CANDIDATE–PRIORITY COUPLING IN DISCRETE-DIFFUSION ROBOT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful robot execution does not reveal whether a policy can recover multiple valid behaviors from the same state. We identify a source of coverage loss in discrete-diffusion decoding: ranking token positions by the probability of the sampled candidates that will be committed. An exact selection identity shows how this candidate–priority coupling can distort the output distribution even with exact conditionals. Sample-split confidence ranks positions with an independent pilot draw and commits a separate draw, removing this dependence. On FORKBENCH-CHOICE, a controlled manipulation task, matched two-draw interventions in the first policy query hold weights and model-call budgets fixed. Sample splitting improves primary-fit Coverage@10 by 20.4 percentage points (95% paired-state bootstrap interval: 11.1–29.6), while satisfying a prespecified validity guard. A second fitted checkpoint shows a consistent gain, with disclosed technical qualifications. A separately trained VLA also shows increased diversity on a fresh Approach-Side panel when choosing between prevalidated fixed-controller routes. Together, the exact identity and matched no-remasking interventions identify candidate–priority coupling as a source of finite-budget behavioral coverage loss. Decoding rules determine which valid behaviors a fixed policy recovers and warrant evaluation alongside task success.

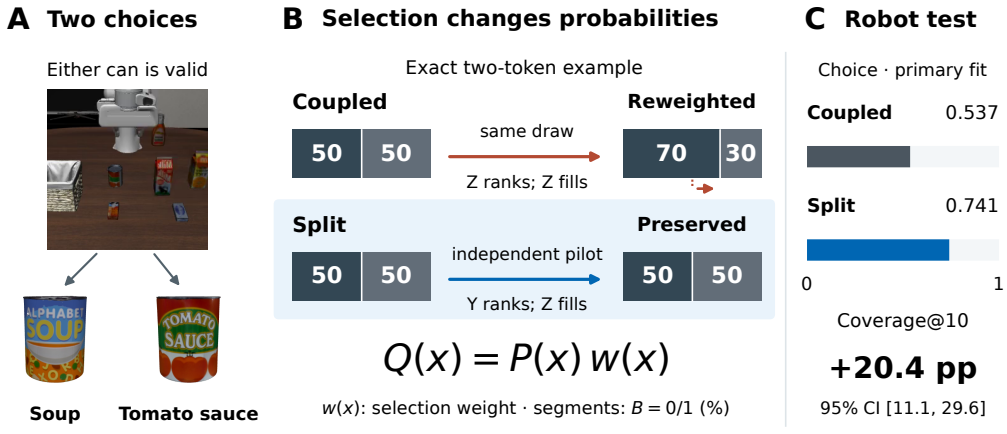


Figure 1: **Selection can change the law of an exact model.** (A) Either can is a valid Choice outcome; images illustrate the task. (B) In the exact two-token example ($\alpha = 3/5$, zero priority noise), coupled selection turns 50/50 branch mass into 70/30; independent-pilot selection preserves 50/50. These are not robot measurements, and the oracle guarantee does not extend to learned parallel decoding. (C) Robot Coverage@10 improves by 20.4 percentage points (95% paired-state bootstrap CI: 11.1–29.6; 27 states), with matched weights, two draws and model-call budgets.

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.

1 INTRODUCTION

Many robot observations admit several correct actions: a manipulator can grasp from either side or choose either of two objects satisfying an instruction. Access to multiple valid solutions can support planning and failure recovery even when single-rollout success is unchanged. Diffusion policies can model multimodal action distributions that regression policies tend to average (Chi et al., 2025); whether decoding preserves these modes is a separate question.

Discrete Diffusion VLA (DD-VLA) brings this idea to tokenized VLA action chunks (Liang et al., 2026). At each denoising iteration it predicts every masked action token, samples candidates in parallel, and decides which candidates to commit. The DD-VLA paper defines max-confidence selection, whereas the released LIBERO implementation ranks positions by the probability assigned to each *sampled* token. We evaluate this released-code rule and call it *sampled-confidence-first*. Published max-confidence success ablations do not evaluate this exact score or repeated-sampling behavioral coverage.

A sampled candidate can bias the output by promoting its own commitment. We call this *candidate–priority coupling*. Our exact selection identity characterizes the resulting reweighting of the output distribution, even with correct conditionals. In exact sequential decoding, ordering independent of fresh candidates preserves the joint; sample splitting removes this dependence (Section 3.1). The analysis isolates value-dependent selection as a source of distortion, distinct from the branch-deferral hypothesis motivated by The Flexibility Trap (Ni et al., 2026).

We freeze a VLA and vary only first-query commitment order on FORKBENCH-CHOICE, a controlled task with two predicate-defined valid modes and paired demonstrations. The prospectively frozen 27-state expansion shows larger coverage under random, temporal-forward, and uncertainty-first priorities, despite similar competence point estimates. A two-draw intervention then isolates whether a candidate promotes its own commitment: coupled confidence ranks by the committed draw, whereas Split ranks by an independent pilot. With draws and model calls matched, Split improves coverage on the primary fit and a technically qualified second fit while passing the validity guard. On a fresh Approach-Side panel, the same intervention also increases diversity of VLA choices between validated fixed-controller routes, extending the selection test to a separately trained policy and fresh contexts.

The contribution connects an exact selection law, a distribution-preserving condition for sequential decoding, and a matched intervention demonstrating the behavioral cost of same-draw coupling in a learned robot policy. The robot intervention changes first-query commitment without remasking, holding weights and model-call budgets fixed. Behavioral coverage is therefore a property of the policy and its decoding rule, not model weights alone.

2 RELATED WORK

Robot multimodality and controlled evaluation. Diffusion Policy evaluates alternate Push-T approaches and block-order choices (Chi et al., 2025). D3IL provides multimodal tasks, behavior descriptors, and conditional behavior entropy from repeated fixed-state simulations (Jia et al., 2024); VQ-BeT evaluates closed-loop mode capture using behavior entropy (Lee et al., 2024). MARS Policy gates stochastic versus history-based flow sources and solver effort across task phases (Jia et al., 2026). These establish multimodal-policy, benchmark, and metric precedents. FORKBENCH-CHOICE is instead a minimal testbed with predicate-defined outcomes for an auditable fixed-checkpoint reveal-order intervention, not a general benchmark or new diversity metric.

Generative robot policies. OpenVLA tokenizes actions and OpenVLA-OFT adds parallel action chunks (Kim et al., 2025c;b). DD-VLA introduces masked discrete diffusion (Liang et al., 2026); MMaDA-VLA jointly denoises goals and actions, whereas DFM-VLA keeps the action sequence revisable (Liu et al., 2026; Chen et al., 2026). Revisability targets early errors; we study reveal-order effects on coverage. Other diversity, temporal-block, and plan-token approaches differ in their learned representation or decoding interface; Appendix T.3 details these comparisons.

Order in masked diffusion and VLAs. MaskGIT popularized confidence-based parallel commitment (Chang et al., 2022). Subsequent work learns or searches reveal paths, identifies early-decision

and confidence-shortcut failures, and scores order by downstream information or process reward (Kim et al., 2025a; Peng et al., 2025; Leang et al., 2026; Kim et al., 2026a; Kim & No, 2026; Yang et al., 2026; Bu et al., 2026). The Flexibility Trap links easy-first order to shrinking solution flexibility (Ni et al., 2026). Position-temperature and confidence-remasking studies connect order to text diversity (Olausson et al., 2026; Frkovic et al., 2026), while an any-order interface need not yield arbitrary-order inference (Kim et al., 2026b). These non-robot results make order sensitivity itself non-novel; none tests fixed-checkpoint closed-loop actions or predicate-defined physical modes.

Sampling bias also has more than one source. Informed correctors address approximate discrete-diffusion sampling and distinguish confidence ordering from the theoretical reverse process (Zhao et al., 2025). Zhao & Cai (2026) analyze dependence-induced error when several conditionally correlated tokens are sampled independently, while Cai & Li (2026) establish efficiency guarantees for entropy-budget adaptive decoding. Our exact identity isolates candidate-dependent selection even with one coordinate revealed at a time; it does not imply that all adaptive or confidence-based schedules are biased.

Two action-decoding precedents constrain our claim. DD-VLA defines max-confidence as a position's largest posterior probability and reports 96.8% LIBERO-Goal success versus 95.8% for random order (Liang et al., 2026). Its released code instead ranks the sampled candidate's probability; that sampled-confidence rule is our baseline. WAM-Diff compares a confidence-driven non-temporal schedule named "Random" with forward- and reverse-causal schedules (Xu et al., 2025); its condition is not uniform random selection. Our contribution connects an exact candidate-selection identity to a matched Split/Coupled intervention on physical-mode recovery. This isolates a particular source of distortion, rather than proposing a new diversity metric or claiming the first alternative action schedule.

3 Candidate–Priority Coupling

3.1 An exact selection identity

Why the distribution changes. Drawing a candidate and deciding whether to keep it are different operations. If that decision depends on the value just drawn, some values are more likely to be kept than others: correct probabilities before selection need not remain correct after selection. Exact completion cannot undo this filtering, because it must respect the value already committed. An independent pilot breaks this link: it chooses the position without inspecting the separate value to be written.

Theorem 1 (candidate-dependent selection). Let P be a probability distribution on $\mathcal{X} = \prod_{i=1}^{L} \mathcal{X}_i$, where each $\mathcal{X}_i$ is finite and $P_i(v) > 0$ for all $v \in \mathcal{X}_i$. Draw candidates V with marginals P_i, select one index $I \in \{1, \dots, L\}$, commit V_I, and draw the remainder from $P(X_{-I} \mid X_I = V_I)$, independently of unused candidates given (I, V_I). Let $a_i(v) = \Pr(I = i \mid V_i = v)$ be the chance that position i is selected when its candidate is v. The output law obeys

$$Q(x) = P(x) \sum_i a_i(x_i). \tag{1}$$

Hence $Q = P$ if and only if $\sum_i a_i(x_i) = 1$ for every x with $P(x) > 0$. The proof, including normalization and the zero-mass case, is given in Appendix O.

Corollary (exact sequential decoding). If each step draws fresh exact conditional candidates and selects one position independently of them given the revealed history, the final law is P. The order may depend on revealed values, entropy, or maximum conditional probabilities. Ranking with an independent pilot and committing a separate draw satisfies this condition. Candidate-dependent ordering need not preserve P, but dependence alone does not imply distortion.

A balanced two-token construction makes this bias explicit and shows that imbalance lowers expected finite-budget coverage at matched validity (Appendices B and T). It proves possibility, not mediation in the parallel learned decoder. Figure 2 shows the corresponding intervention: draw independent pilot Y and candidate Z from the same predictions, then rank by $p_i(Y_i)$ while committing Z_i.

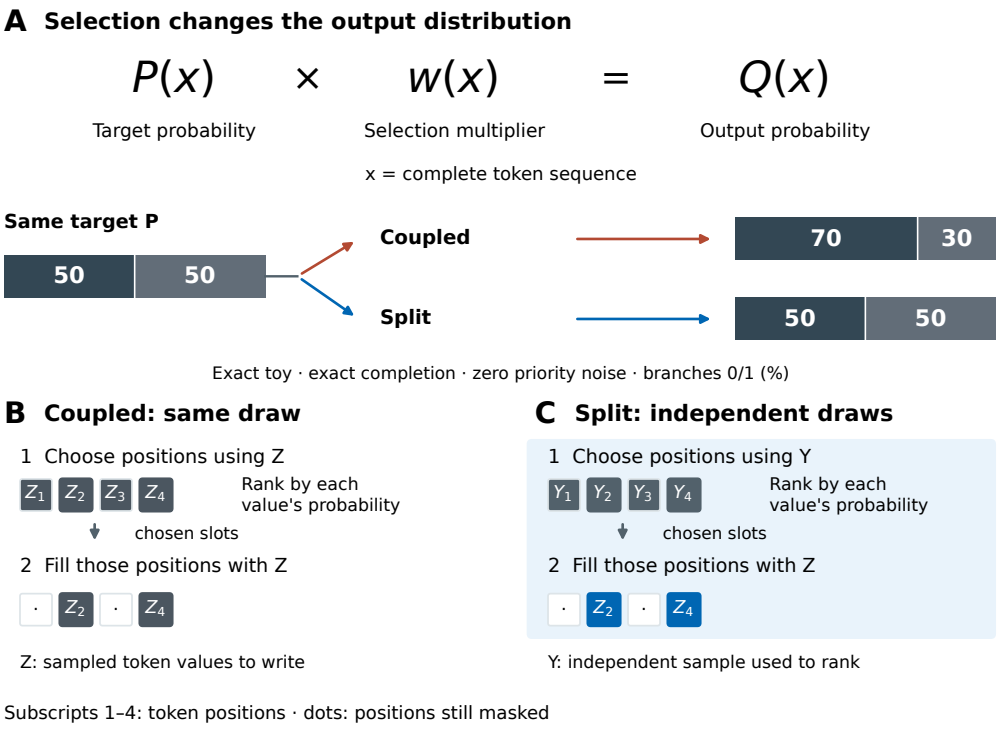


Figure 2: **Candidate-dependent selection reweights the output law.** (A) For a complete token sequence x, selection multiplies target probability $P(x)$ by $w(x)$ to produce output probability $Q(x)$ after one reveal and exact completion. Both paths start from the same target. The exact two-token example in Appendix B, with zero priority noise, gives branch probabilities 70/30 under Coupled and 50/50 under Split, not robot measurements. (B,C) First choose positions by sampled-value probability, then fill them. Z is the candidate draw to write; Y is an independent pilot draw. Coupled uses Z for both operations and leaves Y unused; Split chooses with Y and fills from Z. Subscripts label token positions and dots mark still-masked slots; the four-token diagrams are schematic. The robot comparison matches two draws and 12 forwards. The oracle guarantee does not extend to learned parallel decoding.

3.2 Masked Action Decoding

We use absorbing-mask discrete diffusion (Austin et al., 2021; Shi et al., 2024) with a MaskGIT-style iterative reveal rule (Chang et al., 2022).

Let a tokenized action chunk be $A = (a_{t,d}) \in \mathcal{V}^{T\times D}$, with $T = 8$ action timesteps and $D = 7$ action dimensions. We flatten coordinates into positions $i \in \{1, \dots, L\}$, where $L = TD = 56$. At denoising iteration s, the policy predicts a categorical distribution

$$p_i^{(s)}(v) = p_\theta(a_i = v \mid A_{\neg M_s}, o), \tag{2}$$

for every position still masked by M_s, conditioned on observation o and the currently visible action tokens. Candidate tokens $\tilde{a}_i^{(s)} \sim p_i^{(s)}$ are sampled in parallel. A fixed cosine mask schedule determines *how many* positions remain masked; a commitment scheduler determines *which* positions remain masked.

The released-code confidence rule used in our experiments assigns priority

$$q_i^{\text{conf}} = p_i^{(s)}(\tilde{a}_i^{(s)}), \tag{3}$$

the probability of the sampled token. This differs from Eq. (7) of the current DD-VLA paper, which defines confidence as the top-1 categorical probability. Higher priority commits earlier. Alternatives

use $q_i^{\text{rand}} = 1$, $q_{t,d}^{\text{temp}} = T - t$ for $t = 0, \ldots, T-1$, or entropy $q_i^{\text{uncert}} = -\sum_v p_i^{(s)}(v) \log p_i^{(s)}(v)$. They preserve weights, mask counts, temperature schedule, and random draws. These are stochastic priorities, not deterministic orders: the same annealed Gumbel perturbation remains active. Only the first Choice policy query is intervened on; later queries revert to sampled-confidence-first. Changes to later conditional distributions induced by the visible context are part of the scheduler effect. At zero-indexed iteration s, priorities are clamped and perturbed as

$$\bar{q}_i = \max(q_i, 10^{-20}), \quad r_i = \log(\bar{q}_i + 10^{-20}) + \tau_s G_i, \quad \tau_s = \tau_0 \left(1 - \frac{s+1}{12}\right), \tag{4}$$

where G_i is standard Gumbel, $\tau_0 = 1$, and the lowest scores remain masked. This is a position-selection temperature; categorical candidates are sampled separately from the untempered model probabilities.

3.3 Behavioral endpoints

Let $\mathcal{M}$ be the known valid semantic modes for an observation and let $m(A_k)$ denote the mode achieved by rollout k, if valid. Coverage among K samples is

$$C_K(o) = \frac{1}{|\mathcal{M}|} \left| \{ m \in \mathcal{M} : \exists k \le K,\ m(A_k) = m \} \right|. \tag{5}$$

Task success means achieving at least one goal predicate; mode validity requires a uniquely first-achieved predicate. Failures and ties contribute no mode but remain in the fixed budget. Success@K uses task success rather than mode validity; no discrepancy occurs in the primary rollouts. Coverage measures finite-budget recovery, not model support: non-observation does not imply zero probability. Secondary Jensen–Shannon divergence compares the empirical valid mode distribution with the balanced demonstrations and is undefined if no valid rollout occurs. Entropy and commitment traces are exploratory diagnostics, not evidence of decision timing or mediation (Appendix H).

4 ForkBench-Choice

We build ForkBench-Choice in LIBERO (Liu et al., 2023) with the instruction *put either the alphabet soup or the tomato sauce in the basket*. Either placement satisfies the task. The uniquely first-achieved predicate defines the mode, $\mathcal{M} = \{\text{soup}, \text{tomato}\}$; a tie is mode-invalid. Exact predicates avoid trajectory-clustering ambiguity.

Balanced paired demonstrations. Ten soup and ten tomato demonstrations share ten paired starts, with matched first images, simulator states, and numerically matched proprioception. Every constructed demonstration is replay-verified. Episode-level balance does not imply transition-level balance or a balanced learned distribution (Appendix T.1).

