# USED DOES NOT MEAN NEEDED: SELF-ATTENTION IN INTERLEAVED FLOW-MATCHING ACTION EXPERTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Does a performance drop after deleting self-attention establish that an action expert needs it? We study flow-matching vision-language-action models whose action experts interleave separate cross-attention (CA) and self-attention (SA) modules. With CA retained, we distinguish deleting SA from a fixed dense-adapted policy from adapting a policy with SA disabled throughout. In GR00T N1.7 on LIBERO-10, dense inference, post-hoc deletion, and No-SA adaptation achieve 93.27%, 85.33%, and 93.60% success. In SmolVLA at two integration steps, their four-suite means are 76.40%, 23.84%, and 88.38% under matched adaptation configurations that differ only in the SA switch. Interventions in the fixed dense policies further show that dependence is conditional: early SA changes from a negative to a positive observed contribution when late SA is retained. This characterizes learned reliance without identifying the mechanism of successful adaptation. No-SA deployment reduces measured latency by 16.97% and 23.40% under the two models' respective timing protocols. Together, these results show that post-hoc dependence does not establish downstream architectural necessity in the studied CA–SA action experts.

## 1 INTRODUCTION

Does a performance drop after removing a module imply that the module must be retained? For a fixed trained policy, deletion measures dependence on its existing computation. Designing a downstream policy asks a different question: can effective behavior be learned when that computation is unavailable during adaptation? A deletion experiment can answer the first question while leaving the second open.

We study this distinction in *flow-matching VLA action experts with separate cross-attention (CA) and self-attention (SA) modules*, instantiated by the studied GR00T N1.7 and SmolVLA configurations. Flow-based VLAs generate continuous action chunks conditioned on multimodal representations (Black et al., 2025b; Bjorck et al., 2025; Shukor et al., 2025). In the interleaved experts considered here, CA remains available when SA is disabled. This gives a concrete architectural question: does the dense policy's reliance on SA imply that SA must remain available during downstream adaptation? Our scope is this CA–SA decomposition, rather than all flow-matching VLAs or action experts with fused attention.

The central observation is a contrast between deletion and adaptation (Figure 1). On LIBERO-10, GR00T achieves 93.27% success with dense inference and 85.33% after deleting SA. Yet a policy adapted with SA disabled reaches 93.60%. SmolVLA exhibits an even larger deletion loss at the tested two-step operating point, while its No-SA-adapted policy remains capable across four LIBERO suites. These results support a specific hypothesis: at least part of the dense policy's SA dependence is contingent on adaptation with that computation available. The experiments demonstrate a viable downstream solution without SA; they do not determine how the remaining computation acquires that solution.

We next characterize the dependence that deletion exposes. In SmolVLA's complete two-step intervention, retaining only late SA strongly outperforms retaining only early SA at the same budget. More strikingly, adding early SA lowers observed success when late SA is absent but raises it when late SA is present, in all four suite means. An operation's contribution therefore cannot be summarized



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

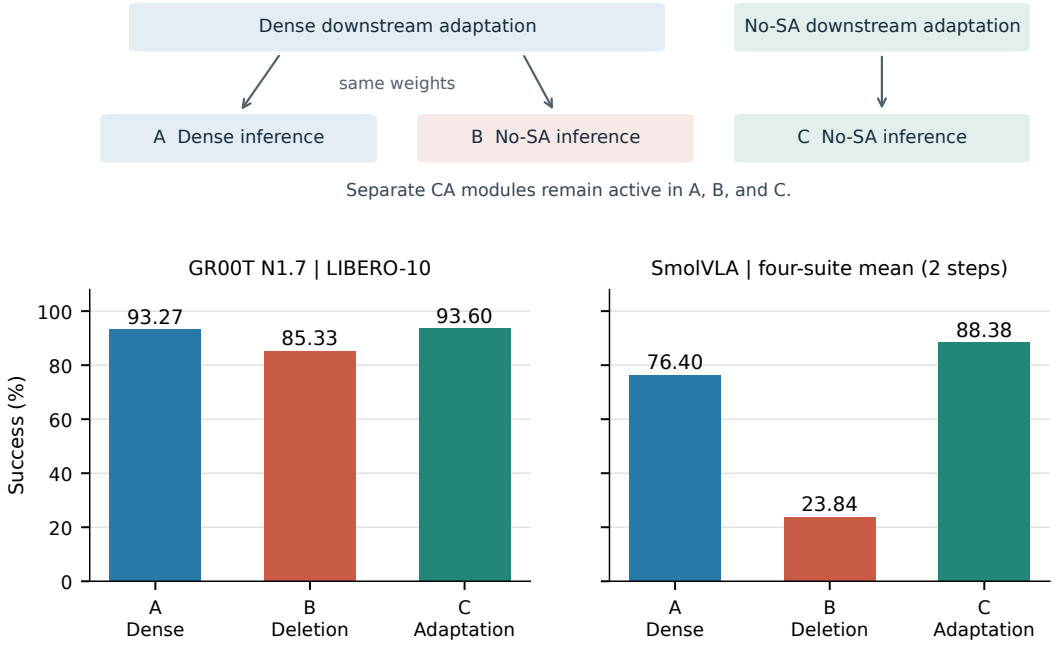


Figure 1: **Deletion tests a fixed policy; adaptation tests a target structure.** A and B share dense-adapted weights. C is separately adapted with SA disabled, not recovery training from B. CA remains active throughout. GR00T uses LIBERO-10; SmolVLA uses four-suite means at two steps. Bars are point estimates; evaluation repeats do not quantify training-run variability.

as a fixed increment independent of the retained computation. GR00T's flow–depth interventions likewise show that placement matters beyond execution count. These results characterize conditional reliance in dense policies; they are not a causal explanation of why No-SA adaptation succeeds.

