	+1.5 [−3.5, +6.0]	−0.0069 [−0.013, −0.001]	+0.0022 [−0.002, +0.007]	67.5	0.68
+16, late steps only	89.0	−1.0 [−6.0, +4.0]	−0.0141 [−0.021, −0.007]	+0.0023 [−0.003, +0.008]	76.5	0.85
+8, gated on feature activity	92.0	+2.0 [−2.0, +6.0]	−0.0057 [−0.013, +0.001]	+0.0044 [+0.000, +0.009]	65.5	0.45

Table 19: Closed-loop behavioral events of the $\pi_{0.5}$ LIBERO policy under steering of `approach-stove-top` (9168) (same episodes as Table 18; paired against SAE-in-loop, bootstrap 95% CIs). Descend onset is the first control step at which the end effector is 3 cm below its initial height; grasp onset is the first step with a closing gripper command; minimum height is the lowest end-effector z in the episode; lateral path is the summed xy displacement (collateral). One control step is 50 ms.

Dose α (σ_j)	n	Δ descend onset (steps) ↓ for $+\alpha$	Δ grasp onset (steps)	Δ minimum height (mm)	Δ lateral path (mm)
−16	200	+2.80 [+2.45, +3.19]	+2.69 [0.00, +4.56]	+2.76 [+0.96, +4.82]	+9 [−18, +35]
−8	200	+1.38 [+1.17, +1.60]	+2.08 [+0.91, +3.89]	+2.29 [+0.59, +4.32]	−5 [−36, +26]
−4	200	+0.67 [+0.53, +0.82]	+1.01 [+0.30, +1.92]	+0.78 [−0.11, +1.89]	+3 [−20, +26]
+4	200	−0.60 [−0.72, −0.49]	−0.41 [−0.96, +0.31]	+0.19 [−0.78, +1.36]	−38 [−67, −11]
+8 (α_{cl})	400	−1.32 [−1.48, −1.17]	−1.47 [−1.91, −1.11]	−2.17 [−3.53, −0.90]	−27 [−44, −10]
+16	200	−2.47 [−2.81, −2.17]	−2.83 [−4.29, −1.82]	−4.63 [−7.03, −2.33]	−34 [−62, −7]
+32 (α_{ol})	200	−5.19 [−5.76, −4.68]	−2.72 [−6.19, +0.91]	−6.40 [−10.15, −2.60]	−97 [−136, −59]
+8, late steps only	200	−0.99 [−1.19, −0.81]	−0.94 [−1.73, +0.25]	−1.02 [−2.73, +0.76]	−24 [−50, −1]
+16, late steps only	200	−2.00 [−2.29, −1.75]	−1.44 [−2.19, −0.61]	−2.17 [−4.41, +0.07]	−37 [−68, −4]
+8, gated on activity	200	−0.37 [−0.48, −0.26]	+0.01 [−1.03, +1.75]	−0.48 [−1.73, +0.83]	−34 [−60, −10]
Frozen VLA (reference)	395	−0.07 [−0.14, 0.00]	−0.10 [−0.51, +0.37]	+0.19 [−0.56, +1.02]	−19 [−35, −4]

Second dictionary of the $\pi_{0.5}$ LIBERO policy. An EX L=8, k=128 dictionary trained with seed 1 under the identical recipe, with the frozen lift rule of the first dictionary (largest $|r|$ with the vertical end-effector velocity among active features), selects `approach-bowl-open` (868), with $r_{v_z} = -0.49$ and $\sigma_j = 1.92$ versus 3.10 for `approach-stove-top` (9168). In open loop it is monotone and sign-symmetric on dz with $R_{\mathrm{spec}} = 1.14$ at $16\sigma_j$ against 1.54 for `approach-stove-top` (9168), the seed-1 grasp row `grasp-object-closing` (15948) reaches 1.8 on the gripper channel, and the three random active controls stay at 0.3–0.7. In closed loop the seed-1 dictionary is faithful (90.5% against 93.5% frozen; $-3.0\,pp$ $[-7.0, +1.0]$) and every grid dose passes the gate, so $\alpha_{\mathrm{cl}} = 16\sigma_j$ for this feature. The descend onset shifts by -0.16, -0.51, and -1.00 steps at +4, +8, and $+16\sigma_j$ and by +1.12 at $-16\sigma_j$, with every interval excluding zero; per unit of injected activation ($\alpha\sigma_j$) this is -0.033 steps against -0.050 for `approach-stove-top` (9168). Gripper timing and minimum height do not change beyond noise, and the commanded-height change is significant only at $+16\sigma_j$ (-0.0123 $[-0.0184, -0.0067]$). The second seed therefore reproduces the direction, monotonicity, sign symmetry, and safety of the handle at a smaller effect size, as expected for a feature with a smaller scale and weaker alignment.