Development and prospectively unopened states. State 13 is for development; states 14–22 form an earlier confirmation. Before opening the remaining default states 23–49, we freeze the terminal step-8,000 checkpoint, $K = 20$, seed 37, decoders, endpoints, and state-level inference with Holm correction. All 27 remaining states are included without outcome-dependent stopping: the target is this fixed census, not a random sample of deployment states (Appendix T.1).

Feasibility and scope. A scripted controller verifies both predicates in all 27 primary states (Appendix A); evaluated Choice rollouts use the closed-loop VLA. Known outcomes make coverage falsifiable, but this controlled testbed does not establish broad task generalization (Section 7).

5 Experimental Design

Policy and training. The released DD-VLA LIBERO implementation (Liang et al., 2026) uses OpenVLA-7B (Kim et al., 2025c) and OpenVLA-OFT action chunking (Kim et al., 2025b). From a 30,000-step two-order LoRA checkpoint, we fine-tune for 8,000 steps on the 20 replay-verified single-object prefixes (rank 32, effective batch 8, learning rate 2×10^{-4}, image augmentation). The configured terminal checkpoint is frozen before states 14–49 are opened; no intermediate is selected on development outcomes. Training details are tabulated in Appendix A.

Paired scheduler intervention. We evaluate SAMPLED-CONF., RANDOM, TEMPORAL, and UNCERTAINTY with 20 rollouts per state, paired seeds, and matched token counts at every iteration: $27 \times 20 \times 4 = 2{,}160$ rollouts. Only first-query commitment priorities differ; later queries use sampled confidence. Remasking is disabled throughout. Full-chunk execution and preprocessing are specified in Appendix A.

Primary and secondary endpoints. The primary endpoint is state-mean Coverage@20, with three paired alternative-minus-confidence contrasts. Secondary endpoints are validity, Success@20, valid-mode JS divergence, entropy–commit correlation, and entropy degradation. Every decoder and failed rollout is retained.

Statistical inference. Inference is conditional on the 27-state panel, not on independent rollouts or tokens. Paired-state bootstrap intervals use 10,000 resamples. Two-sided sign-flip tests omit zeros and enumerate up to 20 nonzero signs; otherwise they use 10,000 Monte Carlo assignments with plus-one correction. Holm adjusts the three alternatives within each endpoint family. Tests condition on the nonzero magnitudes and assume independent equiprobable signs under the null; neither tests nor intervals imply superpopulation inference. Token histories are descriptive.

Same-draw selection-coupling control. The intervention in Figure 2 matches two candidate draws, position noise, mask counts, model calls, and later-query decoding; random and max confidence are additional two-draw controls. The frozen study uses 27 states, $K = 10$, and both fixed fits (2,160 rollouts). Split-minus-coupled Coverage@10 contrasts use 100,000 paired-state bootstrap draws, exact sign-flip tests, and Holm correction across fits, with a confidence-bound validity guard and separate proximal-score family. Fits are not pooled; fit 17 remains technically qualified (Appendix P).

Supporting studies. Trace diagnostics, the token–timestep ablation, and the six-dose position-noise study with separate random anchor are specified in Appendices G, H, R and T.

Fresh Approach-Side validation. Separately, a frozen VLA chooses between left and right fixed-controller routes in 54 fresh contexts, using the same four two-draw decoders and ten proposals per condition/context. Both routes must pass simulator validation before a context is eligible; model outputs do not select the panel. The sign of the first decoded action's y coordinate selects the route. The frozen primary endpoint is unbiased pairwise side diversity, $D = 2n_L n_R/[K(K-1)]$, with $K = 10$, rather than Coverage@10. It was chosen using the earlier completed panel and frozen before fresh-panel generation. We use 100,000 paired-context bootstrap draws and one million two-sided sign flips with plus-one correction. These are learned route choices, not new closed-loop policy executions (Appendix S).

Compute. Training and evaluation use one NVIDIA RTX 5090. Scheduler alternatives add no model calls; they replace only the Gumbel top-k priority. Artifact contents, exclusions, and regeneration are documented in Appendix M.

6 RESULTS

6.1 SAMPLED-CONFIDENCE-FIRST DECODING REDUCES VALID-MODE RECOVERY

Figure 3 shows the central result (exact table: Table 17). Sampled-confidence-first decoding reaches only 0.593 mean Coverage@20. Random and temporal-forward each reach 0.833, a paired gain of 0.241; uncertainty-first reaches 0.815, a gain of 0.222. The state bootstrap intervals exclude zero for all three alternatives: random $[0.148, 0.333]$, temporal-forward $[0.130, 0.352]$, and uncertainty-first $[0.111, 0.333]$. Under nonzero-difference sign exchangeability, the enumerated sign-flip Holm-adjusted p-values are 0.00073, 0.00342, and 0.00418. All three paired gains meet the prespecified Holm-adjusted $\alpha = .05$ criterion (largest adjusted $p = .00418$). Because no noninferiority margin was prespecified for primary validity, competence is evaluated descriptively below.

The effect is distributed across states. Random improves coverage in 13 of 27 states and never worsens it; temporal-forward improves 13 and worsens one; uncertainty-first improves 14 and worsens

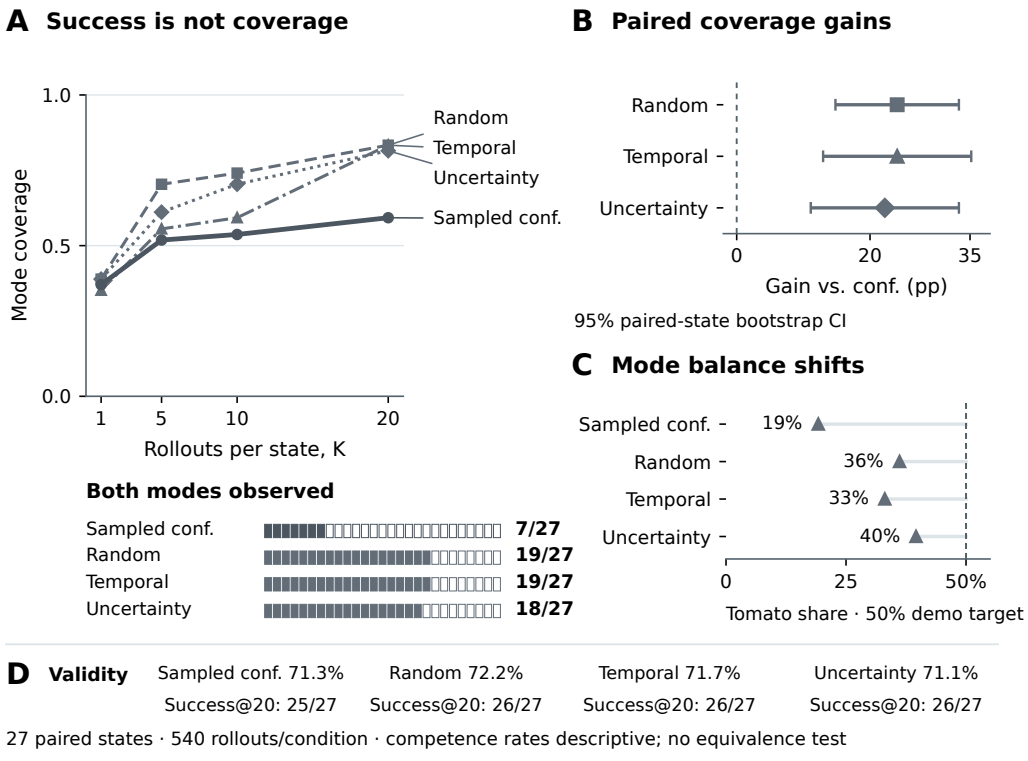


Figure 3: **Finite-budget mode coverage separates from task competence.** (A) Realized coverage prefixes and both-mode counts; only $K = 20$ was prespecified. (B) Paired gains with 95% state-bootstrap intervals over 27 fixed states. (C) Pooled valid-rollout tomato shares versus the 50% demonstration target; states are weighted by valid counts. (D) Descriptive competence, not equivalence. All 540 rollouts per decoder, including failures, enter the coverage budget.

two. Random and temporal recover both modes in 19 states, compared with only 7 for sampled-confidence-first—a 2.71× increase. Figure 9 exposes every rollout.

6.2 Task competence under matched decoding budgets

Validity is 71.3% for sampled confidence and 72.2/71.7/71.1% for the three alternatives; paired differences range from $-.002$ to .009, with intervals spanning zero and no secondary-family rejection. Success@20 is 25/27 versus 26/27. These estimates do not establish equivalence. Pooled mode shares and complete-case JS move toward the balanced demonstrations, providing additional validity-conditioned sensitivity analyses (Figure 3 and Appendix J).

6.3 Breaking same-draw coupling improves recovery on both fits

On the primary fit, split confidence raises Coverage@10 from .537 to .741, a paired gain of .204 [.111, .296] (Holm $p = 0.00195$). On recovered fit 17, coverage rises from .630 to .759, a gain of .130 [.056, .222] (Holm $p = 0.01562$). The corresponding validity differences are .044 [-0.007, .096] and .007 [-0.019, .033], whose lower bounds pass the prespecified $-.05$ guard. Split and coupled have identical Success@10 within each fit (25/27 primary; 26/27 recovered). Both fits meet the frozen gate; the separate proximal-score family supports immediate selection tilt, not behavioral mediation, and fit 17 retains its qualification.

Max confidence has the highest primary-fit coverage (.778), whereas random is highest on fit 17 (.852). Split tests a specific mechanistic contrast: ranking with the committed draw Z_i versus an independent pilot Y_i. Letting Z_i promote itself reduces recovery relative to the independent pilot under matched random-number and model-call budgets (Appendix P).

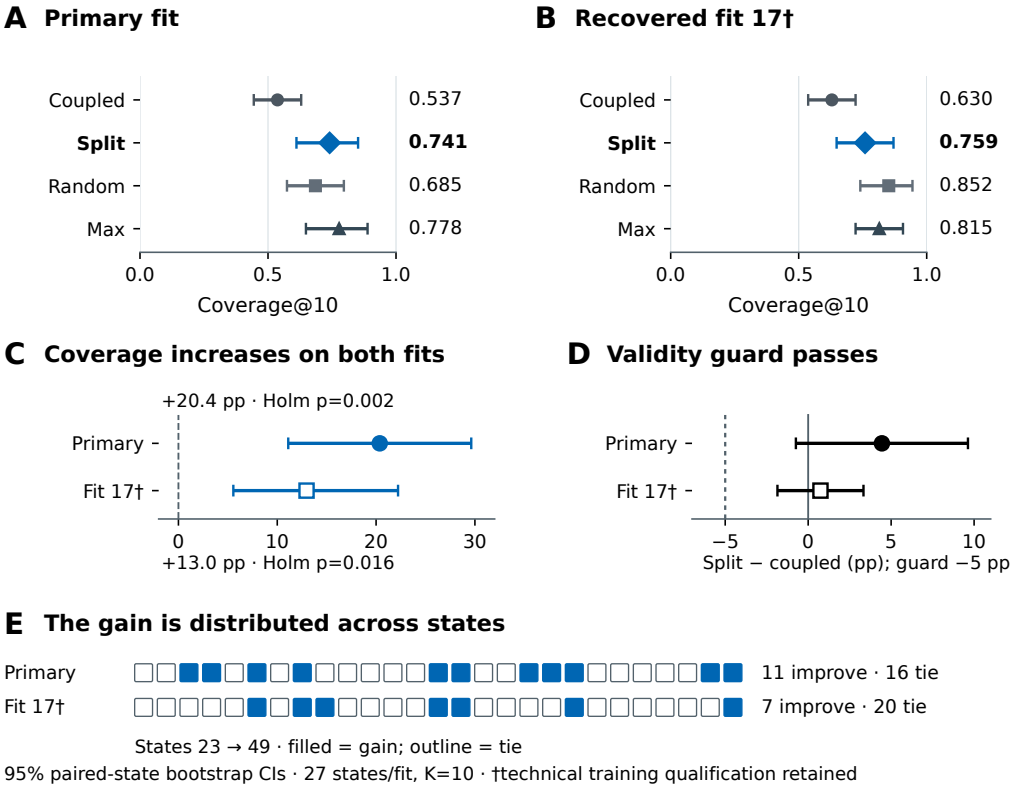


Figure 4: **Breaking same-draw coupling improves coverage.** (A,B) All four two-draw decoders; fit 17 is technically qualified. (C) Coverage gains pass the two-fit Holm family. (D) Validity guard. Whiskers: 95% paired-state bootstrap intervals, 27 states per fit, $K = 10$. (E) All state-paired changes. Fits are analyzed separately; random and max-confidence anchors are shown in (A,B).

6.4 Fresh route choices confirm the diversity gain

Split increases pairwise side diversity from 0.245 to 0.378, a paired gain of +0.133 [0.094, 0.173] ($p \approx 10^{-6}$), passing the frozen fresh-panel criterion (Figure 5). Coverage@10 is 0.870 versus 0.954. Route validity is guaranteed by panel construction; the measured outcome is learned route-selection diversity. Split-minus-random and split-minus-max diversity intervals include zero (Appendix S). The matched split–coupled contrast confirms the selection effect in a distinct route-choice policy on fresh contexts.

6.5 Additional decoder and mechanism controls

The seed-17 scheduler comparison retains the coverage direction, but undefined detached diagnostics and a later verification kernel make it a qualified sensitivity, not a clean replication (Appendix N). The full max-confidence-with-remasking control gives a coverage gain of .056 $[-.019, .130]$, $p = .375$ (Appendix K). Position noise has a positive ordinal trend with a nonmonotone dose response, characterizing decoding sensitivity (Appendix R). A frozen semantic-axis control is null (Appendix Q); exploratory traces remain associational (Figure 7 and Appendix H).

7 Discussion and Conclusion

Decoding is part of the policy. Similar task success can hide coverage loss. The exact identity, controlled Choice intervention, and fresh Approach-Side choices connect value-dependent selection to behavioral coverage. A decoder decides which predictions become actions; this selection can

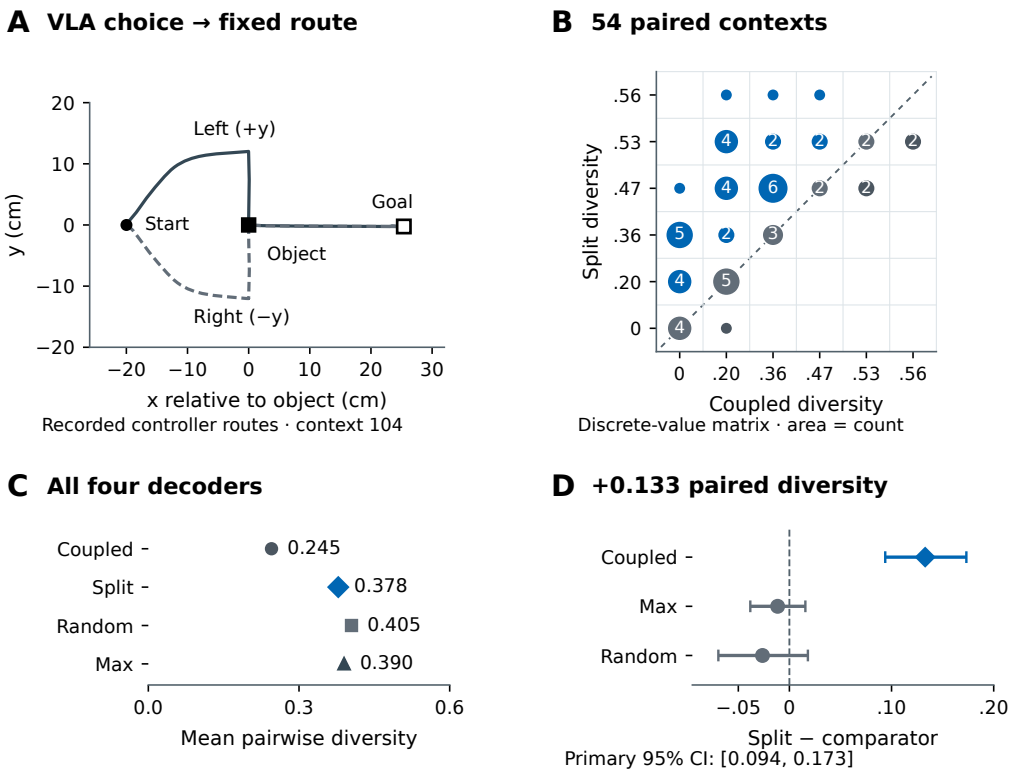


Figure 5: **Split increases diversity of fresh Approach-Side choices.** (A) Recorded controller paths, lowest context ID. (B) All 54 contexts; discrete estimator categories, bubble area proportional to count. (C) All decoder means. (D) Paired-context bootstrap 95% intervals; coupled is the primary comparator, random/max are secondary and unadjusted. The 2,160 VLA proposals select among 108 prevalidated fixed-controller routes.

alter the distribution even when the predictions are exact. Repeated-sampling coverage makes that consequence measurable.

Limitations. The experiments use task-adapted simulation policies: Choice supplies closed-loop rollouts, whereas Approach-Side measures VLA selection among fixed-controller routes. The Choice fits share a state panel, and fit 17 carries technical recovery qualifications. The full max-confidence/remasking control has a coverage-gain interval spanning zero. Further scope and mechanism details are reported in Appendix L.

Conclusion. Candidate–priority coupling can reweight an exact decoded joint, and removing it improves finite-budget behavioral coverage in our controlled robot tests. Preserving a policy's behavioral repertoire requires attention to how predictions are selected for commitment, alongside how they are learned. Selection rules warrant evaluation alongside task success.