The paper makes one central empirical argument: *learned SA dependence does not establish downstream SA necessity in the tested CA–SA flow-matching experts*. The deletion/adaptation comparison establishes the distinction, joint interventions characterize the learned dependence, and measured latency connects structural simplification to deployment. Local output-sensitivity analysis serves as a supporting diagnostic rather than an optimal allocation method. The design implication is correspondingly focused: when assessing whether SA must be retained for downstream adaptation, evaluate the target structure under adaptation as well as under fixed-weight deletion.

## 2 Experimental Design: Deletion versus Adaptation

### 2.1 Scope and Intervention

A flow-matching policy integrates a conditional vector field $v_\theta(x, t; c)$ to generate an action chunk from noise, conditioned on observation and instruction context $c$ (Lipman et al., 2023). Our intervention targets the named SA modules in an action expert that also contains separate CA modules. Disabling SA omits its residual update while retaining the residual path, feed-forward transformations, timestep conditioning, and CA computation. We use *No-SA* only for this intervention on the action expert; it does not remove attention from the VLM or disable CA.

The module semantics differ across the two implementations. GR00T's SA sequence includes state and action tokens, so deletion also removes SA-mediated state–action mixing. In the studied SmolVLA implementation, suffix queries in named SA layers can access cached VLM-prefix keys and values. Deletion thus removes more than action–action edges, even though separate CA and prefix computation remain. The common intervention is removal of the named SA computation with CA retained, not removal of an identical set of attention edges across models.

To locate dependence, let $k$ index integration execution order and $\ell$ logical SA depth. A binary mask specifies the retained executions:

$$M \in \{0,1\}^{T\times L}, \qquad B(M) = \sum_{k=1}^{T}\sum_{\ell=0}^{L-1} M_{k,\ell}. \tag{1}$$

$B$ counts SA executions, not distinct layers, parameters, or FLOPs. A flow-only schedule gates every SA depth at each step. Strings run from the first executed step to the last. GR00T uses four steps and 16 logical SA depths in 32 alternating blocks, giving 64 dense SA executions. SmolVLA uses eight logical SA depths and two steps in the reported comparison, giving 16 executions. Its first and second steps have numerical timesteps 1.0 and 0.5; "late" refers to execution order.

### 2.2 Three conditions, two questions

We compare three conditions: (A) dense downstream adaptation with dense inference; (B) the same dense-adapted weights with SA deleted at inference; and (C) separate downstream adaptation with SA disabled throughout, followed by No-SA inference. Condition C starts from a pretrained base and is not continued recovery training of condition B. Both adaptation and inference retain CA.

Let $J(\theta, M)$ denote closed-loop success, $\theta_D$ dense-adapted parameters, and $\theta_N$ No-SA-adapted parameters. The two contrasts are

$$\Delta_{\text{delete}} = J(\theta_D, \mathbf{0}) - J(\theta_D, \mathbf{1}), \tag{2}$$

$$\Delta_{\text{adapt}} = J(\theta_N, \mathbf{0}) - J(\theta_D, \mathbf{1}). \tag{3}$$

The first measures a fixed policy's reliance on SA. The second compares downstream solutions under different adaptation structures. A capable condition-C policy demonstrates feasibility without SA during downstream adaptation, even if condition B fails. The SmolVLA adaptation comparison holds the training configuration fixed except for the SA switch.

### 2.3 Evaluation and comparability

We evaluate on LIBERO (Liu et al., 2023). GR00T covers LIBERO-10 and Object, using separately adapted Dense/No-SA checkpoints at updates 18k and 26k, respectively. SmolVLA covers LIBERO-10, Goal, Object, and Spatial. Its Dense and No-SA policies use the same pretrained initialization, data, and training configuration, differing only in the SA switch. All four SmolVLA flow schedules use one dense checkpoint; all reported SmolVLA comparisons use two integration steps. Scores describe this operating point rather than a model-family leaderboard comparison.

Each GR00T configuration uses three evaluation seeds, ten tasks, and 50 episodes per task per seed. SmolVLA uses evaluation seeds 0, 1, and 2, with 50 episodes per task per seed and identical initial-state allocation and environment/code versions across configurations. We report suite means and their equal-weight four-suite mean. These are evaluation repeats, not independent training runs. Appendix B reports available repeated values.

GR00T's 85.33% deletion result uses Dense-18k, whose dense baseline is 93.27%; the official-checkpoint budget sweep's 91.87% is a separate baseline. These batches are not interchanged. SmolVLA executes one action before replanning in every closed-loop comparison. Appendix A gives the experimental scope and three key adaptation hyperparameters.

## 3 Deletion Loss Does Not Preclude No-SA Adaptation

### 3.1 GR00T: a deletion-sensitive policy and a capable No-SA solution

On LIBERO-10, deleting all SA executions from Dense-18k changes success from 93.27% to 85.33%, a loss of $7.93\,pp$. Separately adapting with No-SA gives 93.60% (Table 1). Thus, failure to preserve the dense policy under deletion coexists with near-dense performance from the target structure under adaptation.

