HEPT: HUMAN EXECUTION PREFERENCE TRANSFER FOR ROBOT POLICY LEARNING

Anonymous authors
Paper under double-blind review

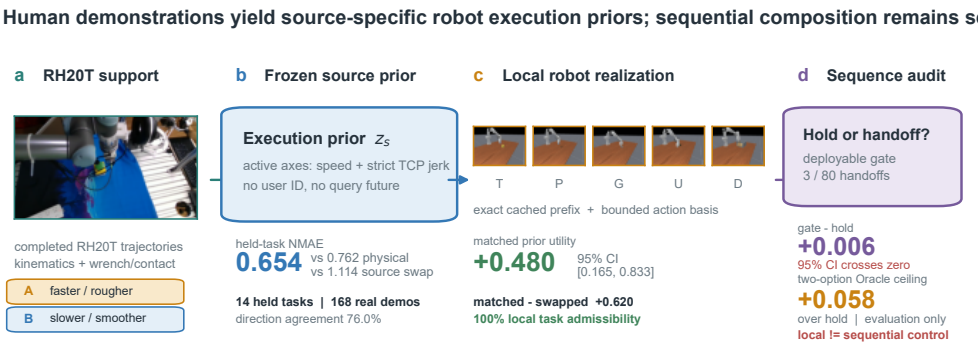

Figure 1: **Human demonstrations provide source-specific execution information that can be carried into task-preserving local robot behavior.** Representations from completed teleoperation predict held-out speed/jerk behavior beyond task, state, and source-swap controls. A physically matched counterfactual basis then exposes source-dependent robot effects. Sequential composition is evaluated separately, so a successful local transfer is not silently promoted to a whole-episode claim.

ABSTRACT

Human demonstrations teach robots more than task completion: they reveal repeatable, source-associated differences in how tasks are executed. We study Human Execution Preference Transfer (HEPT) through a traceable interface from completed teleoperation to a frozen robot policy. Without pairwise preference labels, language annotations, or operator identity as a model feature, a group-aware residual model predicts speed and physical-time tool-center-point jerk from kinematic and contact summaries on held-out tasks. We align these coordinates with bounded interventions around a frozen vision-language-action policy and evaluate local effects through exact-prefix counterfactual rollouts, separating outcome-aware reachability from deployable selection. In an independent RH20T confirmation with 168 demonstrations across 14 held-out tasks, the source representation achieves 0.6537 task-macro normalized error, compared with 0.7619 for an observable-state baseline, 0.8133 for a task mean, and 1.1136 for source swaps. Across five simulation task families, an outcome-aware counterfactual Oracle finds a task-admissible positive local response for both coordinates at all 90 tested anchors. On 20 fresh physical groups, a frozen matched source-profile prior yields utility of 0.4804 [0.1649, 0.8328] and exceeds source swap by 0.6202 [0.0889, 1.1880], with all matched selections locally task admissible. HEPT thus identifies source-specific execution information and demonstrates its task-preserving local realization; whole-episode policy adaptation remains a distinct problem.

1 INTRODUCTION

Robots are usually evaluated by *what* they accomplish: was the drawer opened, the object transported, or the peg inserted? Human demonstrations, however, also vary in *how* the outcome is produced. One demonstrator may move quickly and accept abrupt commands; another may trade time for smoother tool motion. These differences matter whenever a robot must operate around people,

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.

fragile objects, energy limits, or downstream controllers with different dynamics. Yet the prevailing imitation-learning objective tends to absorb such variation into a single policy, discard it as noise, or reproduce a trajectory without identifying what aspect should persist under a new context.

We study Human Execution Preference Transfer (HEPT) through a more precise question: *can completed human teleoperation demonstrations define an execution prior that remains informative on a held-out task and produces a source-specific, task-preserving change in another robot controller?* The learned object is an execution prior and its local intervention interface; the underlying VLA task policy remains frozen. This is not equivalent to cloning a trajectory. A new task geometry may require different actions to express the same tendency. Nor is behavioral variation automatically a subjective preference: speed can reflect task difficulty, interface latency, or skill. We therefore seek source-associated execution information under explicit controls and reserve the term *preference* for the bounded behavior expressed by the observed support set, not an enduring psychological trait.

The central difficulty is compositional. Three links must all be tested, and success at one cannot certify the next. First, **source sufficiency**: correctly associated demonstrations must predict held-out behavior beyond the task mean, deployment-observable physical context, and scene-matched demonstrations from the other source. Second, **coordinate meaning**: source and robot must measure the same quantity with declared sign, scale, sampling rate, and horizon. Third, **realization**: a robot intervention must move that coordinate in the requested direction while preserving task capability. This decomposition turns a vague "learn how humans act" claim into a sequence of falsifiable estimands.