AI USE STATEMENT

OpenAI Codex was used only as an auxiliary tool. Generative AI assisted with design exploration, hypothesis framing, implementation, checks, and text editing. All scientific decisions, problem formulation, results, and conclusions belong to the human authors. AI assistance also included literature retrieval and discovery, drafting manuscript sections, and writing, expanding, and checking mathematical proofs.

Empirical measurements were computed from recorded outputs of the frozen policy and simulator, not invented or imputed by the writing assistant. The evaluated VLA generates actions as part of the experiment. The writing assistant did not create the training data. No AI-generated images appear in the manuscript.

All authors confirm personal responsibility for the final text, claims, code, analyses, and visual artifacts.

ETHICS STATEMENT

This study evaluates simulated manipulation policies and involves no human participants, personal data, physical robot deployment, or autonomous action in the world. Behavioral diversity is not intrinsically beneficial: in a deployed system, alternate modes may differ in safety, cost, or human preference, so our balanced two-mode diagnostic must not be read as a prescription to maximize diversity indiscriminately. We report validity separately and do not establish real-world safety or noninferiority. The review artifact excludes model weights, demonstrations, datasets, simulator assets, and inherited runtime code; its source subset cannot be released until accountable file-level provenance and license review confirms redistribution rights. Authors remain responsible for compliance with upstream software, model, and dataset terms.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the decoder interventions, training, endpoints, and inference. Appendix O provides the proofs; the study appendices report protocols, qualifications, and state-level evidence. Local replay scripts reconstruct the reported behavioral endpoints from recorded outcomes, and figure manifests bind displayed values to their sources. This is endpoint replay, not independent retraining or a repeat of the simulator experiments. No downloadable artifact is claimed here: redistribution of the candidate code/data package has not been approved. Appendix U records its exact contents, excluded dependencies, and the separate status of the fresh Approach-Side and figure packages.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html.

Dake Bu, Wei Huang, Andi Han, Hau-San Wong, Qingfu Zhang, Taiji Suzuki, and Atsushi Nitanda. DPRM: A plug-in Doob h-transform-induced token-ordering module for diffusion language models, 2026. URL https://arxiv.org/abs/2604.24357. arXiv preprint; presented at the ICML 2026 FoGen workshop.

Changxiao Cai and Gen Li. Confidence-based decoding is provably efficient for diffusion language models. arXiv:2603.22248, 2026. URL https://arxiv.org/abs/2603.22248. Preprint.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chang_MaskGIT_Masked_Generative_Image_Transformer_CVPR_2022_paper.html.

Jiayi Chen, Wenxuan Song, Jiaxin Fang, Ruiqing Yin, Jingbo Wang, Shuai Chen, Jieyuan Pei, Yikai Qin, Feifan Chen, Haodong Yan, Zhide Zhong, Wen Chen, Yan Wang, Yuxiang Gao, and Haoang Li. DFM-VLA: Iterative action refinement for robot manipulation via discrete flow matching, 2026. URL https://arxiv.org/abs/2603.26320. arXiv preprint.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

Stipe Frkovic, Metod Jazbec, Dan Zhang, Christian A. Naesseth, Ilija Bogunovic, and Eric Nalisnick. Re-evaluating confidence remasking in masked diffusion language models, 2026. URL https://arxiv.org/abs/2606.12232. arXiv preprint.

Chengyang He, Xu Liu, Gadiel Sznaier Camps, Joseph Bruno, Guillaume Sartoretti, and Mac Schwager. Demystifying robot diffusion policies: Action memorization and a simple lookup table alternative. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10890–10917, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/122ea6470232ee5e79a2649243348005-Abstract-Conference.html.

Jindou Jia, Tuo An, Yuxuan Hu, Gen Li, Jingliang Li, Bohan Hou, Xiangyu Chen, Jiaqi Bai, Bofan Lyu, and Jianfei Yang. MARS Policy: Multimodality only when it matters, 2026. URL https://arxiv.org/abs/2605.29766. arXiv preprint.

Xiaogang Jia, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. Towards diverse behaviors: A benchmark for imitation learning with human demonstrations. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 41666–41686, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b63cb503757f9f31e5d74e3567dd160a-Abstract-Conference.html.

Dueun Kim and Albert No. The confidence shortcut: A reasoning failure mode of masked diffusion models, 2026. URL https://arxiv.org/abs/2605.29123. arXiv preprint.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30749–30768. PMLR, 2025a. URL https://proceedings.mlr.press/v267/kim25ah.html.

Jiyeon Kim, Sungik Choi, Yongrae Jo, Moontae Lee, and Minjoon Seo. Early decisions matter: Proximity bias and initial trajectory shaping in non-autoregressive diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. doi: 10.48550/arXiv.2604.10567. URL https://arxiv.org/abs/2604.10567.

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*, Los Angeles, CA, USA, June 2025b. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025c. URL https://proceedings.mlr.press/v270/kim25c.html.

Seunggeun Kim, Jaeyeon Kim, Taekyun Lee, Yuyuan Chen, Yilun Du, Sham Kakade, and Sitan Chen. From interface to inference: Eliciting any-order inference from any-order models, 2026b. URL https://arxiv.org/abs/2607.26504. arXiv preprint.

Yilun Kong, Yunpeng Qing, Guozheng Ma, Haoyu Wang, Li Shen, Zhi Hou, and Dacheng Tao. ExToken: Structured exploration for efficient vision-language-action reinforcement fine-tuning, 2026. URL https://arxiv.org/abs/2607.12931. arXiv preprint.

Joshua Ong Jun Leang, Yu Zhao, Mihaela Catalina Stoian, Wenda Li, Shay B. Cohen, and Eleonora Giunchiglia. Can i have your order? monte-carlo tree search for slot filling ordering in diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2602.12586. URL https://arxiv.org/abs/2602.12586.

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26991–27008. PMLR, 2024. URL https://proceedings.mlr.press/v235/lee24y.html.

Sung-Wook Lee, Xuhui Kang, and Yen-Ling Kuo. TBD-VLA: Temporal block diffusion vision language action model, 2026. URL https://arxiv.org/abs/2606.07895. arXiv preprint.

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liuao Pei, Tian Nian, Shunbo Zhou, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2508.20072. URL https://arxiv.org/abs/2508.20072.

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*, volume 36, pp. 44776–44791. Curran Associates, Inc., 2023. doi: 10.52202/075280-1939. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Yang Liu, Pengxiang Ding, Tengyue Jiang, Xudong Wang, Wenxuan Song, Minghui Lin, Han Zhao, Hongyin Zhang, Zifeng Zhuang, Wei Zhao, Siteng Huang, Jinkui Shi, and Donglin Wang. MMaDA-VLA: Large diffusion vision-language-action model with unified multi-modal instruction and generation, 2026. URL https://arxiv.org/abs/2603.25406. Accepted at ACM Multimedia 2026.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, and Gao Huang. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2601.15165. URL https://arxiv.org/abs/2601.15165. Outstanding Paper.

Theo X. Olausson, Metod Jazbec, Xi Wang, Armando Solar-Lezama, Christian A. Naesseth, Stephan Mandt, and Eric Nalisnick. A tale of two temperatures: Simple, efficient, and diverse sampling from diffusion language models, 2026. URL https://arxiv.org/abs/2604.09921. Accepted at the Conference on Language Modeling (COLM).

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling, 2025. URL https://arxiv.org/abs/2502.03540. arXiv preprint.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103131–103167. Curran Associates, Inc., 2024. doi: 10.52202/079017-3277. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Xiufeng Song, Yiran Qin, Yan Tai, Li Kang, Heng Zhou, Siqi Luo, Jiwen Yu, Ling Yang, Philip Torr, and Lei Bai. DIVA: Discrete diffusion vision-language-action models for parallelized action

generation. OpenReview preprint, 2026. URL https://openreview.net/forum?id=mNya9d1DA2. Submitted to ICLR 2026.

Derek Ming Siang Tan, Shailesh Shailesh, Srikrishna Iyer, William Wei Jie Teo, Yuanliang Ju, Qiao Gu, and Guillaume Sartoretti. RL^2-VLA: Adaptive RL latent compositional steering with test-time scaling for vision-language-action models, 2026. URL https://arxiv.org/abs/2607.26991. arXiv preprint.

Ruiheng Wang, Shuanghao Bai, Haoran Zhang, Badong Chen, and Xiangyu Xu. BlockVLA: Accelerating autoregressive VLA via block diffusion finetuning, 2026. URL https://arxiv.org/abs/2605.13382. arXiv preprint.

Yuqing Wen, Hebei Li, Kefan Gu, Yucheng Zhao, Tiancai Wang, and Xiaoyan Sun. LLaDA-VLA: Vision language diffusion action models, 2025. URL https://arxiv.org/abs/2509.06932. arXiv preprint.

Jinhao Wu, Shiduo Zhang, Yicheng Liu, Xiaopeng Yu, Sixian Li, Siyin Wang, Hang Zhao, Jing Huo, Yang Gao, Jingjing Gong, Xipeng Qiu, and Yu-Gang Jiang. Coarse-to-control: Action-token planning for vision-language-action models, 2026a. URL https://arxiv.org/abs/2606.07107. arXiv preprint.

Yuhao Wu, Yitian Liu, Weijie Shen, Mishuo Han, Wenjie Xu, Haotian Liang, Zhongshan Liu, Yinan Mao, Lei Xu, Xinping Guan, Ru Ying, Ran Zheng, Wei Sui, Xiaokang Yang, Wenbo Ding, and Yao Mu. dVLA-RL: Reinforcement learning over denoising trajectories for discrete diffusion vision-language-action models, 2026b. URL https://arxiv.org/abs/2606.23623. arXiv preprint.

Mingwang Xu, Jiahao Cui, Feipeng Cai, Hanlin Shang, Zhihao Zhu, Shan Luan, Yifang Xu, Neng Zhang, Yaoyi Li, Jia Cai, and Siyu Zhu. WAM-Diff: A masked diffusion VLA framework with MoE and online reinforcement learning for autonomous driving, 2025. URL https://arxiv.org/abs/2512.11872. arXiv preprint.

Kaisen Yang, Jayden Teoh, Kaicheng Yang, Yitong Zhang, and Alex Lamb. Improving sampling for masked diffusion models via information gain. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2602.18176. URL https://arxiv.org/abs/2602.18176.

Yixiu Zhao, Jiaxin Shi, Feng Chen, Shaul Druckmann, Lester Mackey, and Scott Linderman. Informed correctors for discrete diffusion models. In *Advances in Neural Information Processing Systems 38*, pp. 139126–139154. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-4183.

Yunxiao Zhao and Changxiao Cai. Adaptation to intrinsic dependence in diffusion language models. arXiv:2602.20126, 2026. URL https://arxiv.org/abs/2602.20126. Preprint.

A Frozen Protocol and Implementation Details

Table 1: Primary expansion configuration frozen before inspecting states 23–49.

Component	Fixed value
Backbone	OpenVLA-7B / DD-VLA implementation
Checkpoint	Choice fine-tune, step 8,000
Inputs	Agent-view and wrist images, proprioception, center crop
Action chunk	8 timesteps $\times 7$ dimensions; full chunk between replans
Denoising	12 iterations, cosine mask count, choice temperature 1.0
Action interface	256-bin normalized tokens; checkpoint statistics for unnormalization
Episode	5 settling steps, then at most 520 control steps; predicate/native termination
Remasking	Disabled
Initial states	23–49 (27 states)
Samples	$K = 20$ per state and scheduler
Schedulers	sampled confidence, random, temporal, uncertainty
Later queries	sampled confidence, token granularity
Replicate block	seed 37; rollout seeds paired within state
Primary endpoint	valid-mode Coverage@20
Inference unit	physical initial state
Correction	Holm across three alternatives

The baseline matches the released DD-VLA LIBERO code at upstream commit `d2efbc4`: it ranks the probability assigned to the *sampled* token (permalink). It is distinct from the max-confidence rule in Eq. (7) of the current paper and from the top-1/top-2 gap. We use "sampled-confidence-first" throughout for this reason. Random order assigns equal priority before the same annealed Gumbel perturbation. Temporal-forward assigns higher priority to earlier action timesteps. Uncertainty-first uses categorical entropy as priority. Infinite priority prevents a committed token from being remasked.

All traces cover observation index zero of each episode—the query at which the alternative scheduler is active. Each policy input contains the resized, center-cropped agent-view and wrist images plus normalized proprioception. The checkpoint's `forkbench_order` statistics unnormalize decoded actions; the gripper command is binarized and sign-corrected for LIBERO. After five dummy steps for scene stabilization, control alternates a policy query with execution of all eight decoded actions, unless a valid soup/tomato predicate or the native environment termination occurs first. The task allows at most 520 subsequent control steps. Thus control is closed loop across action chunks, with a full eight-step open-loop interval between replans. Later queries use token-level sampled-confidence-first decoding with the same 12-step cosine schedule and choice temperature.

Primary-state mode feasibility. In a post-primary benchmark audit, a deterministic scripted controller placed each intended object first from every state 23–49. Each resulting first-object action sequence also reproduced its intended predicate in one open-loop replay from the recorded post-settle simulator state. Thus both ForkBench-Choice modes were directly verified in all 27 states (54/54 state–mode checks); this audit used no policy checkpoint and cannot explain the observed scheduler effect.

B Sample-Dependent Order Can Bias Exact Conditionals

The released rule chooses a position after drawing candidate values at every masked coordinate. This sample dependence can change the model distribution even when all categorical predictions are exact.

Proposition (sample-dependent-order bias). Consider two binary tokens: a behavior branch B and context C. For any $\alpha \in (1/2, 1)$, define the joint distribution

$$\begin{array}{c|cc} & C=0 & C=1 \\ \hline B=0 & 1/2 & 0 \\ B=1 & \alpha - 1/2 & 1-\alpha. \end{array} \tag{6}$$

Table 2: Audited reconstruction of the two-stage training and exact-axis evaluation configuration. Evidence tags resolve to the read-only provenance audit included with the artifact. "Unavailable" denotes missing historical evidence, not an inferred difference.

Component	Audited configuration	Evidence / status
Pipeline	OpenVLA-7B (revision `47a0ec7...`) → full-order LoRA, 30k steps → Choice-prefix LoRA, 8k steps.	E02, E06–E08
Data and training states	Ten paired states (IDs 0,2,3,4,5,6,8,9,11,12): 20 full-order trajectories (8,762 transitions) and 20 derived Choice prefixes (3,744 transitions; 10 per mode).	E04–E05
Objective and inputs	Discrete diffusion; L1, continuous diffusion, and FiLM off; agent-view plus wrist images and proprioception.	E07–E08
Batch and shuffle	Order: microbatch 8, accumulation 1; Choice: microbatch 1, accumulation 8; effective batch 8 and shuffle buffer 10,000 in both stages.	E07–E08
LR schedule	Order: 5×10^{-4}, then 5×10^{-5} from step 15k; Choice: 2×10^{-4}, then 2×10^{-5} from step 4k; no warm-up. Historical optimizer family/defaults: *unavailable*. Seed-17 source: AdamW + one-milestone MultiStepLR, $\gamma = .1$.	E07–E08, E14, E27
LoRA	Rank 32, $\alpha = 16$, dropout 0, Gaussian initialization, no bias; all linear layers (the realized primary module-name set is recorded in the audit).	E09–E10, E14
Augmentation	Enabled in both primary stages. Exact historical operators: *unavailable*. Seed-17 freezes scale-.9 square crop, brightness .2, contrast/saturation [.8,1.2], and hue .05.	E07–E08, E15–E16
Training RNG	Historical primary: seed 7 in both stages. Prospective checkpoint: seed 17 in both stages. Byte-identical historical training execution: *unavailable*.	E07–E08, E17, E20–E21
Checkpoint rule	Fixed final order step 30k and final Choice step 8k; no evaluation-based checkpoint selection.	E02–E03, E17
State roles	Training/reference: listed above; development and K=20 screening: state 13; earlier confirmation: 14–22; primary and seed-17 exact-axis panel: 23–49.	E01–E03, E17
Simulator and episode	LIBERO-10 task 0; either-object instruction; 256-pixel environment images, 224-pixel policy inputs; **5** settling actions; full 8-action chunks; at most 520 control steps. Historical run-local simulator/GL receipt: *unavailable*; seed-17 freezes EGL.	E01, E11, E22–E25
Decoder and budget	First-query token scheduling; sampled-confidence follow-up; 12 cosine iterations; remasking off; four schedulers; seed 37, samples 0–19, $K = 20$ on 27 states (2,160 rollouts/checkpoint).	E03, E17, E20, E22–E23
Cross-checkpoint claim	Recorded hyperparameters match, but historical source/environment and restart artifacts are incomplete. Protocol v5 therefore tests robustness across one fresh fitted checkpoint under matched recorded hyperparameters; a training-RNG-only claim would require a fresh seed-7 companion under the same source/environment closure.	E07–E08, E13–E17, E20

Both behavior modes have probability $1/2$, while $P(C = 0) = \alpha$. Starting with both positions masked, draw candidates $\widetilde{B} \sim P(B)$ and $\widetilde{C} \sim P(C)$ independently, score each position by its sampled value's marginal probability, select one with additive Gumbel perturbation, and redraw the other token from its exact conditional. Under the implementation-faithful position score $\log(\max\{q, \delta\} + \delta) + \tau G$, with $\delta = 10^{-20}$, at any finite position-selection scale $\tau > 0$ (categorical candidates are sampled separately from untempered probabilities), the decoded branch marginal satisfies $P_{\mathrm{decode}}(B = 0) > 1/2$.