The No-SA-minus-Dense differences across the three evaluation seeds are $+1.2, -1.2, +1.0\,pp$, with an approximate 95% Student-$t$ interval of $[-2.97, +3.64]\,pp$. The small positive mean is not evidence

Table 1: **Three conditions distinguish dependence from downstream feasibility.** Success rates (%); CA is retained in every column. SmolVLA uses two integration steps and an equal-weight suite mean. GR00T LIBERO-10 uses 18k checkpoints. Evaluation settings are described in Section 2.

| Model | Suite | A: Dense/Full | B: Dense/No-SA | C: No-SA/No-SA |
|---|---|---|---|---|
| GR00T | LIBERO-10 | 93.27 | 85.33 | 93.60 |
| | Object | 100.00 | 100.00 | 100.00 |
| SmolVLA | LIBERO-10 | 42.80 | 1.27 | 72.27 |
| | Goal | 84.33 | 39.80 | 88.47 |
| | Object | 89.47 | 32.47 | 97.87 |
| | Spatial | 89.00 | 21.80 | 94.93 |
| | **Mean** | **76.40** | **23.84** | **88.38** |

of superiority or formal equivalence. On Object, all three conditions—dense inference, post-hoc SA deletion, and No-SA adaptation—achieve 100.00% success. This extends No-SA feasibility to a second suite, but shows no observed deletion penalty there. The dependence exposed on LIBERO-10 is therefore not uniform across suites, and Object saturation limits discrimination between conditions.

An intermediate sparse structure is also feasible: Dense/Uniform-8 reaches 91.80%, and Uniform-8 adaptation/inference reaches 92.27%. Since No-SA reaches 93.60%, the evidence does not make eight retained SA executions a necessary endpoint. The central result is the existence of a capable No-SA-adapted policy, rather than identification of an optimal mask.

### 3.2 SmolVLA: severe deletion loss does not rule out feasibility

At two integration steps, deleting SA reduces the dense policy's four-suite mean from 76.40% to 23.84%. On LIBERO-10 the score falls from 42.80% to 1.27%. These are substantial dependencies of the tested dense checkpoint. Nevertheless, the No-SA-adapted policy achieves 88.38% overall and 72.27% on LIBERO-10, with higher observed means than Dense in every suite.

With training configurations matched and only the SA switch changed, the observed comparison favors No-SA adaptation at this two-step operating point. The evaluation seeds characterize the resulting checkpoints, not variation over independent training runs. The architectural conclusion is that the tested CA-retaining No-SA expert can support effective downstream behavior even where deletion from a dense policy performs poorly. Neither this comparison nor the GR00T result tests SA-free pretraining; both rely on pretrained representations.

## 4 Learned Dependence Is Conditional on Retained Computation

The preceding comparison establishes downstream feasibility. We now hold dense-policy weights fixed and ask what kind of dependence deletion measures. These interventions characterize the existing policy, rather than explaining the training dynamics of the No-SA solution.

### 4.1 A complete two-step test in SmolVLA

Two integration steps give four flow-only schedules: `11` (both), `01` (late only), `10` (early only), and `00` (neither). Both single-step schedules execute eight SA operations. Yet late-only inference remains within 1.27–5.33 points of Dense across suites, while early-only inference is below even No-SA in every suite mean (Table 2). The retained position matters, not just the amount of SA computation.

Let $S_{ab}$ be success with early gate $a$ and late gate $b$. Adding early SA changes success by $S_{10} - S_{00}$ without late SA, and by $S_{11} - S_{01}$ with late SA. Their difference is

$$I = (S_{11} - S_{01}) - (S_{10} - S_{00}). \tag{4}$$

For Goal, the early-SA effect changes from $-11.87\,pp$ to $+2.80\,pp$. All four suite means exhibit this sign reversal. A fixed additive contribution for each step cannot describe these observed scores. The interactions are descriptive; available rounded schedule means do not support a significance

Table 2: **Fixed dense SmolVLA checkpoint, complete two-step intervention.** Three-evaluation mean success rates (%). $I$ is the success-scale interaction in Equation 4, in percentage points, computed from supplied rounded means.

| Suite | `11` (16 calls) | `01` (8) | `10` (8) | `00` (0) | $I$ |
|---|---|---|---|---|---|
| LIBERO-10 | 42.80 | 39.87 | 0.07 | 1.27 | 4.13 |
| Goal | 84.33 | 81.53 | 27.93 | 39.80 | 14.67 |
| Object | 89.47 | 88.20 | 16.27 | 32.47 | 17.47 |
| Spatial | 89.00 | 83.67 | 15.80 | 21.80 | 11.33 |

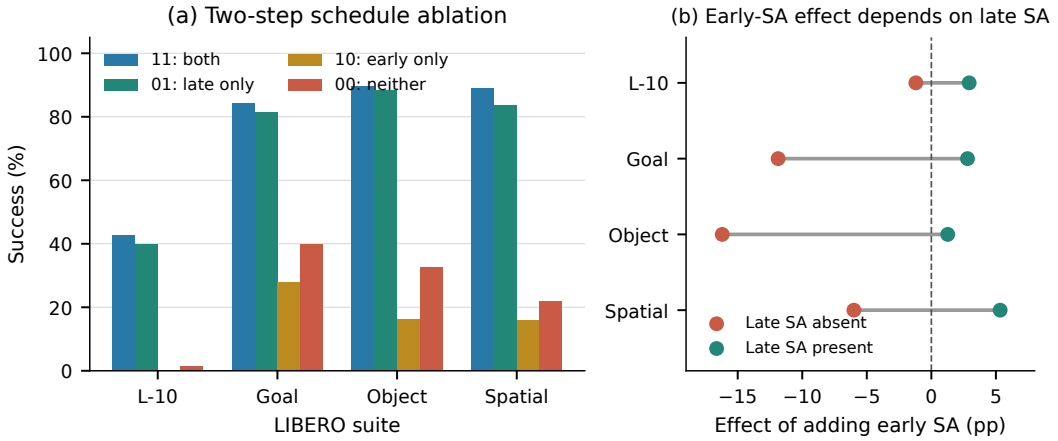


Figure 2: **Early SA has a conditional contribution in the dense policy.** Left: all four schedules. Right: the observed success change from adding early SA, with late SA absent or present. Its sign reverses in every suite mean. These are success-scale contrasts, not measurements of an internal correction mechanism.

test. They also do not locate the source of non-additivity within attention, integration, or closed-loop dynamics.