OpenVLA-OFT LIBERO policy. The layer-31 action-token dictionary (Appendix G) is faithful in closed loop (98.0% against 96.5% frozen; $+1.5\,pp$ $[-1.5, +4.5]$). The height-targeted aligned feature with the highest open-loop specificity, `grasp-object` (7165) (10.4×; $\sigma_j =$

5.34), was frozen as the steering row and `transport-object-closed-gripper` (7962), `grasp-object` (5066), and `approach-object-open-gripper` (3501) ($\sigma_j = 3.54$, 8.04, 8.47; predicted signs $+$, $+$, $-$ on dz) as its population set. The dose grid $\{\pm4, \pm8, \pm16\}\sigma_j$ was extended by ±2 after $\pm8\sigma_j$ failed the gate, and the population set was run at the resulting $\alpha_{\text{cl}} = 4\sigma_j$ in addition to the predeclared $8\sigma_j$. Across all eight doses `grasp-object` (7165) is monotone and sign-symmetric on every event metric; the gate passes at $+2$, -2, and $+4\sigma_j$ and fails at $-4\sigma_j$ by $1.5\,pp$, so the positive direction is usable at $4\sigma_j$ and the negative at $2\sigma_j$. The other three height features move the commanded height and the descend onset with the predicted sign at both doses (all 12 intervals exclude zero), but only `transport-object-closed-gripper` (7962) stays within the margin at $4\sigma_j$; `grasp-object` (5066) and `approach-object-open-gripper` (3501) have the largest σ_j, so the same σ-dose is a larger absolute perturbation, and their tolerable dose lies below the grid. Direction is thus predictable from the screen, while the dose has to be gated per feature.

Table 20: Closed-loop replication on the paired harness: `approach-bowl-open` (868) from the second dictionary of the $\pi_{0.5}$ LIBERO policy, and layer-31 action-token features of the OpenVLA-OFT LIBERO policy. Paired against the respective SAE-in-loop condition (n=200 per row); bootstrap 95% CIs. Commands are in each policy's normalized action space ($\times10^{-3}$); consistent: episodes whose commanded-height change has the sign predicted by the feature's association (negative for `approach-object-open-gripper` (3501)).