Proof. The marginals of Equation (6) are $P(B = 0) = P(B = 1) = 1/2$, $P(C = 0) = \alpha$, and $P(C = 1) = 1 - \alpha$. Thus either branch candidate has score $1/2$. For $u \in (0, 1)$, let $s_{\tau,\delta}(u)$ denote the probability that a context candidate of marginal probability u is selected instead of the branch candidate. For independent standard Gumbel variables G_C, G_B, this probability is

$$s_{\tau,\delta}(u) = \Pr\big(\log g_\delta(u) + \tau G_C > \log g_\delta(1/2) + \tau G_B\big),$$
$$g_\delta(u) = \max\{u, \delta\} + \delta.$$

To evaluate it, put $d = \tau^{-1} \log(g_\delta(u)/g_\delta(1/2))$. The standard Gumbel cumulative distribution and density are $F_G(z) = \exp(-e^{-z})$ and $f_G(z) = e^{-z} \exp(-e^{-z})$. Conditioning on $G_C = z$ and substituting $t = e^{-z}$ yields

$$\begin{aligned} s_{\tau,\delta}(u) &= \int_{-\infty}^{\infty} F_G(z + d) f_G(z)\, dz \\ &= \int_0^{\infty} \exp\big(-(1 + e^{-d})t\big)\, dt = \frac{1}{1 + e^{-d}}. \end{aligned}$$

Consequently,

$$g_\delta(u) = \max\{u, \delta\} + \delta, \qquad s_{\tau,\delta}(u) = \frac{g_\delta(u)^{1/\tau}}{g_\delta(u)^{1/\tau} + g_\delta(1/2)^{1/\tau}}. \tag{7}$$

Let $s_0 = s_{\tau,\delta}(\alpha)$ and $s_1 = s_{\tau,\delta}(1-\alpha)$. The completion probabilities are $P(B=0 \mid C=0) = 1/(2\alpha)$ and $P(B=0 \mid C=1) = 0$. If the branch position is selected, its candidate remains fair: the selection event depends on the context candidate and the independent Gumbel variables, but not on the value of the branch candidate. Conditioning on the context candidate and the selected position gives

$$\begin{aligned} P_{\text{decode}}(B=0) &= \alpha\left[s_0\frac{1}{2\alpha} + (1-s_0)\frac{1}{2}\right] + (1-\alpha)\left[s_1\cdot 0 + (1-s_1)\frac{1}{2}\right] \\ &= \frac{s_0}{2} + \frac{\alpha(1-s_0)}{2} + \frac{(1-\alpha)(1-s_1)}{2} \\ &= \frac{1}{2} + \frac{1-\alpha}{2}(s_0 - s_1). \end{aligned}$$

Equivalently,

$$P_{\text{decode}}(B=0) = \frac{1}{2} + \frac{1-\alpha}{2}\left[s_{\tau,\delta}(\alpha) - s_{\tau,\delta}(1-\alpha)\right] > \frac{1}{2}. \tag{8}$$

The inequality is strict for every finite $\tau > 0$ because the required comparisons satisfy $s_{\tau,\delta}(\alpha) > s_{\tau,\delta}(1/2) > s_{\tau,\delta}(1-\alpha)$, even though the clamp makes the function flat below δ. This proves the strict inequality. □

Zero-noise limit and numerical specialization. Since $g_\delta(\alpha) > g_\delta(1/2) > g_\delta(1-\alpha)$,

$$\lim_{\tau\downarrow 0} s_{\tau,\delta}(\alpha) = 1, \qquad \lim_{\tau\downarrow 0} s_{\tau,\delta}(1-\alpha) = 0.$$

Substitution in Equation (8) gives $P_{\text{decode}}(B=0) = 1 - \alpha/2$; at $\alpha = 3/5$, an exactly balanced model is decoded as 70/30. At the first-iteration temperature used in our 12-step decoder, $\tau = 11/12$, the same example gives 0.5220/0.4780.

For $\alpha = 3/5$ and zero priority noise, the two selection events are

Candidate event	Probability	Committed position	$\Pr(B=0 \mid \text{event})$
$\widetilde{C}=0$	$3/5$	C	$5/6$
$\widetilde{C}=1$	$2/5$	B	$1/2$

The last column refers to the completed output. Hence $P_{\text{decode}}(B=0) = (3/5)(5/6) + (2/5)(1/2) = 7/10$. The context value 0 is committed whenever proposed; the context value 1 is never committed first. This value-dependent selection accounts for the shift without any error in the conditional distributions.

Finite-budget coverage corollary. Let $K \geq 2$ independent, identically distributed rollouts have validity probability $v \in (0,1]$. Conditional on validity, let the two branch probabilities be p and $1-p$, where $p \in [0,1]$. Then expected two-mode Coverage@K is

$$F(p) = 1 - \frac{(1-vp)^K + [1-v(1-p)]^K}{2}, \tag{9}$$

and its unique maximizer at fixed v and K is $p = 1/2$.

Proof. Let H_0 and H_1 indicate that at least one valid rollout realizes the respective branch. Then $C_K = (H_0 + H_1)/2$. Independence across rollouts gives

$$\Pr(H_0 = 0) = (1-vp)^K, \qquad \Pr(H_1 = 0) = [1-v(1-p)]^K.$$

Since $\mathbb{E}[H_j] = 1 - \Pr(H_j = 0)$, taking expectations proves Equation (9). Differentiation gives

$$F'(p) = \frac{Kv}{2}\left[(1-vp)^{K-1} - (1-v+vp)^{K-1}\right].$$

Both bases are nonnegative. Their difference is $v(1-2p)$, and $z \mapsto z^{K-1}$ is strictly increasing on $[0,\infty)$. Therefore $F'(p) > 0$ for $p < 1/2$, $F'(1/2) = 0$, and $F'(p) < 0$ for $p > 1/2$. This proves uniqueness of the maximum. □

Full-support robustness. The effect does not rely on the zero cell in Equation (6): replacing its rows by $(1/2-\epsilon, \epsilon)$ and $(\alpha - 1/2 + \epsilon, 1-\alpha-\epsilon)$ gives full support for $0 < \epsilon < (1-\alpha)/2$, preserves both marginals, and changes the positive bias term in Equation (8) to $((1-\alpha)/2 - \epsilon)\,[s_{\tau,\delta}(\alpha) - s_{\tau,\delta}(1-\alpha)] > 0$. To verify this expression, the modified completion probabilities are $(1/2-\epsilon)/\alpha$ and $\epsilon/(1-\alpha)$. Substituting them into the same two-event decomposition gives

$$\begin{aligned} P_{\text{decode}}(B=0) &= \alpha\left[s_0\frac{1/2-\epsilon}{\alpha} + \frac{1-s_0}{2}\right] + (1-\alpha)\left[s_1\frac{\epsilon}{1-\alpha} + \frac{1-s_1}{2}\right] \\ &= \frac{1}{2} + \left(\frac{1-\alpha}{2} - \epsilon\right)(s_0 - s_1). \end{aligned}$$

Both factors in the added term are strictly positive in the stated range. In contrast, any position order chosen independently of the uncommitted candidate values samples either $P(B)P(C \mid B)$ or $P(C)P(B \mid C)$ and preserves the 50/50 branch marginal; this includes uniform-random order and, for Equation (6), top-1 confidence. Thus exact conditional calibration is not sufficient to make sampled-candidate-dependent order harmless. The construction includes finite Gumbel perturbation but omits parallel multi-token commits, longer sequences, and model error; it establishes a possibility result rather than explaining the empirical effect size. Exact zero-temperature enumeration, a generic finite-temperature enumerator for the numerically indistinguishable $\delta = 0$ idealization, and the full-support perturbation are implemented in `analysis/toy_selection_bias.py` and covered by the test suite.

C Why the Original Two-Object Task Was Repaired

The initial benchmark required both objects to be placed in the basket and treated soup→tomato and tomato→soup as valid modes. On held-out state 13, 400 rollouts across four schedulers produced 23 full successes: 11 sampled-confidence, 6 random, 5 temporal-forward, and 1 uncertainty-first. Every success was soup-first. This result could have been decoder collapse, model collapse, or brittle long-horizon control, so it was not used as confirmatory evidence.

We audited the ambiguity before changing the task. Paired demonstrations had identical first images and simulator states, proprioception equal to numerical precision, balanced 10/10 mode counts, and similar transition counts. Newly generated held-out expert trajectories for both orders replayed successfully three times each. Objective-matched teacher forcing assigned similar likelihood to both modes: across training initial states, soup and tomato negative log-likelihoods were 0.237 and 0.282 nats/token; on held-out full trajectories, they were 0.546 and 0.474. Finally, an eight-action verified tomato expert prefix moved the frozen policy's next predicted chunk to 0.0117 xyz RMS from the tomato expert and completed the tomato predicate. In a one-state, one-sample post-hoc diagnostic, expert prefixes of 224, 288, and 400 actions completed the full tomato-first order, whereas a prefix of 352 actions did not. The checkpoint assigned similar teacher-forced likelihood to both orders and could complete tomato-first when supplied an expert prefix, although this did not establish autonomous recovery from the initial state.

ForkBench-Choice terminates at the first valid object placement. It preserves the same scene, observations, action interface, and exact paired demonstrations while removing the long-horizon conjunction. Every derived single-object prefix was independently replayed before training.

D Nine-State Multi-Seed Replication

The first multi-state confirmation used states 14–22, seeds 7/17/27, $K = 20$, and five decoders (2,700 rollouts). Results below treat the 27 repeated state–seed cells descriptively.

Treating state–seed cells as independent is optimistic because each seed reuses the same physical observation. We therefore amended the analysis to average seed replicates within state before inference. Under this conservative nine-state analysis, alternative-minus-confidence coverage gains were 0.278 for random (95% CI $[0.130, 0.407]$), 0.296 for temporal-forward ($[0.148, 0.426]$), and 0.241 for uncertainty-first ($[0.037, 0.426]$). Raw exact sign-flip p-values were 0.031, 0.031, and 0.094; all Holm-adjusted values were 0.094. This correction motivated the prospectively frozen 27-new-state expansion rather than post hoc significance claims.

Table 3: Earlier multi-seed confirmation. Primary inference later clusters the three seeds within each of nine physical states.

Decoder	Valid	Both	Cov.	S/T	JS
Sampled-conf.	414/540	7/27	0.630	357/57	0.079
Random	409/540	22/27	0.907	274/135	0.015
Temporal	411/540	23/27	0.926	273/138	0.014
Uncertainty	417/540	20/27	0.870	281/136	0.016
Branch-Aware	395/540	9/27	0.667	311/84	0.046

E Post-hoc Finite-Block Sensitivity

The primary test conditions on the realized $K = 20$ rollout block. We therefore performed a post-hoc sensitivity analysis of a newly sampled 20-rollout block on the same fixed 27-state panel. One paired empirical bootstrap preserves the observed four-scheduler seed vectors within state; a secondary two-level version also resamples states; and four Dirichlet–multinomial posterior-predictive settings use symmetric-sparse, Jeffreys, uniform, and invalid-favoring priors. All six settings, fixed before this sensitivity was computed, are reported; none is a second confirmatory test.

Table 4: Finite-block sensitivity. Values are alternative-minus-confidence Coverage@20. Ranges and predictive-interval envelopes span all six reported assumption settings; P_{rev} is the largest estimated probability of a negative contrast.

Scheduler	Observed Δ	Predictive mean range	95% PI envelope	Max P_{rev}
Random	.241	[.108, .223]	[.000, .315]	2.4%
Temporal-forward	.241	[.085, .189]	[-.037, .296]	6.3%
Uncertainty-first	.222	[.090, .190]	[-.037, .296]	5.2%

Every predictive mean and median remained positive, but the uniform-prior interval included zero for all three alternatives, and its reversal probability exceeded 5% for temporal-forward and uncertainty-first. The result therefore supports directional finite-block stability, most strongly for random order, without substituting for an independent rollout block. The empirical bootstrap cannot generate unseen paired outcome vectors; the Dirichlet–multinomial settings assume scheduler independence conditional on observed counts; and the fixed panel does not identify a task or state superpopulation. The frozen state-level paired sign-flip tests with Holm correction remain the sole primary inference.

F Exploratory Branch-Aware Decoder

We explored whether a token should be prioritized when it is both uncertain and influential. For the top two candidate values at token i, the oracle forces each candidate, reruns the model over remaining masked positions, and uses Jensen–Shannon divergence between future token marginals as influence I_i. It prioritizes a sparse coordinate with a high product $H_i I_i$ in the first denoising iteration, then falls back to confidence.

This sparse token oracle did not work. In the nine-state experiment it improved coverage over sampled-confidence by only 0.037, with clustered $p = 1.0$, and reduced valid-rollout rate by 3.5 points. Individual action dimensions are a poor unit for an intervention distributed across correlated Cartesian coordinates and timesteps. We report this negative result to avoid presenting a post hoc method and do not train the proposed influence predictor. Timestep-group interventions are the more plausible next step.

G Token–Timestep Granularity Ablation

We tested whether grouping the seven action coordinates at each action timestep changes the observed coverage gap. The study froze the same 27 physical initial states (23–49), used a new repli-

cate block (seed 47), drew $K = 10$ rollouts per state and scheduler, and crossed four schedulers with token- and timestep-level reveal units (2,160 rollouts total). The semantic endpoint, validity rules, and state-level paired analysis matched the primary experiment. This was a post-primary robustness study, not a new confirmatory endpoint.

Table 5: Granularity robustness results. Coverage is mean valid-mode Coverage@10 over 27 states. Mode shares pool valid rollouts and are descriptive. Holm-adjusted p-values compare each alternative with sampled-confidence *within the same granularity*.

Scheduler	Reveal unit	Valid	Cov.	S/T (%)	Holm p
Sampled-conf.	token	.715	.574	82.9/17.1	–
Random	token	.737	.722	65.8/34.2	.0430
Temporal	token	.719	.778	66.5/33.5	.0222
Uncertainty	token	.696	.704	65.4/34.6	.143
Sampled-conf.	timestep	.711	.685	64.6/35.4	–
Random	timestep	.726	.778	62.8/37.2	.188
Temporal	timestep	.715	.722	65.3/34.7	.688
Uncertainty	timestep	.730	.759	61.4/38.6	.578

For sampled-confidence, timestep grouping increased mean coverage by 0.111 relative to token scheduling (95% paired bootstrap CI $[-0.037, 0.241]$, exact sign-flip $p = .210$). Its pooled mode-distribution JS divergence fell from .063 to .011, but the state-paired difference was also inconclusive in the 24 states where both conditional distributions were defined ($p = .068$). The entropy–commit Spearman correlation weakened from $.446$ at token level to $.011$ at timestep level, while mean entropy degradation remained similar (.369 versus .351). Thus grouping removes the coordinate-level ordering signature and directionally balances modes, but the behavioral improvement is not established. Moreover, cosine mask-count rounding changes the reveal trajectory: mean zero-indexed commitment occurred at iteration 6.375 for timestep groups versus 7.036 for tokens. We therefore do not interpret this comparison as a pure causal effect of granularity.

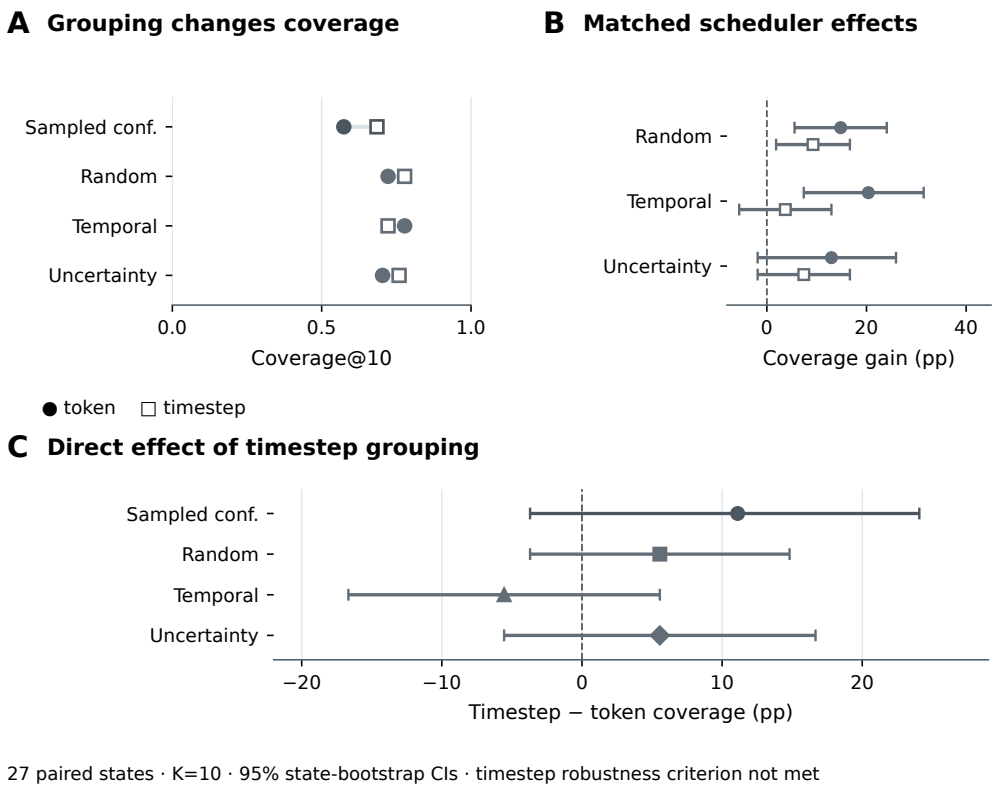


Figure 6: **Token–timestep granularity robustness study** (27 paired initial states, $K = 10$, 270 rollouts per cell). (a) Descriptive state-mean Coverage@10. (b) Gain over sampled-confidence at matched granularity; bars are paired 95% state bootstrap intervals; tests use Holm correction within each three-alternative family, as reported above. Filled circles denote token scheduling and open squares timestep scheduling. Token-level random and temporal-forward replicate the gain, whereas no timestep alternative passes the frozen familywise criterion. (c) Direct timestep-minus-token contrasts; every interval includes zero. Groupwise mask-count rounding changes the reveal trajectory, so panel (c) is not a pure grouping effect.