This conditionality sharpens the interpretation of deletion: a measured contribution belongs to a policy with a particular retained computation. The dependence is real, but it is not an intrinsic, context-independent value of the module, nor a test of what a separately adapted structure can achieve.

### 4.2 GR00T: PLACEMENT MATTERS BEYOND THE BUDGET

A dedicated four-step rerun gives 94.33% success for dense `1111` and 93.40% for final-only `0001`, reducing SA executions from 64 to 16. The paired three-seed difference is $-0.93\,pp$, with an approximate 95% interval of $[-4.46, +2.59]\,pp$. Much of the observed success is retained even though SA is absent from the first three steps; CA, FFNs, and integration at those steps still execute.

In the separate official-checkpoint budget sweep, Uniform-8 reaches 91.93%, compared with 91.87% for Dense and 83.40% for one fixed Random-8 mask. Moving Uniform-8's aggregate depth set entirely to the last step gives 91.07%, while another mask with two random depths per step gives 87.67%. Equal budget or equal flow coverage therefore does not uniquely determine performance. These comparisons establish position dependence rather than a universal preference for uniform placement; additional random masks and full results appear in Appendix C.

### 4.3 LOCAL OUTPUT DEVIATION AND CLOSED-LOOP CONTRIBUTION DIFFER

Single-operation deletion measures the change in the dense policy's action output while retaining all other SA operations. Figure 3 shows non-uniform dependence across flow execution time and depth. GR00T's mean deviations across its four steps are 0.0280, 0.0358, 0.0490, and 0.1522, with

the largest mean at the final step. SmolVLA's synthetic-input diagnostic instead gives depth-averaged relative deviations of 0.08790 and 0.07620 at the first and second steps. The latter uses the script's deterministic synthetic input with no external observation file, rather than task observations.

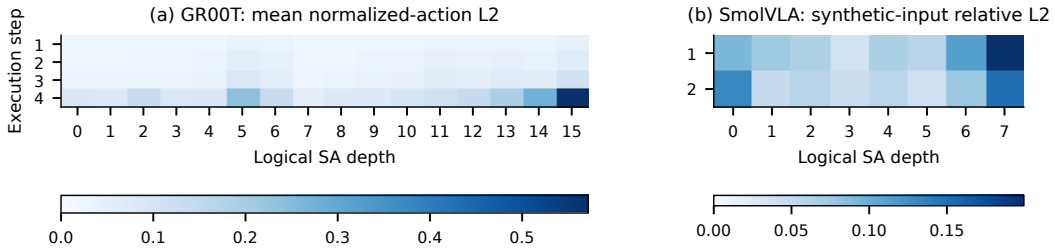


Figure 3: **Flow–depth structure of single-operation output sensitivity.** Each cell disables one SA execution from the dense policy with input and initial noise fixed. GR00T: padding-masked normalized-action L2 averaged over 500 observations. SmolVLA: output-difference L2 divided by dense-output L2 plus $10^{-8}$, for a deterministic synthetic input tensor set. Independent scales reflect different metrics and sampling scopes; magnitudes are not compared across models.

In the closed-loop experiment, late-only SA is much more effective than early-only SA (Table 2). The synthetic-input heatmap does not estimate sensitivity on the LIBERO observation distribution and cannot establish its temporal ordering on real task inputs. More generally, single-operation action deviation and joint-mask task completion probe different properties: they differ in intervention size, outcome metric, and evaluation inputs. The heatmap illustrates dependence under the diagnostic input, while the rollout experiment establishes closed-loop contribution. Neither explains how No-SA adaptation succeeds.

Sensitivity-based Top-$K$ retention also does not consistently outperform uniform allocation in GR00T. Complete scores, depth rankings, and budget comparisons appear in Appendices C–E.

## 5 Deployment Benefit of the No-SA Structure

No-SA adaptation also yields a practical simplification (Table 3). GR00T Object's batch-1 full-policy latency falls from 98.538 to 81.820 ms, a 16.97% reduction; its Dense and No-SA policies each achieve 100% observed success on Object. SmolVLA action-chunk prediction falls from 59.19 to 45.34 ms, a 23.40% reduction at two integration steps. Both latency comparisons use separately adapted Dense and No-SA models.

Table 3: **Measured deployment savings with CA retained.** Latency in ms; timing scopes differ across models. Neither includes simulator or robot-control-loop time.

| Model | Timing scope | Dense | No-SA | Reduction |
|---|---|---|---|---|
| GR00T | Full policy, batch 1 | 98.538 | 81.820 | 16.97% |
| SmolVLA | Action chunk, batch 1 | 59.19 | 45.34 | 23.40% |

GR00T timing includes preprocessing, backbone and action-expert execution, integration, and decoding, with CUDA synchronization. SmolVLA measures `predict_action_chunk` on H200 in FP32. These are within-model deployment comparisons, not a common cross-model benchmark. SA-call reduction is not proportional to total latency reduction because CA, FFNs, the VLM, and other computation remain. Additional timing details are given in Appendix F.

## 6 Related Work

**Continuous-action VLAs and expert architecture.** Diffusion Policy models action sequences through denoising (Chi et al., 2023), while flow matching learns a conditional vector field (Lipman et al., 2023). The $\pi_0$ and $\pi_{0.5}$ families couple multimodal representations to flow-based action generation (Black et al., 2025b;a); GR00T N1 and SmolVLA likewise use an action-generating expert (Bjorck et al., 2025; Shukor et al., 2025). These papers motivate the repeated expert computation studied here, but their attention interfaces are not interchangeable. In particular, our intervention requires separately identifiable CA and SA modules; we do not infer the same result for $\pi$-style joint attention. CogACT studies a diffusion-transformer action module (Li et al., 2024), and recent controlled action-head comparisons examine architecture, initialization, and latency (Sun et al., 2026). OpenVLA-OFT shows that downstream adaptation and action decoding choices can substantially change the speed–success tradeoff (Kim et al., 2025). Our narrower question is whether reliance on a named SA module in a *fixed* flow-matching expert predicts the capability of a separately adapted expert with CA retained and SA absent.