Row	Dose (σ_j)	ΔSR (pp)	Δ commanded z	Δ$\lvert xy\rvert$ collateral	Δ descend onset (steps)	Δ min. height (mm)	Consistent (%)
$\pi_{0.5}$ LIBERO policy, EX dictionary seed 1 (SAE-in-loop 90.5%; frozen 93.5%)							
`approach-bowl-open` (868)	+4	**+2.0** [−2.0, +6.0]	**−4.4** [−9.3, +0.1]	**+1.2** [−2.4, +5.3]	**−0.16** [−0.27, −0.04]	**+1.4** [+0.2, +2.8]	56
	+8	**−1.0** [−5.5, +3.0]	**−3.1** [−8.0, +1.7]	**−2.2** [−6.8, +2.2]	**−0.51** [−0.62, −0.39]	**+0.7** [−0.4, +1.9]	59
	+16 (α_{cl})	**−2.0** [−6.5, +2.5]	**−12.3** [−18.4, −6.7]	**−3.0** [−7.8, +1.9]	**−1.00** [−1.19, −0.84]	**+0.4** [−0.8, +1.7]	68
	−16	**−1.0** [−5.0, +3.0]	**+1.9** [−4.7, +8.2]	**−0.6** [−4.2, +3.4]	**+1.12** [+0.97, +1.28]	**+1.4** [+0.2, +2.7]	60
OpenVLA-OFT LIBERO policy, action-token dictionary L=31 (SAE-in-loop 98.0%; frozen 96.5%)							
`grasp-object` (7165)	−16	**−69.0** [−75.5, −62.5]	**−142.4** [−164.4, −121.5]	**−86.3** [−96.6, −75.9]	**−13.5** [−14.7, −12.3]	**−12.9** [−18.6, −6.8]	92
	−8	**−18.5** [−24.5, −13.0]	**−41.7** [−51.0, −32.8]	**−16.8** [−25.2, −8.6]	**−7.9** [−8.8, −7.0]	**−9.6** [−13.2, −6.4]	88
	−4	**−4.5** [−8.0, −1.5]	**−16.0** [−20.4, −11.5]	**+2.5** [−1.8, +6.8]	**−4.2** [−4.7, −3.8]	**−2.8** [−4.7, −1.3]	88
	−2	**−2.5** [−5.5, +0.5]	**−10.9** [−17.0, −6.4]	**+1.7** [−2.0, +5.1]	**−2.2** [−2.5, −1.9]	**−0.8** [−1.8, +0.3]	76
	+2	**−2.5** [−5.0, +0.0]	**+12.0** [+9.3, +15.0]	**−8.6** [−11.7, −5.8]	**+2.2** [+2.0, +2.5]	**+0.8** [−0.3, +1.8]	83
	+4 (α_{cl})	**−1.5** [−4.5, +1.5]	**+17.3** [+13.7, +20.6]	**−10.2** [−13.9, −6.3]	**+4.6** [+4.1, +5.2]	**+3.0** [+1.9, +4.1]	88
	+8	**−20.0** [−26.0, −14.0]	**+40.7** [+34.0, +47.4]	**−36.0** [−42.3, −30.1]	**+10.2** [+9.1, +11.4]	**+9.7** [+7.1, +12.6]	95
	+16	**−78.5** [−84.0, −72.5]	**+130.1** [+114.9, +146.2]	**−104.0** [−112.7, −95.8]	**+23.5** [+21.1, +26.2]	**+44.9** [+36.9, +53.5]	94
`transport-object-closed-gripper` (7962)	+4	**−1.0** [−3.0, +1.0]	**+9.4** [+6.9, +12.0]	**−4.6** [−7.7, −1.8]	**+1.5** [+1.3, +1.8]	**+1.0** [+0.2, +1.8]	78
	+8	**−5.0** [−9.0, −1.5]	**+18.2** [+14.2, +22.3]	**−13.6** [−18.6, −9.2]	**+3.1** [+2.7, +3.6]	**+4.2** [+2.8, +5.6]	86
`grasp-object` (5066)	+4	**−23.5** [−30.0, −17.5]	**+29.6** [+23.7, +35.2]	**−37.7** [−44.4, −31.2]	**+5.8** [+5.2, +6.6]	**+7.6** [+5.1, +10.2]	87
	+8	**−80.0** [−85.5, −74.5]	**+64.3** [+48.5, +79.1]	**−101.1** [−109.0, −93.2]	**+15.9** [+13.4, +18.7]	**+30.5** [+23.5, +37.7]	85
`approach-object-open-gripper` (3501)	+4	**−30.5** [−37.0, −24.0]	**−28.2** [−37.8, −18.9]	**−37.0** [−45.0, −29.3]	**−5.7** [−6.5, −4.7]	**−6.6** [−10.3, −2.9]	79
	+8	**−84.0** [−89.0, −79.0]	**−76.1** [−98.8, −55.4]	**−92.6** [−101.7, −83.5]	**−10.8** [−12.0, −9.7]	**−5.0** [−12.4, +3.1]	70

J TASK UTILITY UNDER DISTRIBUTION SHIFT

This appendix reports what a validated handle does to task success when it is held at a constant dose under distribution shift. The question is deliberately harder than the one the main text answers: a handle is a lever on one motor variable, whereas a perturbed scene fails for many different reasons, so a constant bias can only help where the perturbation happens to call for the motion the handle produces. We test the handle `approach-stove-top` (9168) of the $\pi_{0.5}$ LIBERO policy at both doses and on two shift benchmarks with paired initial states; the feature, sign, and dose were frozen before any test outcome existed.

What the experiments show. On LIBERO-plus, six perturbation families with a 79% success floor leave room to move, and SAE-in-loop stays within noise of the frozen policy ($-1.1\,pp$, n=720), so the interface itself transfers across shifts. The gated dose $+\alpha_{\text{cl}}$ shifts success by $-2.9\,pp$, with no family improving (Table 21). On LIBERO-X, where the zero-shot success is only 6.4%, the gated dose gives $-2.0\,pp$ (n=450) and the open-loop dose α_{ol} gives $-2.9\,pp$ (Fig. 17). Held in distribution for entire rollouts, α_{ol} costs $26.8\,pp$ against the frozen policy on the official protocol, whereas α_{cl} costs nothing; this contrast motivates the closed-loop gate. Across all perturbation families and levels, the per-condition effects are small and homogeneous: no family benefits and none collapses at the gated dose, so the handle acts on one motor variable rather than perturbing the policy globally.