H Additional Mechanism Diagnostics

Mechanism measurements and hypotheses. For each token, we record its initial entropy

$$H_i^{(0)} = -\sum_v p_i^{(0)}(v) \log p_i^{(0)}(v), \tag{10}$$

first-commit iteration c_i, and entropy at commitment $H_i^{(c_i)}$. The entropy lost while waiting is $\Delta H_i = H_i^{(0)} - H_i^{(c_i)}$. The flexibility-trap hypothesis predicts

$$\rho_S\Big(H_i^{(0)}, c_i\Big) > 0 \quad \text{(sampled confidence)}, \tag{11}$$

with especially late commitment for action coordinates that define the behavioral fork and greater entropy loss before commitment. Random and structured schedules should weaken the global association; uncertainty-first should reverse it by construction. Because this statistic does not localize a physical decision, we also select a semantic action dimension from paired expert chunks on states disjoint from evaluation. Its committed-evidence score averages revealed values over the action horizon, assigning neutral zero to masked coordinates. We measure when that score retrospectively recovers the final first-chunk direction; the diagnostic remains associational without a token intervention.

Semantic-axis construction and support. The semantic dimension is selected from paired expert trajectories on state IDs 0, 2, 3, 4, 5, 6, 8, 9, 11, and 12; none appears in the primary panel. These are training/reference demonstrations rather than an independently sampled test set, but primary rollout outcomes play no role in axis selection. The rule chooses the translation dimension with maximum squared energy in the mean paired tomato-minus-soup first-chunk difference, selecting y with 99.737% of translation-axis energy. The achieved-mode validation conditions on success because the soup/tomato label is undefined for failures. Internal-plan timing instead uses all rollouts. Its common-support state IDs are 24, 27, 30, 37, 38, 44, and 48; requiring both internal directions under every scheduler avoids comparing AUCs on different state sets but is a post-primary restriction. For semantic dimension d^*, iteration-s committed evidence is $z^{(s)} = T^{-1}\sum_t \tilde{a}^{(s)}_{t,d^*} \mathbb{1}\{(t, d^*) \notin M_{s+1}\}$; masked coordinates contribute neutral zero. The retrospective target is each scheduler's own sign of final mean y, not a shared latent decision. Final-score AUC is therefore one by construction and is not itself evidence of a mechanism.

Table 6: Outcome-linked semantic-axis trace summary. Entropy and commitment statistics average within each of 27 states before averaging across states. Commit iterations are zero-indexed.

Scheduler	Entropy at commit	Entropy lost	Commit step
Sampled-conf.	.236	.783	6.965
Random	.316	.702	7.062
Temporal	.319	.700	7.103
Uncertainty	.432	.586	5.192

Over iterations 0–5, state-macro AUC is .718 for sampled-confidence-first, versus .843/.838/.885 for random/temporal/uncertainty. Conditional on the seven selected states, paired differences are .125/.121/.167 and the three within-endpoint Holm values are .046875. These values do not adjust over the post-hoc support rule, early window, threshold, or other analysis choices and are therefore exploratory. For semantic-axis entropy loss, corresponding confidence-minus-alternative differences are .080/.083/.196 nats; the three-contrast exploratory Holm values are .00030. Confidence-minus-alternative mean commit-step intervals are $[-.409, .242]$ for random and $[-.444, .199]$ for temporal-forward, so these data do not establish later semantic-coordinate commitment for either contrast.

Sampled-confidence-first commits x and z translation coordinates at mean iterations 8.80 and 9.35, versus 6.95/7.10 for random, 6.97/7.06 for temporal-forward, and 4.00/3.48 for uncertainty-first. These are the globally highest-entropy translation dimensions (1.87 and 1.79 nats), but they are not the expert-defined semantic y axis. This distinction is why the main text uses outcome-linked evidence rather than treating generic uncertainty as a ground-truth decision point.

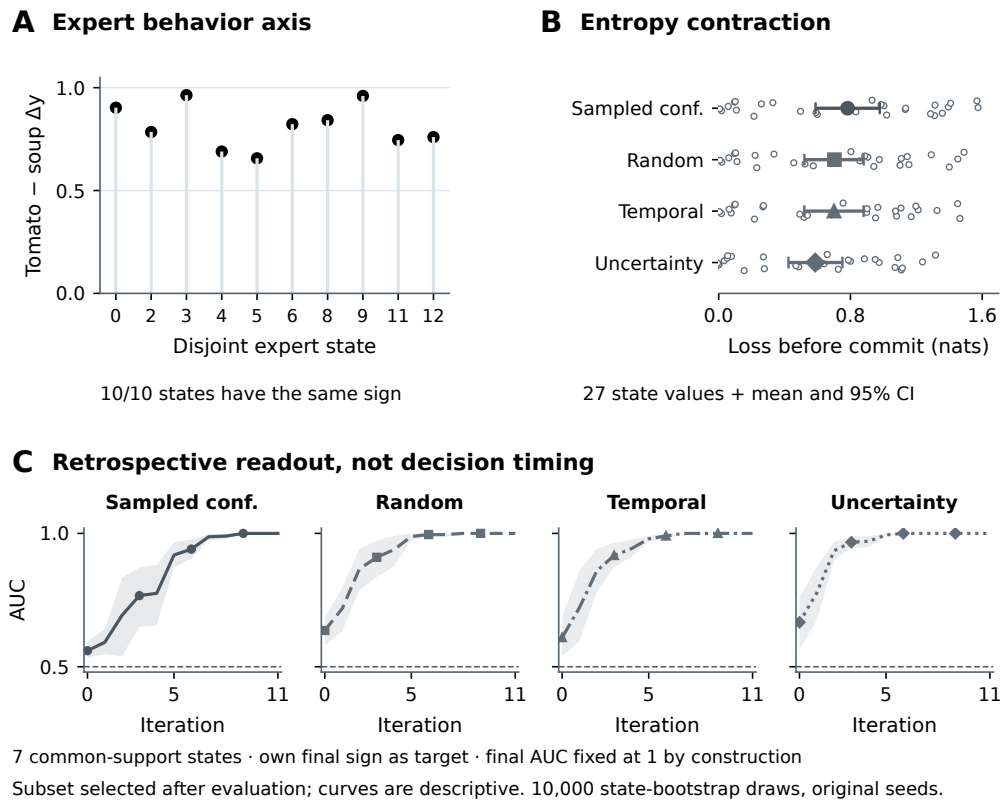


Figure 7: **Complete semantic-fork audit.** (A) Paired chunks from ten disjoint training/reference states select the end-effector y axis; these are not evaluation replicates. (B) Semantic-axis entropy loss before commitment for all 27 primary states; markers and whiskers give means and bootstrap 95% intervals. (C) Retrospective state-macro AUC for each scheduler's own eventual y sign on the seven post-hoc common-support evaluation states, with state-bootstrap 95% pointwise intervals, in separate scheduler panels. Final AUC is one by construction. This exploratory diagnostic links the trace to behavior but does not identify token-level mediation.

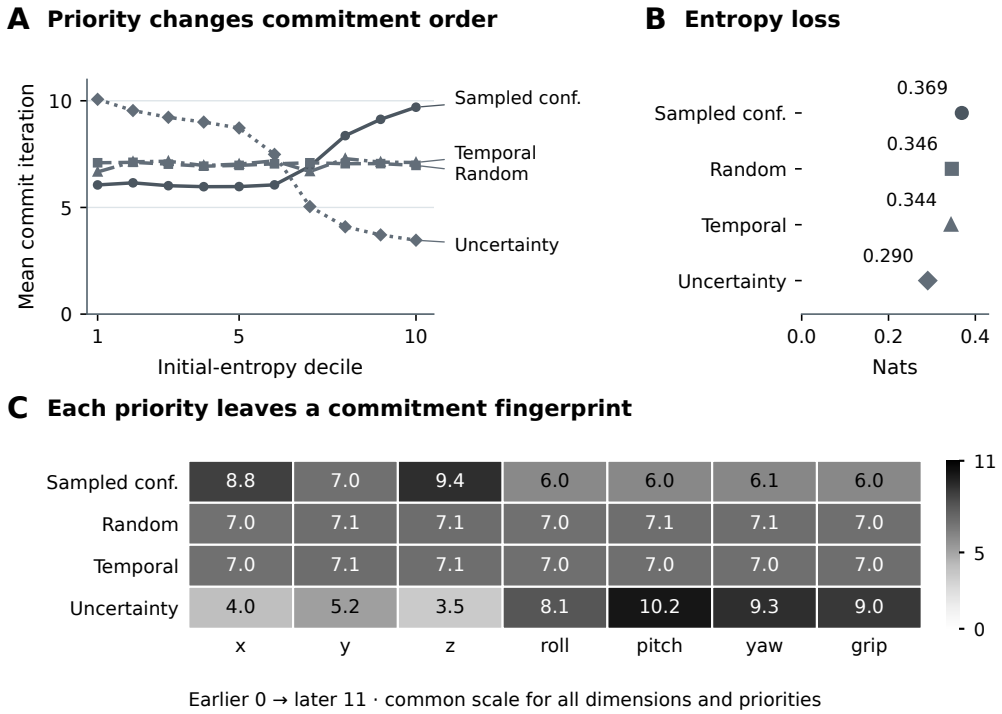


Figure 8: **Scheduler-wide trace diagnostics.** (A) Under sampled-confidence-first, mean commitment time rises with initial entropy ($\rho = 0.446$); random and temporal-forward are near zero, while uncertainty-first reverses the association by construction. Each curve summarizes 30,240 token histories by initial-entropy decile. (B) Descriptive mean entropy lost before commitment across all coordinates. (C) Commitment iteration by action dimension on a common 0–11 color scale. The identical mask-count schedule gives every method the same mean commitment iteration (7.036), but allocates those slots differently. This figure verifies scheduler behavior; unlike Figure 7, it does not localize a behavior-linked branch.

I COMPLETE STATE-RESOLVED COVERAGE

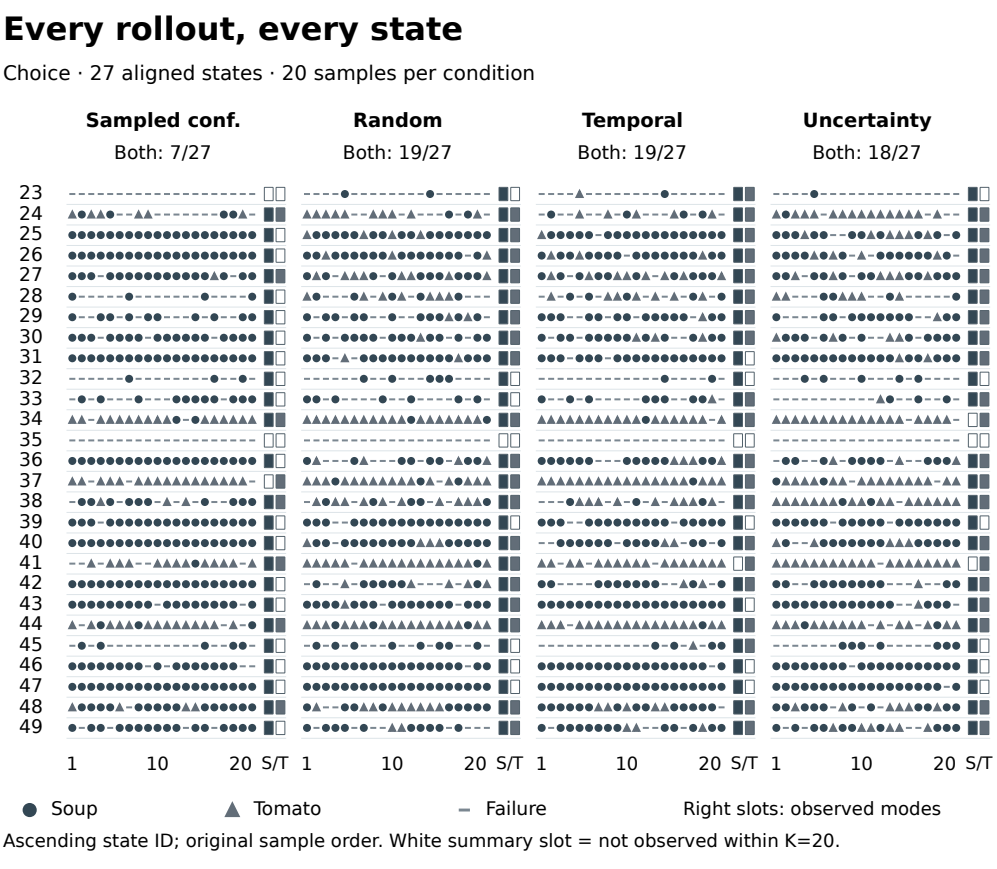


Figure 9: **Every primary rollout, every state.** All 2,160 outcomes in ascending state-ID order (23–49), with the original 20 paired sample positions within each decoder. Teal circles denote soup, orange triangles tomato, and gray dashes failure. The two right-hand slots summarize whether soup and tomato were observed in each state; an empty slot means non-observation within $K = 20$, not missing data or zero mode probability. No states or rollouts are omitted.

J STATISTICAL DETAILS

For each state j and alternative a, the primary paired value is $d_{j,a} = C_{20}(j, a) - C_{20}(j, \text{confidence})$. Confidence intervals are percentile intervals from 10,000 state resamples generated by NumPy's default generator initialized at seed 2026 for the coverage endpoint. The two-sided test statistic is the absolute mean of the nonzero paired differences. Under the sign-exchangeability null, the primary contrasts have $m = 13, 14, 16$, so all 2^m sign assignments are enumerated; zeros are discarded because their signs are invariant. For $m > 20$, we use 10,000 Monte Carlo signs and the plus-one estimate $(b + 1)/(B + 1)$. Holm correction is applied across the three alternatives separately for each endpoint.

Pooled mode shares and JS are descriptive because states contribute unequal valid-rollout counts. State-paired JS excludes states where either conditional distribution is undefined, never substituting an all-zero vector. Over the 25 complete states per primary contrast, reference-minus-alternative JS gains are .078/.064/.071 with enumerated sign-flip Holm $p = .00050/.00282/.00405$. This is a sensitivity analysis because inclusion depends on observed validity. No inference treats rollouts, repeated seeds, or tokens as independent physical environments.

K Paper-Defined Decoder Fidelity Control

The post-primary control uses the same checkpoint, 27 states, and first-query-only scheduler intervention. Its four new cells share rollout seeds $37 + 10s + k$ for state s and sample $k = 0, \ldots, 9$. The reused sampled-confidence/off baseline instead uses $37 + 20s + k$ from the primary $K = 20$ block: all 270 nominal first-ten pairs differ in seed, contrary to the protocol's pairing statement. Comparisons involving that baseline are paired by physical state, not by rollout RNG; this affects the score-off and sampled-confidence remasking descriptives. The prespecified max-confidence/on versus random/on primary comparison uses only new, correctly seed-paired cells and is unaffected by this discrepancy. No recorded outcome or estimate was changed. The complete descriptive grid is:

Table 7: Decoder-fidelity outcomes. Coverage and Success use the $K = 10$ prefix. Validity and pooled soup share include every recorded row in each analysis cell; the reused sampled-confidence/off primary block contains 20 seeds per state, while each new cell contains 10. Soup and tomato shares sum to one among valid rollouts.

First-query priority	Remask	N	Valid	Success@10	Coverage@10	JS to target	Soup share
Sampled confidence	off	540	.713	.926	.537	.054	.808
Sampled confidence	on	270	.726	.926	.593	.053	.806
Max confidence	off	270	.722	.963	.759	.024	.713
Max confidence	on	270	.678	.926	.722	.006	.612
Random	on	270	.689	1.000	.778	.005	.597

Descriptively, replacing sampled confidence with max confidence changes Coverage@10 by $+.222$ with remasking off and $+.130$ with remasking on; enabling remasking changes it by $+.056$ under sampled confidence and $-.037$ under max confidence. The score contrasts are factorial diagnostics, not independent replications. For max confidence with remasking versus random with remasking, random's gain is $+.056$ $[-.019, .130]$, $p = .375$; hence the frozen full-decoder coverage criterion is not met. Independent trace auditing verifies all four newly executed priority formulas exactly and verifies the declared remasking histories.

L Expanded Limitations and Negative Results

Primary evidence covers one adapted checkpoint; the additional fitted model in Appendix N provides a technically recovered sensitivity, not an unqualified prospective replication. Both use the same instruction, scene family, first action chunk, state panel and rollout block. Transfer of learned closed-loop execution across tasks, backbones, mode counts, and continuous diffusion is untested. The separate Approach-Side study confirms route-choice diversity under a fixed-controller interface (Appendix S), not closed-loop transfer. All behavioral results are simulation-only after task-specific adaptation on a small, deliberately balanced dataset, so real-robot transfer and safety are untested. Claims about the complete DD-VLA decoder require the paper-defined max-confidence score with secondary remasking. Our completed control has the same directional point estimate as the primary result but does not meet its prespecified coverage criterion (Appendix K); an ordinary LIBERO control is also required. Timestep grouping is inconclusive (Appendix G), validity lacks an equivalence test, and Success@20 is partly saturated.