**Efficiency at different points in the VLA pipeline.** EfficientVLA combines backbone pruning, visual-token selection, and action-head feature reuse without additional training (Yang et al., 2025); LightVLA learns visual-token pruning during fine-tuning (Jiang et al., 2025). DySL-VLA trains a skip-aware layer-routing policy (Yang et al., 2026). These approaches alter the observation path or skip broader layers, whereas we isolate the expert's SA residual update and leave its CA, FFN, and other paths intact. For flow-based policies, SnapFlow distills the iterative trajectory to fewer expert calls (Luan et al., 2026), and Coda learns an endpoint correction after a shorter trajectory (Tang et al., 2026). These reduce or reallocate integration steps; our flow–depth schedules instead vary *which SA computations* execute within the retained expert calls. In related diffusion-transformer settings, DiTFastAttn compresses attention after training (Yuan et al., 2024), and test-time action-diffusion sparsity selects residual computation dynamically (Ji et al., 2026). Their post-hoc efficiency goals do not test No-SA downstream adaptation.

**SA-free designs and the deletion–adaptation distinction.** MergeVLA constructs a cross-attention-only action expert to support cross-skill model merging (Fu et al., 2026); SmolVLA-Mamba replaces causal SA with selective state-space modeling (Mohsen et al., 2026). These are particularly close demonstrations that alternative expert designs are viable, but neither asks whether a large deletion loss from the *same dense-adapted checkpoint* rules out an SA-disabled adaptation. A recent VLA pruning study reports a severe immediate success loss after pruning the language backbone and recovery through offline hidden-state distillation (Kim et al., 2026). It reinforces the need to evaluate an adapted structure, while targeting a different component and recovery procedure. Attention-head ablations measure which computations a trained model currently uses (Michel et al., 2019; Voita et al., 2019). Structured pruning distinguishes retaining trained weights from training a target architecture (Liu et al., 2019), and lottery-ticket results show that initialization and training route matter (Frankle & Carbin, 2019). These are methodological precedents, not evidence about robot SA itself. We make the contrast explicit in two CA–SA flow-matching experts: post-hoc deletion probes learned dependence, and target-structure adaptation probes downstream feasibility. Our joint-mask rollouts further show that the apparent contribution of early SA depends on whether late SA remains available; local action deviation is therefore not a substitute for closed-loop success.

## 7 Discussion and Limitations

Post-hoc deletion correctly identifies a dependence of the current policy. The unsupported inference would be to treat that dependence as sufficient evidence that the architecture must retain SA during downstream adaptation. Our results supply counterexamples in two CA–SA flow-matching experts. The joint interventions further show that the measured dependence is conditional on retained computation, without establishing the mechanism by which adaptation succeeds.

The evidence covers two interleaved configurations and simulated LIBERO tasks. It does not extend directly to fused-attention experts or establish SA-free pretraining. Three evaluation repeats do not quantify training-run variation. Object saturation limits discrimination between architectures, and the synthetic-input SmolVLA sensitivity diagnostic is not a task-distribution estimate. We have

not established a minimal architecture, optimal mask, or advantage over smaller dense experts or alternative solvers. Real-robot behavior, distribution shift, and training-compute savings remain outside the measured scope.

## 8 CONCLUSION

In flow-matching VLA action experts with separate CA and SA modules, a dense policy can depend strongly on SA while a policy adapted without it remains capable. Fixed-policy interventions reveal conditional reliance; target-structure adaptation establishes downstream feasibility; measured latency demonstrates a deployment benefit. For the tested architectures, used does not mean needed: post-hoc dependence alone does not establish that SA must be retained during downstream adaptation.

### AI USE STATEMENT

Generative AI tools were used to polish language and correct grammar, assist with literature search, and assist with writing and checking figure-generation and numerical verification code. During manuscript revision, they also helped draft parts of the related-work section and phrase the interpretation of supplied results, including the distinction between post-hoc SA deletion and adaptation with SA disabled. No experimental outcomes were generated by AI. The authors reviewed the AI-assisted text and code and take responsibility for the final manuscript, citations, results, and interpretation.

### ETHICS STATEMENT

This study analyzes robot policies in simulation. Success on LIBERO and reduced policy latency do not establish safe behavior on physical robots or under distribution shift. Deployment would require independent validation of control behavior and safety constraints. The study does not introduce new human-subject data or a new demonstration dataset.

### REPRODUCIBILITY STATEMENT

Section 2 defines the intervention and the checkpoint distinctions. Appendices A–F report documented settings, repeated results, masks, and diagnostic values. The accompanying source package includes the numerical tables, both supplied sensitivity CSVs, and scripts that regenerate every figure. These plotting materials reproduce the reported figures and tables; they are not a complete training and rollout implementation.

## REFERENCES

Johan Bjorck, Fernando Castaneda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40, 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke,

Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, 2025b. doi: 10.15607/RSS.2025.XXI.010.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems*, 2023.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*, 2019.