What follows for the method. These results locate the boundary between causal control and task utility. SAE-VLA establishes which features move which motor variable, with what specificity, and at what dose the policy tolerates the change; it does not by itself decide *when* that change is wanted. Because the interface is faithful under shift and the handles are graded and sign-symmetric,

the natural next step is a state-dependent controller that engages a handle only when the episode's own signals, such as a detected height or grasp error, indicate that its motion is the missing one. The constant-dose results here are the baseline such a controller must beat.

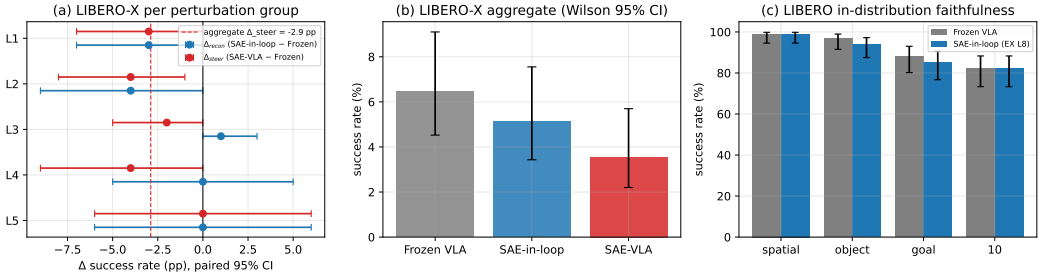

Figure 17: Task-level effect of the predeclared open-loop-dose intervention under distribution shift. (a) Per-group paired changes on LIBERO-X (Δ_{recon}: SAE-in-loop minus Frozen VLA; Δ_{steer}: SAE-VLA minus Frozen VLA; paired bootstrap 95% CIs; dashed line: aggregate Δ_{steer}). (b) Aggregate LIBERO-X success of the three systems over $n = 450$ paired episodes (Wilson 95% CIs). (c) In-distribution faithfulness of the $\pi_{0.5}$ LIBERO policy per suite ($n = 100$ each).

Table 21: LIBERO-plus: the closed-loop-safe intervention under six perturbation families. Each family perturbs 30 tasks per LIBERO suite (4 suites, one initial state; $n = 120$ paired episodes per family, 720 in total). SAE-VLA uses `approach-stove-top` (9168) at $+\alpha_{\text{cl}} = 8\sigma_j$; feature, sign, and dose were frozen from the in-distribution dose sweep before any LIBERO-plus outcome was observed. Δ columns are paired against the Frozen VLA (bootstrap 95% CI).

Perturbation family	n	Frozen VLA SR	SAE-in-loop SR	Δ_{recon}	SAE-VLA SR	Δ_{steer}
Camera viewpoint	120	50.8	52.5	$+1.7$ $[-5.8, +9.2]$	49.2	-1.7 $[-6.7, +3.3]$
Robot initial state	120	69.2	67.5	-1.7 $[-8.3, +5.0]$	65.0	-4.2 $[-12.5, +4.2]$
Language instruction	120	87.5	85.8	-1.7 $[-7.5, +4.2]$	84.2	-3.3 $[-9.2, +2.5]$
Lighting	120	90.0	86.7	-3.3 $[-10.0, +3.3]$	87.5	-2.5 $[-8.3, +3.3]$
Background texture	120	95.8	96.7	$+0.8$ $[-3.3, +5.0]$	94.2	-1.7 $[-6.7, +3.3]$
Object layout	120	83.3	80.8	-2.5 $[-6.7, +0.8]$	79.2	-4.2 $[-9.2, +0.0]$
Aggregate	720	79.4	78.3	-1.1 $[-3.6, +1.3]$, $p = 0.43$	76.5	-2.9 $[-5.4, -0.4]$, $p = 0.026$