The post-primary mechanism analysis conditions its achieved-mode check on success and uses seven post-hoc common-support states for recoverability. Its state-level statistics prevent token pseudoreplication but do not identify mediation. Primary inference is conditional on one realized $K = 20$ seed block and the fixed 27-state panel. Finally, the sparse Branch-Aware oracle reduced validity and did not establish a coverage gain (Appendix F); we therefore claim neither a learned scheduler nor a deployment-ready replacement.

M Trace Schema and Reproducibility

Each JSONL token-step record contains the episode, initial-state ID, replicate and rollout seeds, environment/query index, denoising iteration, action timestep, action dimension, sequence position,

sampled and top-1 tokens, sampled/top-1/top-2 probabilities, categorical entropy, priority, mask state before and after selection, first-commit iteration, and remask count. The schema also reserves fields for exploratory branch influence. Primary trace files contain exactly $540 \times 56 \times 12 = 362{,}880$ records per decoder.

In the full research workspace, `scripts/regenerate_paper_statistics.sh` rebuilds the primary/granularity analyses and exploratory semantic analysis from their original inputs. `paper/scripts/generate_figures.py` rebuilds the configured figures, primary table, and result macros; it does not regenerate every appendix table. The visual teaser is regenerated separately with `paper/scripts/render_visual_teaser.py`; its CSV and provenance manifest bind every primary state–schedule block and the three native demonstration frames, including the documented outcome-frame crops. `make -C paper/iclr2027` compiles the official PDF. These full-workspace commands require inputs excluded from the compact artifact and must not be confused with its portable, read-only audit entry points.

The candidate review artifact contains five statistical replay entry points in `paper/artifacts/iclr2027_review/payload/`: `recompute_statistics.py`, `recompute_recovered_seed17.py`, and `recompute_decoder_fidelity.py`, plus `recompute_selection_coupling.py` and `recompute_semantic_axis_intervention.py`. Together they reconstruct behavioral endpoints from 10,800 unique recorded rows without overwriting caches or pooling studies. The qualified seed-17 failure status and decoder reused-baseline pairing deviation remain explicit; selection coupling retains that same fit-17 qualification, while the semantic-axis replay preserves its unsupported primary family; both exclude raw traces. The separate `regenerate.py` renders the included central displays and primary table; exclusions are recorded in its manifest. All 125 candidate files are byte-bound by the allowlist and pass the technical preflight; the replay and figure checks do not re-execute training or raw-trace treatment audits.

Candidate membership is not an authorship or redistribution-rights conclusion. The archive remains blocked until its file-level human provenance, license, and anonymity gates pass. Checkpoints, demonstrations, robot datasets, raw denoising traces, logs, videos, and third-party assets remain excluded. Internal hash consistency is not independent proof of authenticity without an externally anchored artifact digest.

N Fresh-Checkpoint Sensitivity with Disclosed Technical Recovery

Design and reporting status. A separately fitted checkpoint repeats the recorded two-stage adaptation with training seed 17, the same base model and datasets, matched recorded hyperparameters, 30,000 order-stage and 8,000 Choice-stage optimizer updates, and the prospectively selected terminal checkpoint. Its evaluation retains states 23–49, rollout block 37, $K = 20$, four schedulers, and all 2,160 rollouts. This is not a training-RNG-only comparison: the historical primary run lacks an equivalent trainer/environment provenance record. It also reuses the physical state panel and evaluation block, rather than adding independent tasks or evaluation seeds.

The original completion contract failed. We therefore report this as a *technically recovered checkpoint sensitivity*, not an unqualified prospective replication pass. A post-execution exception was documented before inspecting its effect estimates; no rollout, training record, model weight, scientific protocol, or original completion status was changed.

Technical exceptions and checks. In 4,030 of 8,000 Choice-stage records, the detached next-action accuracy and next-action L1 diagnostics are jointly undefined. The frozen helper divides by the selected-token count or averages over that selection; an empty mask reproduces both NaNs on finite inputs. These diagnostics do not enter the optimized loss, which is finite at every recorded update in both stages. Historical microbatch masks were not saved, so the explanation cannot be verified mask by mask. The recovery rejects every additional nonfinite field, infinity, unpaired NaN, schema mismatch, or schedule/range violation. The machine also rebooted after the run: the recorded and verification environments differ only in the Linux kernel release. The original environment record is retained, not overwritten with the later one.

The separate audit verifies original training configurations and schedules, checkpoint provenance, merge receipts and inventories, finite merged-model, adapter and projector weights, the complete rollout and trace grids, all four scheduler-priority formulas, and regeneration of the analysis outputs. Original scientific-output checksums bind the recovered summaries. These checks support a qualified sensitivity analysis; they do not retroactively satisfy the original completion contract.

Table 8: Recovered seed-17 checkpoint; 540 rollouts per scheduler over the same 27 states. Coverage differences and intervals are alternative minus sampled confidence. Holm adjustment uses the original three-contrast coverage family. All four schedulers attain Success@20 of 27/27. Recovery status is qualified as described above.

Scheduler	Valid rollouts	Coverage@20	Gain [95% CI]	Holm p
Sampled confidence	473/540	.704	—	—
Random	470/540	.870	.167 [.074, .259]	.00781
Temporal-forward	472/540	.870	.167 [.074, .259]	.00781
Uncertainty-first	471/540	.926	.222 [.130, .315]	.00146

All three coverage contrasts retain the primary direction. Their respective validity differences are $-.006$ $[-.041, .030]$, $-.002$ $[-.028, .024]$, and $-.004$ $[-.046, .039]$. The numerical coverage and validity conditions in the frozen analysis are met, but the original completion requirement is not. The result is consistent with sensitivity to reveal priority beyond the primary fitted checkpoint; it is neither an additional task replication nor an excuse to generalize to paper-defined max-confidence decoding with remasking. We do not pool the two checkpoints to manufacture an overall effect or omit the technical exceptions from the interpretation.

O CANDIDATE-DEPENDENT SELECTION AND EXACT CONDITIONAL DECODING

The construction in Appendix B is an instance of the following selection identity. We state its assumptions explicitly to distinguish same-candidate selection from adaptive order alone, parallel factorization, and prediction error. These are oracle-distribution statements, not a claim that the learned robot decoder has compatible exact conditionals.

Theorem 1 (one reveal followed by exact completion; restatement). Let $\mathcal{X} = \prod_{i=1}^{L} \mathcal{X}_i$, where $L \geq 1$ and each $\mathcal{X}_i$ is finite. Let P be a probability distribution on $\mathcal{X}$, and let P_i denote its ith coordinate marginal. Assume $P_i(v) > 0$ for every i and $v \in \mathcal{X}_i$. Let Y be a candidate vector satisfying $\Pr(Y_i = v) = P_i(v)$; its coordinates need not be independent. Select $I \in \{1, \dots, L\}$ by a randomized rule that may inspect Y, commit $X_I = Y_I$, and draw the remaining coordinates from $P(X_{-I} \mid X_I)$. This completion draw is independent of the uncommitted candidates and the selection event conditional on the committed index and value. Write $a_i(v) = \Pr(I = i \mid Y_i = v)$. The output law Q satisfies

$$Q(x) = P(x)w(x), \qquad w(x) = \sum_i a_i(x_i). \tag{12}$$

Consequently $\mathbb{E}_P[w] = 1$, and $Q = P$ if and only if $w = 1$ P-almost surely.

Proof. Fix $x \in \mathcal{X}$. Conditional on $I = i$, the output equals x only if $Y_i = x_i$. The assumed completion kernel is $P(X_{-i} = x_{-i} \mid X_i = x_i)$, irrespective of unused candidates. The law of total probability therefore gives

$$\begin{aligned} Q(x) &= \sum_{i=1}^{L} \Pr(I = i, Y_i = x_i) P(x_{-i} \mid x_i), \\ \Pr(I = i, Y_i = x_i) &= \Pr(Y_i = x_i) \Pr(I = i \mid Y_i = x_i) = P_i(x_i) a_i(x_i), \\ P(x_{-i} \mid x_i) &= \frac{P(x)}{P_i(x_i)}. \end{aligned}$$

Every denominator is positive by assumption. Substitution yields

$$Q(x) = \sum_{i=1}^{L} P_i(x_i) a_i(x_i) \frac{P(x)}{P_i(x_i)} = P(x) w(x).$$

This calculation also applies when $P(x) = 0$, in which case $Q(x) = 0$. Moreover,

$$\begin{aligned}\mathbb{E}_P[w] &= \sum_{i=1}^{L} \sum_{x \in \mathcal{X}} P(x) a_i(x_i) = \sum_{i=1}^{L} \sum_{v \in \mathcal{X}_i} P_i(v) a_i(v) \\ &= \sum_{i=1}^{L} \sum_{v \in \mathcal{X}_i} \Pr(I = i, Y_i = v) = \sum_{i=1}^{L} \Pr(I = i) = 1.\end{aligned}$$

If $w(x) = 1$ on the support of P, the identity gives $Q(x) = P(x)$ there and both laws vanish elsewhere. Conversely, if $Q = P$, division by $P(x) > 0$ gives $w(x) = 1$ at every point in that support. □

Corollary (event probabilities and total variation). For any event $B \subseteq \mathcal{X}$,

$$Q(B) - P(B) = \mathrm{Cov}_P(\mathbf{1}_B, w), \qquad \|Q - P\|_{\mathrm{TV}} = \tfrac{1}{2}\mathbb{E}_P|w - 1|. \tag{13}$$

Here $\mathbf{1}_B$ denotes the indicator of B, and $\|Q - P\|_{\mathrm{TV}} = \frac{1}{2}\sum_x |Q(x) - P(x)|$.

Proof. Using $Q(x) = P(x)w(x)$ and $\mathbb{E}_P[w] = 1$ gives

$$\begin{aligned}Q(B) - P(B) &= \sum_x P(x)\mathbf{1}_B(x)(w(x) - 1) \\ &= \mathbb{E}_P[\mathbf{1}_B w] - \mathbb{E}_P[\mathbf{1}_B]\mathbb{E}_P[w], \\ \|Q - P\|_{\mathrm{TV}} &= \tfrac{1}{2}\sum_x P(x)|w(x) - 1| = \tfrac{1}{2}\mathbb{E}_P|w - 1|.\end{aligned}$$

The first expression is the stated covariance. □

Example (candidate dependence without distortion). Let P be uniform on $\{0,1\}^2$, and draw independent fair bits Y_1, Y_2. Set $I = 1$ if $Y_1 = Y_2$, and $I = 2$ otherwise. For each $v \in \{0,1\}$,

$$\begin{aligned}a_1(v) &= \Pr(Y_2 = v \mid Y_1 = v) = \tfrac{1}{2}, \\ a_2(v) &= \Pr(Y_1 \neq v \mid Y_2 = v) = \tfrac{1}{2}.\end{aligned}$$

Consequently $w(x) = 1$ for all x and $Q = P$, despite candidate dependence.

Proposition (sequential selection identity). Let $h = ((i_1, v_1), \ldots, (i_{t-1}, v_{t-1}))$ be an ordered history of distinct revealed positions and their values, and let $R(h)$ be the set of remaining positions. Write $P_i(v \mid h)$ for the conditional marginal under P given these revealed values. At each step, draw fresh candidates satisfying $\Pr(Y_i = v \mid h) = P_i(v \mid h)$, select $I \in R(h)$, and commit Y_I. Define $a_i(v \mid h) = \Pr(I = i \mid Y_i = v, h)$ on positive-probability events. Let $\mathfrak{S}_L$ be the set of permutations of $\{1, \ldots, L\}$, and put $h_{\sigma,<t}(x) = ((\sigma_1, x_{\sigma_1}), \ldots, (\sigma_{t-1}, x_{\sigma_{t-1}}))$. For $P(x) > 0$, the output law satisfies

$$\frac{Q(x)}{P(x)} = \sum_{\sigma \in \mathfrak{S}_L} \prod_{t=1}^{L} a_{\sigma_t}(x_{\sigma_t} \mid h_{\sigma,<t}(x)). \tag{14}$$

Proof. At a reachable history h, the probability of selecting position i and committing value v is

$$\Pr(I = i, Y_i = v \mid h) = P_i(v \mid h) a_i(v \mid h).$$

Fix x with $P(x) > 0$ and an order σ. Multiplication of the successive transition probabilities gives

$$\begin{aligned}&\Pr\big(X = x, (I_1, \ldots, I_L) = \sigma\big) \\ &= \prod_{t=1}^{L} P_{\sigma_t}(x_{\sigma_t} \mid h_{\sigma,<t}(x)) a_{\sigma_t}(x_{\sigma_t} \mid h_{\sigma,<t}(x)) \\ &= P(x) \prod_{t=1}^{L} a_{\sigma_t}(x_{\sigma_t} \mid h_{\sigma,<t}(x)).\end{aligned}$$

The last equality uses the chain rule in the fixed order σ. The order events are disjoint and exhaustive; summing over σ and dividing by $P(x)$ proves the formula. An unreachable order prefix has probability zero, so its later selection kernels may be specified arbitrarily without changing its contribution. In addition, a positive-probability transition has $P_i(v \mid h) > 0$; starting from the empty history, induction shows that each reached partial assignment has positive probability under P. Thus $Q(x) = 0$ whenever $P(x) = 0$. □

Corollary (candidate-independent sequential selection). Under the preceding assumptions, suppose I is conditionally independent of the fresh candidate vector given h. Then $Q = P$.

Proof. Conditional independence gives $a_i(v \mid h) = \Pr(I = i \mid h) = a_i(h)$, where $\sum_{i \in R(h)} a_i(h) = 1$. Fix x in the support of P. Let $T_x(h)$ denote the sum of products of the selection probabilities over all continuations from a history h consistent with x. With $h \parallel (i, x_i)$ denoting history concatenation, this sum obeys

$$T_x(h) = \begin{cases} 1, & R(h) = \varnothing, \\ \displaystyle\sum_{i \in R(h)} a_i(h) T_x(h \parallel (i, x_i)), & R(h) \neq \varnothing. \end{cases}$$

Backward induction on $|R(h)|$ gives $T_x(h) = 1$: the terminal case is defined above, and if every child has value one, the displayed sum equals $\sum_{i \in R(h)} a_i(h) = 1$. In particular, the path sum in Equation (14) is $T_x(\varnothing) = 1$. Hence $Q(x) = P(x)$ on the support, and both vanish off it. □

The corollary permits selection based on revealed values, position identities, conditional entropies, or maximum conditional probabilities. Such dependence on the history is distinct from dependence on the fresh candidates.

Proposition (parallel selection identity). Assume $P(x) > 0$ for every $x \in \mathcal{X}$. Draw independent candidates $Y_i \sim P_i$, select a nonempty random set S, and commit Y_S. Complete the remaining coordinates from $P(X_{-S} \mid X_S = Y_S)$, independently of unused candidates and the selection event given (S, Y_S). For each nonempty deterministic set $A \subseteq \{1, \ldots, L\}$, let P_A be its joint marginal and define $b_A(v_A) = \Pr(S = A \mid Y_A = v_A)$. Then

$$\frac{Q(x)}{P(x)} = \sum_{A \neq \varnothing} b_A(x_A) \frac{\prod_{i \in A} P_i(x_i)}{P_A(x_A)}. \tag{15}$$

Proof. Independence of the candidates and the definition of b_A imply

$$\Pr(S = A, Y_A = x_A) = b_A(x_A) \Pr(Y_A = x_A) = b_A(x_A) \prod_{i \in A} P_i(x_i).$$

Conditioning on the selected set and committed values therefore yields

$$\begin{aligned} Q(x) &= \sum_{A \neq \varnothing} \Pr(S = A, Y_A = x_A) P(x_{-A} \mid x_A) \\ &= \sum_{A \neq \varnothing} b_A(x_A) \prod_{i \in A} P_i(x_i) \frac{P(x)}{P_A(x_A)}. \end{aligned}$$

Full support ensures that $P_A(x_A) > 0$ and $P(x) > 0$. Division by $P(x)$ gives the stated identity. □

If selection is candidate-independent, $b_A(x_A) = \Pr(S = A)$, but the ratio $\prod_{i \in A} P_i(x_i)/P_A(x_A)$ need not equal one for correlated coordinates. Thus removing selection dependence does not, by itself, remove parallel factorization error. This is distinct from the dependence-induced parallel error analyzed by Zhao & Cai (2026). Guarantees for entropy-budget adaptive decoding likewise concern a different selection rule (Cai & Li, 2026). Our identity does not contradict those guarantees. Nor is confidence-order mismatch with a diffusion reverse process itself new (Zhao et al., 2025).

Corollary (sample splitting). Under the sequential proposition's assumptions, draw candidate vectors Y and Z independently conditional on h, both with exact conditional marginals. Let U be auxiliary randomness independent of (Y, Z) given h. Choose $I = f(h, Y, U)$ and commit Z_I. Then $Q = P$.

Proof. The stated independence implies $I \perp Z \mid h$. For every remaining position i and value v,

$$\Pr(I = i, Z_i = v \mid h) = \Pr(I = i \mid h) \Pr(Z_i = v \mid h) = \Pr(I = i \mid h) P_i(v \mid h).$$

Thus the selection kernel is independent of the committed candidate. The candidate-independent sequential corollary applies. □

Sample splitting and the robot control. Our split rule takes f to rank positions by $P_i(Y_i \mid h)$, with independent position noise included in U, while committing Z_i. For the parallel learned decoder, the intervention removes the same-draw selection dependence, not factorization error or model error, and may change both coverage and validity. A matched comparison must consume both categorical draws in every condition, keep the position-noise and model-call budgets fixed, and report null and adverse results. Our prospectively frozen robot control follows these requirements. Split confidence improves Coverage@10 over the matched same-draw condition on both fixed fits while the validity guard passes (Section 6.3 and Appendix P). Random or max confidence is descriptively better than split on each fit, so the result identifies harmful same-draw coupling without establishing sample splitting as a universal remedy.