Yuxia Fu, Zhizhen Zhang, Yuqi Zhang, Zijian Wang, Zi Huang, and Yadan Luo. MergeVLA: Cross-Skill Model Merging Toward a Generalist Vision-Language-Action Agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kangye Ji, Yuan Meng, Jianbo Zhou, Ye Li, Chen Tang, and Zhi Wang. Test-time Sparsity for Extreme Fast Action Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Titong Jiang, Xuefeng Jiang, Yuan Ma, Xin Wen, Bailin Li, Kun Zhan, Peng Jia, Yahui Liu, Sheng Sun, and Xianpeng Lang. The Better You Learn, The Smarter You Prune: Towards Efficient Vision-Language-Action Models via Differentiable Token Pruning. *arXiv preprint arXiv:2509.12594*, 2025.

Chiyoung Kim, Sanghyuk Roy Choi, and Minhyeok Lee. Recovering Aggressively Pruned Vision-Language-Action Models with Offline Hidden-State Distillation. *arXiv preprint arXiv:2609.19579*, 2026.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.017.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation. *arXiv preprint arXiv:2411.19650*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

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, Datasets and Benchmarks Track*, volume 36, 2023.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the Value of Network Pruning. In *International Conference on Learning Representations*, 2019.

Wuyang Luan, Junhui Li, Weiguang Zhao, Wenjian Zhang, Tieru Wu, and Rui Ma. SnapFlow: One-Step Action Generation for Flow-Matching VLAs via Progressive Self-Distillation. *arXiv preprint arXiv:2604.05656*, 2026.

Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better than One? In *Advances in Neural Information Processing Systems*, 2019.

Farida Mohsen, Thowayba Elkaffash, Mohammad Reza Chalak Qazani, Mohamed Mabrok, Nader Meskin, and Ali Safa. Mamba-based Selective State Space Modeling Improves the Accuracy-Complexity Tradeoff of SmolVLA Vision-Language-Action Experts. *arXiv preprint arXiv:2608.21407*, 2026.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Luoyang Sun, Guoyang Xia, Fengfa Li, Lei Ren, Xinyu Cui, Haifeng Zhang, Fangxiang Feng, Kaike Zhang, Kun Zhan, Yan Xie, Jun Wang, and Cheng Deng. What Makes an Efficient VLA? Navigating Action-Head Design, Scaling, and Latency. *arXiv preprint arXiv:2609.13984*, 2026.

Zhipeng Tang, Xinda Chen, Weining Rao, Xiao Li, Wenting Tan, Yuning Wang, Xiao Shi, and Xiaofang Zhao. Fewer Steps, Better Actions: Rethinking Flow-Matching Inference for VLA Policies. *arXiv preprint arXiv:2609.21216*, 2026.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Yantai Yang, Yuhao Wang, Zichen Wen, Zhongwei Luo, Chang Zou, Zhipeng Zhang, Chuan Wen, and Linfeng Zhang. EfficientVLA: Training-Free Acceleration and Compression for Vision-Language-Action Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zebin Yang, Yijiahao Qi, Tong Xie, Bo Yu, Shaoshan Liu, and Meng Li. DySL-VLA: Efficient Vision-Language-Action Model Inference via Dynamic-Static Layer-Skipping for Robot Manipulation. *arXiv preprint arXiv:2602.22896*, 2026.

Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. DiTFastAttn: Attention Compression for Diffusion Transformer Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

## A Experimental Scope and Provenance

Table 4 separates the experimental batches used in the paper. Identical model-family names do not imply identical checkpoints. In particular, the official-checkpoint budget sweep and the independently fine-tuned 18k architecture comparison have different dense success rates. The post-hoc No-SA result uses the Dense-18k checkpoint; its success rates are compared with that checkpoint's full-inference baseline.

Table 4: Documented evaluation scope. "Three evaluations" is not a statement about three independently trained policies.

| Experiment | Checkpoint / intervention | Evaluation scope |
|---|---|---|
| GR00T flow-only rerun | Official LIBERO-10; `1111/0001` | 3 seeds; 10 tasks; 50 episodes/task/seed |
| GR00T budget sweep | Official LIBERO-10; 13 masks | Same episode counts |
| GR00T architecture | Dense/No-SA, update 18k | Same episode counts |
| GR00T post-hoc No-SA | Dense-18k; all SA off | Same episode counts |
| GR00T Object comparison | Dense/Full, Dense/No-SA, No-SA/No-SA; update 26k | Same episode counts |
| SmolVLA schedules | One Dense checkpoint; `11/01/10/00` | 2 integration steps; 4 suites; seeds 0/1/2 |
| SmolVLA No-SA | No-SA adaptation from `smolvla_base`, update 25k | 2 integration steps; same 4 suites; seeds 0/1/2 |

**Matched SmolVLA adaptation.** Dense and No-SA use identical training configurations aligned with the official recipe, with the SA switch as the sole change. Both start from `smolvla_base`; Table 5 lists three key hyperparameters. All reported SmolVLA closed-loop evaluations use two integration steps and `n_action_steps=1`, with seeds 0, 1, and 2. Each task is evaluated for 50 episodes per seed, giving 1,500 episodes per suite per configuration. Initial-state allocation and environment/code versions are identical across configurations.

Table 5: Three key SmolVLA adaptation hyperparameters, shared by Dense and No-SA.

| Parameter | Value |
|---|---|
| Batch size | 16 |
| Training updates | 25,000 |
| Peak learning rate | $10^{-4}$ |

Both latency comparisons use H200 at batch size one. GR00T uses a fixed observation, CUDA synchronization, 100 warm-up iterations, and 1,000 timed iterations. Its full-policy scope includes preprocessing and action decoding, excluding simulator and communication time. SmolVLA measures action-chunk prediction in FP32 at two integration steps. The scopes differ and are compared only within each model.