K AGILEX CLOSED-LOOP PROTOCOL

Hardware and control. We use an Agilex Cobot Magic dual-arm manipulator and three calibrated cameras (top, left wrist, right wrist). The $\pi_{0.5}$ breakfast policy is the PyTorch conversion of its JAX fine-tune and runs on one RTX 4090. Each call uses 5 denoising steps and emits a 50-step action chunk. The controller runs at 30 Hz, the training-data rate, and executes commands at 30% speed. Real-time chunking conditions the next chunk on a 5-step prefix and executes 10 steps before the following call; at most 25 steps of a chunk are used open-loop. Inference of the frozen policy takes 215–231 ms per call on this machine, inside the 10-step execution window (333 ms at 30 Hz). The SAE encodes and decodes the 1024-dimensional action-expert residual at the final denoising step only, leaving the language prefix unchanged; an intervention is ramped over 15 steps at the chunk seam, and successive chunks are blended with a 5-step window. The toaster is unpowered and pouring uses room-temperature colored water. Persistent oscillation, arm–arm contact, arm–fixture contact, or an e-stop terminates a trial and counts as failure.

Reset and pairing. Both arms return to a recorded home configuration. Objects are placed on marked poses within 2 cm, and paired conditions share the same starting photograph and layout identifier. Condition order is randomized. Operators cannot alter α_{cl}, timeouts, or success criteria after inspecting outcomes.

Success criteria. For *load bread*, at least one slice must remain in the toaster slot after release. For *activate toaster*, the button must travel to its latched state. For *pour drink*, the cup level must visibly rise without a spill outside the 5 cm desk ring. For *place bread*, the bread projection must lie

on the plate. Full-breakfast success requires all four steps in order. Timeouts are 45 s per subtask and 180 s for the full sequence.

Conditions and trial counts. Faithfulness compares Frozen VLA with SAE-in-loop using 20 trials per subtask. Steering compares $\alpha = 0$, $+\alpha_{\mathrm{cl}}$, and $-\alpha_{\mathrm{cl}}$ with 12 paired trials per cell for `grasp-bottle-left` (3659), `tilt-wrist-pour-left` (8176), and `release-bottle-left` (8988); the negative controls, the random active expert row `hold-bottle-left` (10099) and the inert language-locus feature `grasp-bottle-left` (7779), are run at $+\alpha_{\mathrm{cl}}$ with 12 paired trials each. The server logs the 64 largest SAE codes at every policy call in every condition, including Frozen VLA, so the readout of Sec. 5.2 is computed offline for every trial against the gripper, arm-motion, and subtask labels of the client logs, without additional trials. The dose-ladder and distribution-shift blocks of the frozen schedule were not run.

Planner prompt for steering by language. The planner (GPT-4.1-mini) is called once per request with the prompt below and the current top-camera frame attached as an image. The prompt and the operator's requests are written in Chinese; we give an English translation. Fields in braces are filled at each call: `{task}` is the current subtask, `{style}` is the operator's request, and `{candidates}` lists one vocabulary name per line with the fields of Sec. 3.5, including `dir_tested`, which marks names whose direction has been measured; `{MAX_CONCEPTS}` caps the number of proposals. The operator interface shows the planner's summary and the proposed names with their directions (Fig. 6a); the operator ticks the names to apply and sets the strength λ of Eq. (4).

```
You are the steering planner for a dual-arm robot (Agilex Cobot Magic,
breakfast scene). The robot is controlled by a VLA policy; inside its
action expert we can amplify (direction=+1) or suppress (direction=-1)
some labeled action concepts, which changes the style of its motion.
The operator only states the desired style. Using the current subtask
and the image, give 2-{MAX_CONCEPTS} suggestions from the candidate
concepts, ordered from most to least relevant, and mark the 1-3 most
relevant ones with recommend=true.
Rules: use only labels from the candidate list; for "do not do X",
suppress X with direction=-1; for "more Y / keep Y", use +1; prefer
the arm and objects that the current subtask uses; concepts with
dir_tested=yes have a measured direction, so prefer them when
relevance is similar.
Output JSON only: {"concepts": [{"label": str, "direction": 1 or -1,
"recommend": bool, "reason": str}], "summary": str}

Current subtask: {task}
Operator's desired style: {style}

Candidate concepts:
{candidates}
```

Full hardware results. Table 22 gives the per-placement success of steering by language, with 20 paired trials per placement, and Table 23 gives the complete closed-loop steering results, including both signs of each row of the $\pi_{0.5}$ breakfast policy.