Computational checking and scope. Exact rational enumeration checks the one-reveal and sequential path identities on 36 full-support binary distributions in dimensions 2, 3, and 4, under four rules (144 distribution–rule cases). The rules are sampled-probability-weighted selection with a 10^{-20} regularizer, maximum-probability selection, history-adaptive candidate-independent selection, and uniform selection. Sample splitting and all three candidate-independent rules preserve the joint in these checks; every tested sampled-probability case is distorted. This finite computational check supplements the proofs; it does not certify novelty, learned-model compatibility, or mediation of the observed robot coverage differences.

P Prospective Same-Draw Selection-Coupling Control

Question and status. This post-primary intervention asks whether a sampled candidate's ability to raise its own reveal priority changes valid-mode coverage. The protocol was frozen on September 7, 2026 after the original scheduler and recovered-fit results, but before this intervention was implemented or queried on either checkpoint. Both fixed checkpoints and all four conditions were required regardless of the first cell's outcome. The run finished on September 18 with 2,160 rollouts and passed the frozen treatment, provenance, grid, trace, and analysis-regeneration checks. It reuses states 23–49 and is not an independent task or state-panel replication.

Two-draw intervention. At each of the 12 first-query denoising iterations, logits define categorical weights p_i. Every condition consumes two successive full-array multinomial draws, a pilot Y and committed candidate Z, followed by the same position noise. Previously visible tokens are preserved, and only newly selected positions commit Z. The conditions differ only in priority:

$$q_i^{\text{coupled}} = p_i(Z_i), \qquad q_i^{\text{split}} = p_i(Y_i), \tag{16}$$

$$q_i^{\max} = \max_v p_i(v), \qquad q_i^{\text{random}} = 1. \tag{17}$$

At a common revealed history, Y_i and Z_i have the same marginal law, so split and coupled priorities have the same conditional marginal distribution. Only coupled priority reads the value that is committed. Treatment-dependent histories may diverge after a reveal; this is not a replay of a fixed order. All conditions use token granularity, the cosine mask counts, annealed Gumbel position noise at the upstream scale, no remasking, no critic, 12 ordinary model forwards, and the original single-draw decoder on later policy queries.

The frozen protocol described categorical weights as bfloat16. Before any outcome analysis, a precision audit recorded that the actual softmax weights passed to multinomial sampling are float32 under the frozen source. The amendment changed no checkpoint, condition, seed, endpoint, decision rule, or run byte; both the original statement and the corrected precision record are retained.

Design and inference. Each checkpoint uses evaluation seed 1,092,026, ten rollouts at each of 27 states, and paired rollout seeds across conditions. The primary contrasts are split minus coupled

Coverage@10, analyzed separately by checkpoint. Percentile intervals use 100,000 paired-state bootstrap draws. Two-sided sign-flip tests discard exact zero state differences and enumerate all sign assignments here; Holm controls the two-checkpoint coverage family. The null requires exchangeable signs of the nonzero paired state effects. Intervals describe resampling stability on the fixed panel, not population uncertainty. The frozen per-fit gate requires coverage gain at least .10, coverage-interval lower bound above zero, Holm $p < .05$, and validity-difference interval lower bound at least $-.05$. Checkpoints are never pooled.

Table 9: All conditions in the same-draw coupling control. Coverage and validity are state means; Success@10 counts states with at least one task success. Random and max are descriptive. The dagger marks the fit whose original training completion contract remains failed.

Fit	Priority	Coverage@10	Validity	Success@10
Primary fit	Coupled	.537	.685	25/27
	Split	.741	.730	25/27
	Random	.685	.693	25/27
	Max	.778	.700	25/27
Recovered fit 17†	Coupled	.630	.863	26/27
	Split	.759	.870	26/27
	Random	.852	.841	26/27
	Max	.815	.867	27/27

Table 10: Prespecified split-minus-coupled contrasts. Coverage p-values use Holm correction across the two checkpoints; intervals are unadjusted paired state-bootstrap intervals. The validity criterion is a confidence-bound guard, not a claim of equivalence.

Fit	Coverage gain [95% CI]	Holm p	Validity diff. [95% CI]
Primary fit	.204 [.111, .296]	0.00195	.044 [-0.007, .096]
Recovered fit 17†	.130 [.056, .222]	0.01562	.007 [-0.019, .033]

Complete behavioral result. For the primary fit, coupled and split coverage are .537 and .741; the paired gain is .204 [.111,.296], raw $p = .00098$, Holm $p = .00195$. Eleven states improve, 16 tie, and none worsen. Validity changes from .685 to .730, a difference of .044 $[-.007, .096]$, and Success@10 is 25/27 in both cells. For recovered fit 17, coverage is .630 versus .759; the gain is .130 [.056,.222], raw and Holm $p = .01563$. Seven states improve, 20 tie, and none worsen. Validity changes from .863 to .870, a difference of .007 $[-.019, .033]$, and Success@10 is 26/27 in both. Both fits meet every frozen behavioral predicate.

The descriptive controls are material adverse-to-simplicity evidence. On the primary fit, random and max-confidence coverage are .685 and .778; on recovered fit 17 they are .852 and .815. Split is therefore neither uniquely effective nor best in either fit. These cells were not additional confirmatory contrasts, and no post-hoc pairwise p-values are assigned to them.

Proximal diagnostic and mechanism boundary. At iteration zero, the prespecified score averages over newly committed coordinates the log probability of Z_i minus its expectation under the actual categorical sampling weights. Coupled-minus-split means are .144 [.067,.234] and .075 [.012,.176], with separate-family Holm $p = .00012$ and .00586. This confirms the intended immediate selection tilt on both fits. It does not show that this score mediates final behavior, identify when a semantic branch is decided, or exclude parallel factorization and model error.

Treatment and provenance audit. For every checkpoint–state–sample triple, iteration-zero logits, pilot and committed candidates, and RNG digests match across conditions. Each of the eight cells contains 270 rollouts and 181,440 first-query token-step rows. The audit checks both candidate arrays, committed tokens, priorities, normalized log-probability centering, mask counts, remasking state, and model-call budget over 1,451,520 trace rows. The completion verifier regenerates the sealed analysis and binds the source, environment, checkpoint inventories, protocol, precision

amendment, outputs, and receipt. The compact paper tree retains all eight outcome CSVs, the analysis JSON, protocol, amendment, and completion receipt; raw traces remain outside the small review payload.

Qualification and claim boundary. Recovered fit 17 still has 4,030 undefined detached training diagnostics and a later verification kernel, as detailed in Appendix N. New, fully verified rollouts do not repair that original training contract. The cross-fit result is therefore evidence that same-draw coupling matters for two fixed fits of this one task, with the second explicitly qualified. It is not a random-effects model sample, a cross-task replication, a proof that sample splitting preserves a learned parallel decoder's joint distribution, or a deployment recommendation.

Q Prospectively Frozen Semantic-Axis Priority Control

This control was frozen before its four seed-73 conditions were evaluated. It uses the same primary checkpoint and reused states 23–49, with ten paired rollouts per state and condition (1,080 rollouts total). The first query either uses sampled confidence, prioritizes every token on the expert-selected y axis, prioritizes the movement-scale-matched nonsemantic x axis, or delays the y axis; later queries use sampled confidence. Candidate sampling, mask counts, 12 model calls, and non-target priorities are fixed. This is an axis-priority intervention, not a value clamp or token-level mediation test.

Table 11: Semantic-axis intervention outcomes on 27 reused states at $K = 10$. Coverage, validity, and Success are state-macro means.

First-query priority	Coverage@10	Valid rate	Success@10
Sampled confidence	.556	.711	.889
Semantic y first	.667	.696	.926
Nonsemantic x first	.648	.800	.963
Semantic y last	.630	.707	.889

Table 12: Prespecified state-paired contrasts, oriented as semantic-y-first minus comparator. Intervals are 10,000-draw state-bootstrap 95% intervals; Holm adjustment covers the three Coverage@10 contrasts. Validity intervals are descriptive.

Comparator	ΔCoverage [95% CI]	Holm p	Δvalidity [95% CI]
Sampled confidence	+.111 [.000,.222]	.328	−.015 [−.093,.052]
Nonsemantic x first	+.019 [−.111,.130]	1.000	−.104 [−.167,−.048]
Semantic y last	+.037 [−.074,.148]	1.000	−.011 [−.085,.067]

None of the three primary contrasts meets the frozen family-wise criterion; the comparison with the x-axis sham is nearly zero, so the study does not support semantic-axis specificity. The sham also has a descriptively higher valid-rollout rate than semantic-y-first. We therefore retain the earlier semantic trace analysis as associational and do not use this control to expand the paper's causal headline. A separately dated pre-run validity addendum was not included in the completion source manifest; no validity-qualified positive claim depends on it. The canonical completion verifier nevertheless passes for the frozen protocol, 1,080 outcome rows, traces, analyses, and output manifest.

R Prospective Position-Noise Intervention

Question and prospective scope. This experiment asks whether increasing auxiliary position-selection noise changes valid-mode Coverage@10 in the prespecified positive ordinal direction. It follows the primary schedule comparison, using the same terminal checkpoint, task, and states 23–49. The protocol was frozen and amended on September 3, 2026 before any seed-83 rollout; the complete seven-condition run finished on September 9. This is a prospectively specified intervention on a reused panel, not an untouched-state or independent-checkpoint replication. All six doses, the random anchor, null contrasts, and the nonmonotone response are reported.

Intervention and fixed quantities. For first-query denoising step $s \in \{0, \dots, 11\}$, position selection uses

$$r_{i,s} = \log(\max\{p_{i,s}(\widetilde{x}_{i,s}), \epsilon\} + \epsilon) + T\left(1 - \frac{s+1}{12}\right) G_{i,s}, \qquad G_{i,s} \sim \text{Gumbel}(0, 1). \tag{18}$$

Here $\epsilon = 10^{-20}$ retains the implementation's clamp and regularizer. The finite doses are $T = .05, .25, .5, 1, 2, 4$. Here T scales auxiliary *position-selection* noise, not categorical action temperature: untempered softmax followed by multinomial sampling is unchanged. $T = 1$ is the exact upstream position-temperature setting; $T = .05$ is the predeclared statistical reference. $T = 4$ still uses sampled-confidence priorities and is not identical to random order. The separate anchor uses equal priorities with Gumbel ranking; it is excluded from the primary finite-dose test.

All conditions use token granularity, an 8×7 chunk, 12 ordinary decoder forwards, the same cosine mask-count schedule, and no secondary remasking, critic, or extra counterfactual calls. All later policy queries use sampled-confidence selection with $T = 1$. The final denoising step has zero effective position noise, but its hypothetical next mask is discarded and its categorical samples are returned. Identical evaluator-derived seeds pair conditions within each state and sample index. There are ten samples for each of 27 states, or 270 rollouts per condition and 1,890 total. Execution order was shuffled with NumPy seed 2027 before outcomes: $T = .5, 1, 4$, random anchor, $T = .25, .05, 2$.

Pre-run zero-temperature amendment. An audit of archived primary traces found exact sampled-priority ties at the first-step $k = 55$ cutoff in 540/540 groups. At $T = 0$, the implementation's strict threshold would violate the scheduled mask count in every group (minimum reconstructed count 24; maximum cutoff multiplicity 32). The protocol therefore replaced empirical $T = 0$ with $T = .05$ before its output root existed or any seed-83 rollout began, leaving the checkpoint, panel, budget, primary test, multiplicity, and report-all-outcomes rule unchanged. The positive floor supplies tie-breaking noise; it is not deterministic decoding. $T \to 0$ remains an analytic limit only. Further pre-run amendments strengthened GPU serialization and provenance/completion checks without changing the scientific design.

Estimand, inference, and decision rules. For state j, Coverage@10 is the number of the two prespecified valid modes observed among ten rollouts, divided by two. Failures remain in the sampling budget and validity denominator. The primary statistic is the mean of the 27 least-squares slopes of this outcome on fixed ordinal scores $0, \dots, 5$, in ascending finite-dose order. It is *not* a slope per unit T. The one-sided test independently permutes the six labels within each state over 100,000 Monte Carlo draws with a plus-one correction. Its reference null requires within-state dose-label exchangeability, with independently sampled permutations across states; randomized wall-clock order does not itself establish that assumption. A single prespecified primary test uses $\alpha = .05$. Percentile intervals use 100,000 state-bootstrap draws. The analysis random seed is 2029. Inference remains conditional on this fixed panel and realized rollout block; state resampling is a stability analysis, not superpopulation uncertainty.

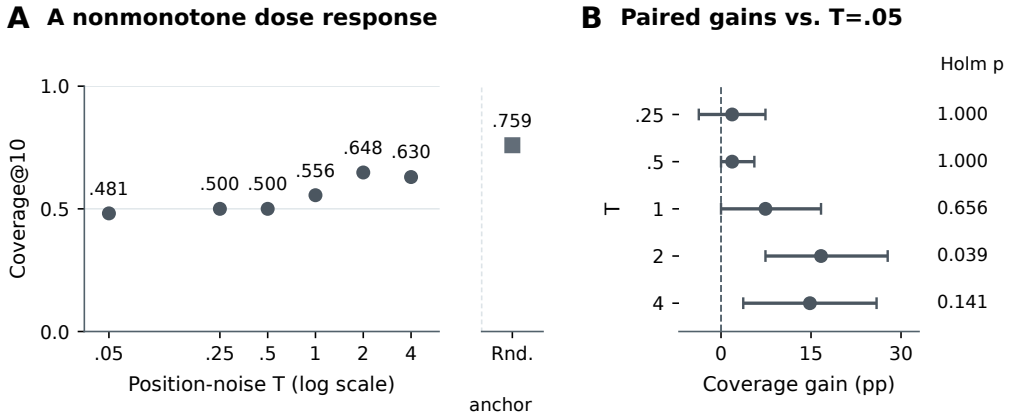
The five higher-dose coverage contrasts against $T = .05$ form a separate prespecified secondary family: two-sided paired sign-flip tests with Holm correction, assuming exchangeable nonzero difference signs. Their 95% bootstrap intervals are unadjusted and need not agree with family-wise test decisions. They cannot rescue a failed primary trend. For every higher dose with a coverage gain, a predeclared guard flags an absolute validity-rate drop exceeding .05 versus $T = .05$. This guard uses point estimates and is not a confidence-bound noninferiority test. The protocol requires the positive primary trend, higher coverage at $T = 4$ than at $T = .05$, and a passing guard for its coverage-improvement criterion.

Complete response and primary result. Table 13 and Figure 10 report all outcomes. The mean ordinal slope is .03545, with 95% state-bootstrap interval [.01376,.05979] and one-sided permutation $p = 1/100001 \approx .00001$, at the Monte Carlo resolution floor. Ten state slopes are positive, 15 are zero, and two are negative. The observed means rise from .481 at $T = .05$ to .648 at $T = 2$, then decline to .630 at $T = 4$; they are not monotonically increasing. The primary criterion passes, but neither strict monotonicity nor a universally beneficial noise increase is supported.

Secondary contrasts and validity guard. Only $T = 2$ meets the five-comparison Holm criterion: gain .167, 95% interval [.074,.278], adjusted $p = .03906$. Although the $T = 4$ gain is .148 with

Table 13: Complete position-noise response: 27 states and 270 rollouts per row. Both-mode and no-mode counts use all states; failures remain in the budget. The random anchor is descriptive and is not a seventh finite dose.

T / condition	Coverage@10	Valid rate	Success@10	Both modes	No mode
0.05	0.481	0.681	24/27	2/27	3/27
0.25	0.500	0.707	25/27	2/27	2/27
0.5	0.500	0.704	24/27	3/27	3/27
1	0.556	0.674	24/27	6/27	3/27
2	0.648	0.674	25/27	10/27	2/27
4	0.630	0.670	24/27	10/27	3/27
Random anchor	0.759	0.700	25/27	16/27	2/27



All six doses retained; no smoothing or inferred optimum. Random is a separate condition.

27 paired states · K=10 · A descriptive means; B 95% paired-state bootstrap CIs; five-test Holm family.

Figure 10: **Position-selection noise changes coverage without changing categorical action temperature.** (A) All six finite doses on a logarithmic horizontal axis; points are observed means, with no fitted curve. The random anchor is separate and excluded from the ordinal-score trend test. (B) All five prespecified contrasts against $T = .05$; bars are unadjusted 95% state-bootstrap intervals, whereas p-values use five-comparison Holm correction. Only $T = 2$ passes that secondary criterion. Each condition has 27 paired states and ten rollouts per state; failures remain in the budget.

Table 14: All five prespecified secondary Coverage@10 contrasts against $T = .05$. Intervals are unadjusted 95% state-bootstrap intervals; Holm p controls the five-test family.

T versus .05	Coverage gain	95% bootstrap CI	Raw p	Holm p
0.25	0.019	$[-0.037, 0.074]$	1.00000	1.00000
0.5	0.019	$[0.000, 0.056]$	1.00000	1.00000
1	0.074	$[0.000, 0.167]$	0.21875	0.65625
2	0.167	$[0.074, 0.278]$	0.00781	0.03906
4	0.148	$[0.037, 0.259]$	0.03516	0.14062

interval [.037,.259], its adjusted $p = .14062$ does not pass (Table 14). Observed validity differences from $T = .05$ are +.02593, +.02222, $-.00741$, $-.00741$, and $-.01111$ in ascending higher-dose order; none exceeds the five-point drop guard. Thus the protocol's coverage-improvement criterion passes without establishing preserved competence or equivalence. Random-anchor coverage is .759, exceeding $T = .05$ by .278 with descriptive interval [.167,.389], and $T = 4$ by .130 [.037,.241]. No anchor p-values were specified. These comparisons do not select a deployment temperature or extend the finite-dose trend to an empirical infinite dose.