## B Repeated Architecture Results

Table 6: GR00T LIBERO-10 at update 18k. Each seed aggregates 500 episodes. No-SA deletion uses the same dense checkpoint.

| Adaptation | Inference | Seed 0 | Seed 1 | Seed 2 | Mean |
|---|---|---|---|---|---|
| Dense | Full-64 | 93.0 | 94.4 | 92.4 | 93.27 |
| Dense | No-SA | 85.0 | 86.0 | 85.0 | 85.33 |
| Dense | Uniform-8 | 91.6 | 92.6 | 91.2 | 91.80 |
| Uniform-8 | Uniform-8 | 93.2 | 93.6 | 90.0 | 92.27 |
| No-SA | No-SA | 94.2 | 93.2 | 93.4 | 93.60 |

Table 7: SmolVLA No-SA-adapted 25k checkpoint at two steps. The three evaluations use seeds 0, 1, and 2. SD is the sample standard deviation over the three evaluation rates, not training uncertainty.

| Suite | Seed 0 | Seed 1 | Seed 2 | Mean (%) | SD (pp) |
|---|---|---|---|---|---|
| LIBERO-10 | 72.6 | 72.0 | 72.2 | 72.27 | 0.31 |
| Goal | 87.8 | 88.2 | 89.4 | 88.47 | 0.83 |
| Object | 98.2 | 97.6 | 97.8 | 97.87 | 0.31 |
| Spatial | 95.0 | 94.8 | 95.0 | 94.93 | 0.12 |

For GR00T paired evaluation-seed comparisons, a descriptive paired-seed interval uses $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$, with $t_{0.975,2} \approx 4.303$. Such an interval conditions on the trained checkpoints and is not an interval over training randomness. We do not pool the four SmolVLA suites into an episode-level binomial test or treat the reported suite means as independent training runs. The schedule interaction values are computed from rounded means and are descriptive.

## C Flow, Depth, and Budget Controls

The matched GR00T flow-only rerun gives Dense `1111` seed rates 93.4/95.4/94.2% and final-only `0001` rates 94.0/94.2/92.0%. These belong to that dedicated rerun, not to the 18k architecture comparison or the budget sweep below.

Table 8: Official-checkpoint GR00T budget sweep on LIBERO-10. Each row contains three seeds and 1,500 episodes. Top-$K$ uses the full-grid single-operation sensitivity scores. Masks at different budgets are not necessarily nested.

| Allocation | SA calls | Seed 0 | Seed 1 | Seed 2 | Mean |
|---|---|---|---|---|---|
| Dense | 64 | 91.2 | 90.8 | 93.6 | 91.87 |
| Top-$K$ | 32 | 91.6 | 90.4 | 90.6 | 90.87 |
| Uniform | 32 | 91.2 | 92.4 | 93.4 | 92.33 |
| Fixed random | 32 | 92.6 | 89.2 | 90.2 | 90.67 |
| Top-$K$ | 16 | 90.4 | 90.8 | 90.4 | 90.53 |
| Uniform | 16 | 89.0 | 88.8 | 90.4 | 89.40 |
| Fixed random | 16 | 90.2 | 90.8 | 87.0 | 89.33 |
| Top-$K$ | 8 | 88.4 | 91.2 | 90.0 | 89.87 |
| Uniform | 8 | 93.8 | 90.8 | 91.2 | 91.93 |
| Fixed random | 8 | 84.0 | 82.6 | 83.6 | 83.40 |
| Top-$K$ | 4 | 89.6 | 92.2 | 88.4 | 90.07 |
| Uniform | 4 | 90.0 | 91.8 | 91.6 | 91.13 |
| Fixed random | 4 | 82.2 | 83.2 | 84.6 | 83.33 |

Uniform-8 retains depths $\{0, 9\}$, $\{2, 11\}$, $\{4, 13\}$, and $\{6, 15\}$ in execution steps 1–4, respectively. Its exact mask is

```
1000000001000000 / 0010000000010000 /
0000100000000100 / 0000001000000001.
```

Moving the same aggregate depth set to the last step gives Late-only-8 (91.07%). A balanced random mask with two depths per step gives 87.67%. A subsequent three-mask robustness sweep with mask seeds 101, 102, and 103 produces means of 88.20%, 94.13%, and 89.87%. These additional masks come from a separate evaluation batch and are reported as a robustness check. One random mask exceeds Uniform-8's earlier mean; uniform placement is not established as globally optimal or superior to random placement in distribution.

The final-step-only sensitivity check is a different intervention baseline from the full $4 \times 16$ grid. At fixed outer schedule `0001`, seed rates are 92.4/92.2/93.2% for all 16 depths, 92.2/91.8/90.8% for depth 7 off, and 89.2/90.0/89.2% for depth 15 off. The most-minus-least-sensitive deletion contrast is $-2.13\,pp$, with an approximate paired interval $[-4.01, -0.25]\,pp$. This checks two selected extremes; it does not establish the rank's optimality under multiple simultaneous deletions.

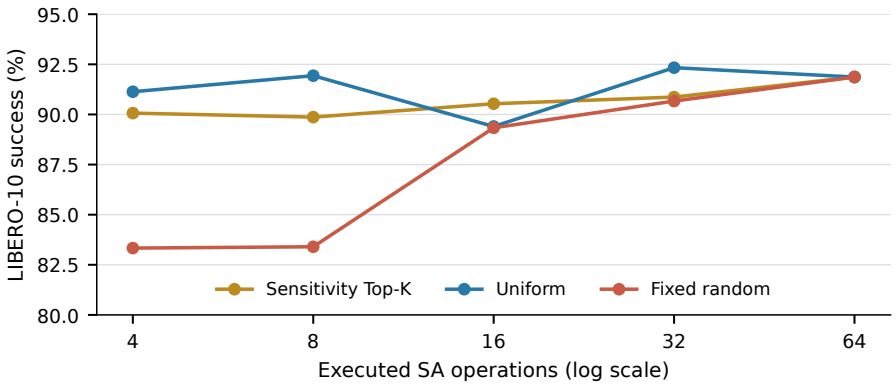


Figure 4: GR00T budget curve. Connecting lines organize different masks; they are not a trajectory obtained by progressively deleting a nested set. A fixed random mask per budget does not characterize a random-mask distribution.