Logging and scoring. Every trial stored synchronized RGB, timestamps, instruction, raw and executed actions, joint and end-effector state, gripper state, SAE codes, reconstruction, intervention, task and condition identifiers, success, collision, timeout, and duration.

Table 22: Recovering failures of the frozen $\pi_{0.5}$ breakfast policy on the robot: success (%) over 20 paired trials per placement (grasp: slice lifted clear of the slot or rack; task: criterion of Appendix K). Place bread amplifies the rows under the three names ticked in Fig. 6a. Load bread amplifies `approach-bread-rack-right` (7724), `grasp-bread-right` (10565), and `hold-bread-right` (12396). Both at λ=4 ($32\sigma_j$ per row; Eq. (4)). The last row pools the four other placements (n=80).

		Frozen $\pi_{0.5}$		SAE-VLA		Δ (pp)	
Subtask	Placement	Grasp	Task	Grasp	Task	Grasp	Task
Place bread	upright (demonstrated)	85	80	100	95	+15	+15
	sideways	5	5	100	95	+90	+90
Load bread	L / R	80	55	95	90	+15	+35
	R / L	5	5	90	85	+85	+80
	R / R (demonstrated)	90	80	95	80	+15	0
	L / L	10	5	90	80	+80	+75
Beyond demonstrations (pooled)		25	17.5	93.8	87.5	+68.8	+70

Table 23: Complete closed-loop steering by name, paired against SAE-in-loop; the main-text Table 3 keeps the gated-dose rows. Name: verified on held-out episodes; OL R_{spec}: open-loop specificity on the predeclared target at $16\sigma_j$ ($\pi_{0.5}$ breakfast policy: left gripper, except left-arm joints for `tilt-wrist-pour-left` (8176) and left wrist joints for `release-bottle-left` (8988)); LM control: a language-locus feature, outside the table. Target / collateral: change in the commanded target / collateral channels ($\times 10^{-3}$); event: change in the named motor event (LIBERO policies: descend onset, steps); consistent: episodes moving in the predicted direction. LIBERO policies: n=400 at α_{cl}, 200 otherwise; breakfast policy: n=12 paired trials per cell. Confidence intervals: Appendix I.

Row	Name	OL R_{spec}	Dose	Δ success (pp)	Target	Collateral	Event	Consistent (%)
$\pi_{0.5}$ *LIBERO policy*								
`approach-stove-top` (9168)	✓	1.5×	$+\alpha_{\text{cl}}$ ($8\sigma_j$)	−1.0	−13.5	+1.4	−1.3	74
			$+16\sigma_j$	−4.5	−20.2	−4.1	−2.5	82
			$-16\sigma_j$	−2.0	+11.4	−4.6	+2.8	82
			$+\alpha_{\text{ol}}$ ($32\sigma_j$)	−24.5	−65.5	−29.0	−5.2	87
Replications: $\pi_{0.5}$ LIBERO policy, second dictionary; OpenVLA-OFT LIBERO policy, layer 31								
`approach-bowl-open` (868)	×	1.1×	$+\alpha_{\text{cl}}$ ($16\sigma_j$)	−2.0	−12.3	−3.0	−1.0	68
`grasp-object` (7165)	×	10.4×	$+\alpha_{\text{cl}}$ ($4\sigma_j$)	−1.5	+17.3	−10.2	+4.6	88
$\pi_{0.5}$ *breakfast policy (Agilex robot)*								
`grasp-bottle-left` (3659)	✓	2.3×	$+\alpha_{\text{cl}}$	−8.3	+4.2	+1.8	−3.0	83
			$-\alpha_{\text{cl}}$	0.0	−3.8	+1.9	+2.7	83
`tilt-wrist-pour-left` (8176)	×	1.9×	$+\alpha_{\text{cl}}$	−8.3	+5.0	+2.7	−3.5	75
			$-\alpha_{\text{cl}}$	0.0	−4.6	+2.5	+3.2	75
`release-bottle-left` (8988)	×	1.1×	$+\alpha_{\text{cl}}$	0.0	+1.5	+1.4	−1.5	67
			$-\alpha_{\text{cl}}$	−8.3	−1.4	+1.5	+1.3	67
`hold-bottle-left` (10099), random control	×	0.8×	$+\alpha_{\text{cl}}$	−8.3	+0.8	+2.0	+0.2	50
`grasp-bottle-left` (7779), LM control	×	0.95×	$+\alpha_{\text{cl}}$	0.0	+0.1	+0.2	0.0	50