Our implemented pipeline is shown in Figure 2. Completed RH20T demonstrations are summarized by direct speed/jerk/wrench descriptors and a structured kinematic–contact block. A nested, group-aware ridge procedure predicts held-out execution coordinates without using source identity or query future. Only speed and strict physical-time tool-center-point (TCP) jerk pass independent axis checks; wrench remains an input describing contact context but is not transferred as an output. On the robot side, exact cached prefixes branch into a fixed library of bounded action-chunk interventions around a frozen SmolVLA policy (Shukor et al., 2025). Paired rollouts measure the induced behavioral response and task change. Source coordinates weight those measured effects, whereas task admissibility is an independent constraint.

The resulting evidence is unusually diagnostic. The source representation passes every required control on an independent RH20T configuration. The robot candidate basis is locally reachable for both coordinates at every tested anchor across five manipulation families. Most importantly, a simple matched source-profile prior produces positive, source-specific local behavior on fresh physical groups while retaining task admissibility. A later sequence audit shows why this does not yet imply arbitrary closed-loop adaptation: learned handoff decisions do not improve upon holding the fixed prior, even though they suppress harmful alternatives. This boundary does not erase the positive result; it distinguishes the demonstrated object—a transferable local execution prior—from temporal composition, a separate learning problem.

Our contributions are:

- We formulate **HEPT** as three independently testable links: source sufficiency, matched physical coordinates, and task-preserving robot realization.
- We introduce a leakage-controlled source representation and confirmation protocol showing that completed real teleoperation demonstrations add source-associated information beyond task, observable-state, direct-descriptor, and source-swap controls.
- We construct an exact-prefix counterfactual response basis around a frozen VLA policy and show complete local task-safe reachability for speed and strict TCP jerk across five action families.
- We demonstrate source-specific local realization on fresh groups and provide a controlled sequence audit that localizes the remaining gap to temporal composition and option coverage, rather than to missing source information or local action reachability.

2 Related Work

Learning from human demonstrations. Imitation and one-shot learning treat demonstrations as examples of what action sequence should be reproduced (Finn et al., 2017; Chi et al., 2023; Zhao et al., 2023). Large robot datasets such as RH20T and DROID broaden tasks, scenes, sensors, and embodiments (Fang et al., 2024; Khazatsky et al., 2024); recent work shows that dataset composition and demonstration retrieval materially affect downstream generalization (Saxena et al., 2025). Multi-human imitation studies how to elicit or preserve compatible strategies across demonstrators (Gandhi et al., 2023), while Guided Strategy Discovery learns task-relevant latent behavior factors that interpolate or extrapolate beyond observed demonstrations (Sreeramdass et al., 2025). Our question is complementary: whether variation associated with the *source* remains predictive after task and observable physical context are controlled, and whether a physically named coordinate can parameterize execution on another controller.

Objectives, preferences, and interpretable coordinates. Inverse reinforcement learning infers objectives that rationalize demonstrated behavior (Abbeel & Ng, 2004; Ziebart et al., 2008; Levine & Koltun, 2012). Motion-style costs and concept bottlenecks motivate explicit intermediate variables (Zhou & Dragan, 2018; Koh et al., 2020). Preference-alignment methods often use comparisons or human feedback (Tian et al., 2024b;a; Li et al., 2025), while multi-objective formulations model heterogeneous experts (Yang et al., 2025). We receive neither preference labels nor online human judgments. Our representation is identified by cross-task prediction and source swaps, then grounded by robot interventions. Accordingly, we claim an *execution prior expressed by demonstrations*, not subjective intent or a complete model of human style.

VLA control and test-time steering. OpenVLA and SmolVLA demonstrate generalist vision-language-action control from large-scale pretraining and task-specific adaptation (Kim et al., 2024; Shukor et al., 2025). Test-time methods improve action chunks through resampling, predictive models, residual steering, or time-optimal replanning (Liu et al., 2025; Qi et al., 2025; Su et al., 2026; Kim et al., 2026). These methods usually optimize task success, policy likelihood, execution time, or a supplied reward. We instead hold the VLA task policy fixed and ask whether a prior inferred from independent human demonstrations can select bounded interventions by their measured physical consequences. Exact-prefix branching and source-swapped controls isolate execution transfer from task-policy retraining.

3 Problem Formulation

Let s index a demonstration source, q a held-out query task-context, and $\mathcal{S}_s = \{d_i\}$ the set of completed demonstrations available from source s on non-query tasks. Each d_i contains timestamped TCP, joints, gripper, and wrist wrench signals plus validity metadata. Source identity is used only to form matched and swapped support sets; its numeric value is never a feature.