## D Supporting Single-Operation Sensitivity Analysis

For a fixed observation and matched initial flow noise, the GR00T diagnostic compares the dense normalized action chunk $\hat{a}_{\mathbf{1}}$ with $\hat{a}_{-(k,\ell)}$, obtained by disabling one SA execution:

$$D(k,\ell) = \mathbb{E}_{o,\epsilon}\left[\left\|\hat{a}_{\mathbf{1}}(o,\epsilon) - \hat{a}_{-(k,\ell)}(o,\epsilon)\right\|_2\right]. \tag{5}$$

The full grid uses 500 observations balanced across ten tasks. This measures deviation from the dense action output, not task reward or adaptation feasibility.

GR00T's mean full-grid deviations across four steps are 0.0280, 0.0358, 0.0490, and 0.1522 (Figure 3); the final-step mean is 5.43 times the first. A separate final-only diagnostic gives 0.0575 for depth 7 and 0.5615 for depth 15, a 9.8-fold ratio. They are the least and most sensitive depths in all ten task-conditioned summaries. The corresponding closed-loop comparison is reported in Appendix C.

SmolVLA's synthetic-input diagnostic instead gives first- and second-step means of 0.08790 and 0.07620. Depth 7 is most sensitive at both steps, but only two of eight depths have a larger second-step value. This diagnostic does not replicate GR00T's late-step concentration. It is a synthetic-input relative-deviation measurement, rather than the observation-averaged absolute deviation in Equation 5; it does not establish a task-population temporal profile.

At eight and four calls, GR00T sensitivity Top-$K$ achieves 89.87% and 90.07%, versus 91.93% and 91.13% for uniform allocation. Top-$K$ is better at 16 calls but worse at 32; all corresponding paired-seed 95% intervals include zero. These results do not validate single-operation ranking as an optimal joint allocator. They also do not identify whether metric choice, joint intervention, or closed-loop dynamics explains the gap between output sensitivity and task outcomes.

## E SmolVLA Diagnostic Values and Metric Scope

The diagnostic keeps the input tensors and initial flow noise fixed. Starting from dense schedule `11111111/11111111`, it disables exactly one logical SA layer at one of the two flow steps, for 16 interventions. The reported statistic is

$$D_{\mathrm{rel}}(k,\ell) = \frac{\|a_{-(k,\ell)} - a_{\mathbf{1}}\|_2}{\|a_{\mathbf{1}}\|_2 + 10^{-8}}. \tag{6}$$

The norm is over the entire action-output tensor returned by `sample_actions`, without a padding mask or per-observation averaging. Step means average the eight depth scores. The reported run uses no `--input` file and therefore uses the script's deterministic synthetic input: a zero-valued image, the instruction "pick up the object," a linearly spaced state vector in $[-0.25, 0.25]$, and a linearly spaced initial flow tensor in $[-1, 1]$. These tensors remain identical for the dense run and all 16 interventions. This is a synthetic diagnostic, not a sampled LIBERO observation.

Table 9: SmolVLA synthetic-input single-operation-off diagnostic. Each cell records the reported normalized deviation after deleting one depth at one step from the otherwise dense $2 \times 8$ grid.

| Depth | Step 1 ($t = 1.0$) | Step 2 ($t = 0.5$) | Step 2 minus step 1 |
|---|---|---|---|
| 0 | 0.09164 | 0.13245 | $+0.04081$ |
| 1 | 0.07382 | 0.04896 | $-0.02486$ |
| 2 | 0.06396 | 0.05870 | $-0.00526$ |
| 3 | 0.03747 | 0.04502 | $+0.00755$ |
| 4 | 0.06608 | 0.05614 | $-0.00994$ |
| 5 | 0.05777 | 0.04185 | $-0.01592$ |
| 6 | 0.11293 | 0.07580 | $-0.03713$ |
| 7 | 0.19953 | 0.15067 | $-0.04886$ |
| Mean | 0.08790 | 0.07620 | $-0.01170$ |

The descending depth rankings are $(7, 6, 0, 1, 4, 2, 5, 3)$ and $(7, 0, 6, 2, 4, 1, 3, 5)$, with Spearman correlation $6/7 \approx 0.857$. The maximum-to-minimum ratios are 5.33 and 3.60. These describe the synthetic-input export, not uncertainty over observations or tasks.

GR00T instead provides 500 observations per location, with mean, standard deviation, median, and 95th-percentile summaries. Its primary metric uses padding-masked normalized action coordinates. Decoded physical-unit L2 is auxiliary because clipping can suppress differences. The two models therefore use separate heatmap scales and their absolute values are not compared. GR00T's final-only depth study has a different baseline mask from its full-grid diagnostic.

## F Additional Timing Measurements

For GR00T, the 98.538/81.820 ms deployment comparison corresponds to a $1.204\times$ speedup. A separate same-checkpoint mask comparison reports 98.462 ms for Full-64 and 82.411 ms for Uniform-8; it corroborates a latency benefit from skipping SA with weights fixed, but is not the No-SA adaptation comparison.

For SmolVLA, the 59.19/45.34 ms means correspond to a $1.305\times$ speedup. Reported medians are 56.48/44.45 ms and 95th percentiles are 72.00/50.10 ms. These measure the Dense and No-SA-adapted models, not the late-only schedule. Hardware and measurement scopes are described in Appendix A. No timing confidence intervals or proportional training-compute savings are inferred from these summaries.