Conditional distributional descriptives. Table 15 reports the remaining prespecified descriptive endpoints. JS uses natural logarithms and an equal-weight mean over states with at least one valid rollout; its available-state set differs by condition. Pooled tomato shares instead weight states by valid counts. Neither comparison is an unconditional coverage measure or a mediation test.

Table 15: All conditional distributional descriptives. Missing JS values are undefined, not zero. Tomato share is among pooled valid rollouts; alphabet-soup share is its complement.

T / condition	Mean conditional JS	Undefined states	Tomato share	Valid n
0.05	0.2020	3/27	0.1739	184
0.25	0.2025	2/27	0.1466	191
0.5	0.1899	3/27	0.1368	190
1	0.1736	3/27	0.1758	182
2	0.1534	2/27	0.2253	182
4	0.1492	3/27	0.2762	181
Random anchor	0.1054	2/27	0.3016	189

Trace and completion checks. The completion receipt binds the frozen protocol, source/checkpoint/environment manifests, and rollout/analysis inventories. The treatment audit passes on all 1,270,080 first-query token-step records, checking matching iteration-zero categorical candidates and logits, equal mask counts at every decoder step, the ordinary model-call budget, token granularity, and disabled remasking. Later candidates may differ because changed reveal order changes conditioning context. Raw Gumbel variates are not serialized: their pairing follows the shared seeded code path and is not independently reconstructed from traces. The intervention identifies sensitivity to position-selection noise, not a mediating token, decision-formation time, exact VLA conditionals, or quantitative agreement with the two-token construction.

Source and reproduction boundary. The full-workspace study directory is `experiments/flexibility_trap_runs/forkbench_choice/position_temperature_states23_49_seed83_k10`. The summary is `position_temperature_dose_response.json`; the protocol, trace audit, seven compact mode CSVs, and completion receipt retain their original bytes. The dedicated renderer `paper/scripts/render_temperature_result.py --verify-replay` can verify exact in-memory statistical reconstruction from the CSVs and regenerate Figure 10 and Tables 13 to 15. Its generated provenance records source, script, figure, and table hashes. These files are not in the separately gated 125-file candidate capsule; temperature packaging and accountable redistribution approval remain pending.

S Fresh Approach-Side Route-Choice Validation

Question and execution boundary. This study tests whether decoupling a committed candidate from its reveal priority increases diversity of VLA choices in a different setting from Choice. The model chooses a left or right approach to a manipulated object; a fixed controller supplies the corresponding route. The learned model generates one 8×7 action chunk with 12 denoising forwards. A positive first-action y selects the left route and a nonpositive value selects the right route. The routes were executed and validated in the simulator during panel construction. Evaluation looks up the selected route's recorded outcome; it does not execute each learned proposal as a new closed-loop rollout. Consequently, route success is guaranteed by eligibility and is not an estimate of learned closed-loop competence.

Task and training. The instruction is *pick up the cream cheese box from either side and place it in the black bowl*. The training split contains 40 episodes and 15,800 actions. Starting directly from OpenVLA-7B, the discrete-diffusion adapter is trained for 30,000 optimizer steps with seed 39, LoRA rank 32, effective batch 8 (microbatch 1, accumulation 8), learning rate 5×10^{-4}, two images, proprioception, and image augmentation. The recorded decay threshold is 15,000 steps; there is no warmup, validation-set checkpoint selection, or resumed training. Only the final checkpoint is eligible. This is a separate fitted model from the two Choice fits, not retraining for any decoder condition. The fresh evaluation's common seed block is 392121.

Frozen panel and treatment. The 2026-09-21 protocol requires the first 54 eligible candidates in consecutive candidate order beginning at index 68, with simulator seed $3903 +$ candidate index. Eligibility requires a fixed ungrasped prefix, both-camera visibility, and successful correctly classified left and right 395-action controller routes with every waypoint reached. The retained context IDs are 104–157; panel generation makes no learned-model query. The same fixed 30,000-step seed-39 Approach checkpoint is used for every condition. The four decoders are coupled confidence, Split, random, and max confidence, each using the same two-draw construction described in Appendix P. Token/cosine decoding and no remasking are held fixed. Ten paired proposals per context and condition yield $54 \times 10 \times 4 = 2160$ learned route choices over 108 validated routes. The retained feasible panel, rather than a deployment population, is the inference target; it is not pooled with Choice or the earlier Approach panel.

Development of the execution interface. An earlier direct closed-loop development check produced no successful rollouts (three contexts, ten rollouts per decoder, four decoders), and its development gate did not pass. The subsequent fixed-route interface tests the learned selection separately from controller execution. This change preceded both the earlier 27-context route-choice panel and the present fresh-panel protocol; it does not repair or replace the failed direct closed-loop result.

Endpoint and chronology. For left and right counts n_L, n_R in a block of $K = 10$ independent proposals, the primary endpoint is

$$D = \frac{2n_L n_R}{K(K-1)}. \tag{19}$$

This is the fraction of ordered pairs of distinct proposals that choose different sides, an unbiased estimator of the corresponding two-sample disagreement probability. Its realized maximum for $K = 10$ is $5/9$, although the population probability is at most $1/2$. Coverage@10 measures whether each side appears at least once and is a different, secondary endpoint.

On the earlier completed 27-context panel, the Coverage@10 gain was positive but did not pass its confirmatory gate (approximately $+.074$, 95% interval $[-.019, .167]$, $p = .289$). Pairwise diversity was selected using that completed panel to avoid the Coverage@10 ceiling. The endpoint, fresh-panel selection, sample size, inference, and positive criterion were then frozen before any new candidate or learned query existed. The present result is therefore fresh-panel confirmation of a development-selected endpoint, not a reinterpretation of the earlier coverage test as positive.

Inference. All conditions and contrasts use the same 100,000 paired-context bootstrap resamples (seed 392122); intervals are percentile 95% intervals. Two-sided paired sign-flip tests drop zero differences, enumerate at most 20 nonzero signs, and otherwise use one million Monte Carlo sign patterns with plus-one correction (seed 392222). Inference assumes sign symmetry or exchangeability for the paired fixed-panel contrasts. The primary gate requires a Split-minus-coupled diversity gain of at least .05, an interval lower bound strictly above zero, $p < .05$, and success of all selected routes. Split-minus-random and Split-minus-max comparisons are secondary; their reported intervals and tests are not multiplicity-adjusted superiority tests.

Table 16: **Complete fresh Approach-Side comparison.** All 54 paired contexts and 540 proposals per condition. Left/right counts are pooled descriptives; diversity and Coverage@10 are context means. Success belongs to the previously validated selected routes. Intervals resample contexts, not individual proposals; the primary and secondary contrasts remain separate.

Decoder	Left/right	Diversity	Coverage@10	Route success
Coupled	458/82	0.245	0.870	1.000
Split	395/145	0.378	0.954	1.000
Random	371/169	0.405	0.963	1.000
Max	393/147	0.390	0.972	1.000

Diversity contrast	Gain	Paired 95% CI	Two-sided p
Split − coupled (primary)	$+0.133$	$[0.094, 0.173]$	$\approx 10^{-6}$
Split − random	-0.026	$[-0.070, 0.018]$	0.255
Split − max	-0.012	$[-0.038, 0.016]$	0.430

Outcome and scope. All primary checks pass. Split increases diversity by +0.133 [0.094, 0.173]; all four conditions and secondary comparisons appear in Table 16 and Figure 5. The result supports candidate–priority decoupling in this route-choice interface. It does not establish superiority over random/max, equivalence to them, complete-decoder transfer with remasking, or general learned closed-loop success on a second task.

Display and provenance. Figure 5A uses the lowest retained context ID (104), not a context selected for a large effect. Recorded end-effector coordinates are translated to the initial-object origin and converted from meters to centimeters, without smoothing. Panel B groups identical discrete estimator pairs by count, with area proportional to count; its category spacing is not a uniform numerical scale. All 54 contexts are retained. The result and completion receipt are byte-bound before rendering manuscript macros and the table. Full-workspace reconstruction is separate from the candidate anonymous artifact, which does not yet include this study; redistribution approval is not implied by successful local verification.

T Exact Construction and Additional Design Details

T.1 Choice data and chronological split

We derive ten soup and ten tomato demonstrations from independently replay-verified manipulation trajectories. Each soup demonstration has a tomato counterpart with the same first image and simulator state; proprioception matches to numerical precision. The 20 demonstrations contain 3,744 transitions and define $P_D(\text{soup}) = P_D(\text{tomato}) = 0.5$ at the episode level. Variable-length prefixes yield 2,051 soup-mode and 1,693 tomato-mode transitions, so the optimizer does not see a

transition-balanced dataset. Every truncated demonstration is replayed again after construction. Pairing removes between-mode differences in the demonstrated initial-observation distribution; unequal transition weighting, trajectory lengths, optimization, and tokenization can still bias the learned distribution.

State 13 is held out from training but used for method development, qualitative task repair, and a $K = 20$ screening gate; it is excluded from independent confirmation, and no intermediate checkpoint is selected from it. States 14–22 form a nine-state, three-seed confirmation with a large but family-wise underpowered state-clustered effect. Before inspecting the rest, the protocol is frozen for all default states 23–49. These 27 states exhaust the remaining list, without outcome-dependent stopping. The protocol fixes the terminal step-8,000 checkpoint, $K = 20$, seed 37, scheduler family, endpoints, and state-level inference with Holm correction. Exhaustive inclusion prevents within-suffix state selection, but this suffix is not a random deployment sample.

T.2 Exact construction and supporting studies

Balanced branch construction. For a concrete balanced binary branch, let

$$P(B, C) = \begin{pmatrix} 1/2 & 0 \\ \alpha - 1/2 & 1 - \alpha \end{pmatrix}, \quad \alpha \in (1/2, 1), \quad P(B = 0) = 1/2. \tag{20}$$

With independent marginal candidates and sampled-probability selection using Equation (4), exact conditional completion gives

$$P_{\text{decode}}(B = 0) = \frac{1}{2} + \frac{1 - \alpha}{2}[s_{\tau,\delta}(\alpha) - s_{\tau,\delta}(1 - \alpha)] > \frac{1}{2}, \tag{21}$$

where $s_{\tau,\delta}$ includes the 10^{-20} clamp and regularizer (Appendix B). At validity $v > 0$ and valid branch mass p, $\mathbb{E}[C_K] = 1 - \{(1 - vp)^K + [1 - v(1 - p)]^K\}/2$, uniquely maximized at $p = 1/2$ for $K \geq 2$. Imbalance therefore lowers expected finite-budget coverage at matched validity. The exact one-coordinate construction proves possibility, not mediation in the parallel 56-token learned decoder.

Trace instrumentation. At the first query we log entropy, token/probability summaries, mask state, first commit, and remasking for all 56×12 coordinate–iterations. Each primary scheduler yields 30,240 token histories and 362,880 token-step records; artifact checks confirm cardinalities and no unexpected valid mode. Expert chunks on ten disjoint states select a behavior-linked translation axis. Its final sign is validated against achieved modes; retrospective sign recoverability uses all rollouts on seven post-hoc common-support states. This analysis is exploratory (Figure 7).

Post-primary granularity study. The frozen 4×2 ablation uses the same 27 states, a new seed, $K = 10$, and 2,160 rollouts. Timestep scheduling groups seven coordinates; mask-count rounding changes trajectories across, but not within, granularities, precluding a pure grouping contrast (Appendix G).

Position-noise intervention. The prospectively frozen post-primary sweep varies first-query Gumbel scale $T \in \{.05, .25, .5, 1, 2, 4\}$, leaving categorical action sampling untempered. $T = 1$ is the upstream setting; $T = .05$ is the predeclared statistical reference. The same checkpoint and 27 states with seed 83 and $K = 10$ give 1,890 rollouts, including a separate random anchor excluded from the trend. The primary mean state-level Coverage@10 slope on ordinal scores 0–5 uses 100,000 within-state label permutations and state-bootstrap draws. Five secondary contrasts use Holm-adjusted sign-flip tests. The pre-run amendment, point-estimate validity guard, and assumptions appear in Appendix R.

T.3 Additional generative-policy context

Continuous Diffusion Policy may retrieve training chunks (He et al., 2026), a representation question our fixed-checkpoint comparison cannot resolve. Other VLA work learns diversity steering, behavior tokens, or denoising-trajectory RL (Tan et al., 2026; Kong et al., 2026; Wu et al., 2026b). BlockVLA and TBD-VLA impose temporal blocks; LLaDA-VLA hierarchically ranks actions and

Table 17: Primary prospective expansion over the complete census of 27 previously unopened default initial states and 540 rollouts per decoder. Holm p compares state-paired Coverage@20 against sampled-confidence-first.

Decoder	Valid /540	Success@20 /27	Coverage@20	ΔC_{20} [95% CI]	Holm p
Sampled-confidence	385/540	25/27	0.593	–	–
Random	390/540	26/27	0.833	+0.241 [0.148,0.333]	0.0007
Temporal-forward	387/540	26/27	0.833	+0.241 [0.130,0.352]	0.0034
Uncertainty-first	384/540	26/27	0.815	+0.222 [0.111,0.333]	0.0042

coordinates by confidence, while DIVA commits the highest-aggregate-confidence consecutive action group (Wang et al., 2026; Lee et al., 2026; Wen et al., 2025; Song et al., 2026). Coarse-to-Control predicts plan tokens before executable action tokens, but changes architecture and training rather than frozen-model reveal order (Wu et al., 2026a).

U Artifact Scope and Local Reproduction

The scheduler definitions, controlled intervention, endpoints, and statistical design are specified in Sections 3–5. Dataset construction and the prospective evaluation split are documented in Section 4; the complete frozen protocol, repaired-benchmark audit, earlier replication, negative Branch-Aware result, state-resolved outcomes, inference procedure, and trace schema appear in the appendix. The anonymous review artifact is a candidate allowlisted capsule containing frozen configurations, compact rollout CSVs, canonical summaries, the semantic-figure input, and a narrow subset of analysis utilities, protocols, tests, and renderers. Its five read-only statistical entry points independently reconstruct the reported behavioral endpoints from 32 unique CSVs containing 10,800 recorded rollout rows: primary/granularity, qualified seed-17 sensitivity, the complete-decoder control, the selection-coupling control, and the semantic-axis priority control. The decoder's reused primary baseline is not counted twice. These are separate fixed-panel study blocks, not pooled evidence from independent tasks. A separate script regenerates the included quantitative displays and primary table. The exact 125-file candidate passes isolated replay and focused tests. Candidate membership is not an authorship or redistribution-rights conclusion, and archive construction remains blocked pending accountable file-level provenance, license, and anonymity approval.

The capsule excludes inherited DD-VLA runtime files, checkpoints, demonstrations, robot datasets, large traces, logs, videos, and third-party visual assets. It instead records the upstream repository and exact revision with installation and integrity instructions. Its regeneration manifest identifies excluded displays, including the simulator-frame graphical abstract and raw-trace mechanism panel. It does not reconstruct the exploratory semantic-axis selection from the excluded traces or demonstrations. The seed-17 extension retains its original failed completion and separate technical-recovery audit (Appendix N). The decoder extension retains its machine receipt and the post-execution reused-baseline seed-pairing correction (Appendix K); one local checkpoint-path alias is explicitly distinguished from the original protocol bytes. Endpoint replay does not re-run excluded training, weight, historical-environment, or treatment-trace checks. The semantic-axis extension reconstructs behavioral endpoints, not its excluded raw traces or expert demonstrations; the earlier semantic diagnostic therefore remains a cached associational analysis. Other queued studies are not included in this endpoint subset. Every included component is byte-bound by the explicit allowlist; the archive manifest is generated only during an approved build. No archive has been approved for redistribution.

The completed position-temperature study is not included in that 125-file candidate. In the full workspace, the dedicated `paper/scripts/render_temperature_result.py` reads its frozen summary, renders the complete dose profile and three tables, and records input/output hashes. Its optional `-verify-replay` check recomputes the canonical analysis in memory from all seven CSVs (1,890 rows), without overwriting frozen results. This statistical replay does not repeat simulation or independently reconstruct raw Gumbel draws. Temperature-source packaging and redistribution approval remain outstanding; the main text and appendix report the full study regardless of that artifact boundary.

The selection-coupling extension includes all eight compact outcome CSVs, portable path-sanitized copies of the frozen result, completion receipt, protocol, and precision amendment, plus a read-only replay of coverage, validity, Success@10, bootstrap intervals, sign-flip tests, and Holm correction. It excludes the large raw traces, so it cannot independently re-run the treatment audit or proximal-score analysis. The second fit's original failed completion contract remains explicit in both the payload and replay output.

The semantic-axis extension includes all four compact outcome CSVs, the frozen protocol and completion receipt with only the local checkpoint locator aliased, and exact replay of both cached endpoint summaries. Its prespecified primary family is not supported. Raw traces are excluded, and a separately dated validity addendum was not bound by the completion source manifest; no validity-qualified positive claim relies on it.

The fresh Approach-Side study is documented in Appendix S, including the fixed-route execution boundary, development history, endpoint chronology, panel eligibility, and all four conditions. Its 2,160 learned proposals are route choices rather than additional closed-loop rollouts and are not part of the 10,800-row capsule count above. The full workspace binds its result and completion receipt before regenerating the manuscript table and macros. The new ten-figure system separately records source/output hashes, display transformations, and underlying values; it supersedes the older displays in this PDF, not the frozen contents of the candidate capsule. Neither the Approach source package nor the revised figure package is represented as already approved for anonymous redistribution. Those release decisions and the final author review remain outstanding.