Source representation. The source model maps completed support and deployment-visible query context to an execution coordinate,

$$\hat{\mathbf{z}}_s(q) = f_\theta\big(\phi(\mathcal{S}_s),\, x_q^{\text{prefix}}\big) \in \mathbb{R}^2, \tag{1}$$

where the two active axes are log mean-squared TCP speed and strict physical-time TCP jerk. The target is the source's held-out behavior on q, not an annotated preference. Source sufficiency requires the matched support to beat: (i) a task mean, (ii) a Ridge model using only the query's first-second physical observation, and (iii) a scene-matched support set from the other source. The source swap is crucial: a model can exploit task content without using the correct source.

Robot response and local utility. At robot history h, the frozen base policy proposes an action chunk. A bounded candidate $\delta \in \mathcal{D}$ modifies only the arm actions and leaves the gripper byte-identical. Exact-prefix paired rollouts define

$$\Delta\mathbf{b}(h,\delta) = \mathbf{b}(\tau_{h,\delta}) - \mathbf{b}(\tau_{h,0}), \tag{2}$$

where $\mathbf{b}$ contains speed and strict TCP-jerk changes over a fixed response window. Axis scales are frozen from development data. With signs oriented toward the source coordinate, local utility is

$$U_s(h,\delta) = \langle \mathbf{z}_s, \Delta\mathbf{b}(h,\delta) \oslash \boldsymbol{\sigma} \rangle\,, \quad \text{s.t.} \quad A(h,\delta) = 1, \tag{3}$$

where A is an independently evaluated task-admissibility constraint. We do not fold task preservation into the preference score.

Sequence composition. A locally useful intervention may change the state distribution seen by a later decision. For a second decision $a \in \{\text{HOLD}, \text{HANDOFF}\}$, we therefore evaluate complete post-entry sequence value,

$$\Delta U_{\text{seq}}(h) = U_s(\text{HANDOFF} \mid h) - U_s(\text{HOLD} \mid h). \tag{4}$$

This estimand is deliberately separate from local reachability. A sequence gate may hand off only when a cross-fitted lower-confidence estimate of ΔU_{seq} is positive.

4 METHOD

4.1 RECOVERING A SOURCE EXECUTION PRIOR

We compute three direct whole-trajectory descriptors: log mean-squared TCP speed, log mean-squared physical-time TCP jerk, and log mean-squared zeroed translational wrist wrench. A 49-dimensional structured block augments these with robust kinematic, joint, wrench, phase, and contact summaries. RGB and gripper-derived coordinates are retained for audit but excluded from the frozen transfer interface because development ablations did not establish incremental value. Wrench remains useful as contact context but is not an active output axis.

Predictions use a direct Ridge backbone plus a structured residual: $p = p_{\text{direct}} + \gamma(p_{\text{structured}} - p_{\text{direct}})$. For each held source–task unit, regularization and γ are selected only by leave-one-task validation inside the outer training bank. All centering and feature scaling are fit inside that bank. The complete source scene is excluded from both query centering and support construction. The model never receives numeric operator identity, query duration, endpoint, rating, future signal, or robot outcome.

Learned and frozen components. The source predictor and deployment-visible response selectors are learned. The physical coordinate definitions, action-candidate library, task-admissibility test, and V23 profile-to-candidate mapping are frozen protocol components; the outcome-aware Oracle is never used at deployment. This inventory identifies which conclusions arise from prediction, controlled intervention, or evaluation and keeps the implemented estimator visible behind the notation.

4.2 FREEZING THE CROSS-EMBODIMENT COORDINATE CONTRACT

Names such as "smoothness" are insufficient for transfer. We match the source and simulator definitions by physical derivative order, sampling rate, window, sign, and scale. Speed uses mean-squared TCP velocity. Jerk is the third physical-time derivative of TCP position estimated at 10 Hz with a nine-point cubic local polynomial and evaluated over a 0.9 s window. The two coordinates are standardized within domain and interpreted directionally; we do not claim equality of force sensor units or raw cross-robot magnitudes. Axis-specific source audits activate speed and jerk and reject wrench as an output.

4.3 EXACT-PREFIX COUNTERFACTUAL RESPONSE BASIS

The robot evaluator uses ManiSkill3 (Tao et al., 2024) and a frozen SmolVLA task policy. Five task families—transport, pick-place, peg insertion, push, and drawer—span free-space, contact, and near-terminal phases. For each anchor, the simulator and policy states are cached once, then restored before executing the base chunk or one of 19 bounded candidates. Candidates implement small scale, bridge, and joint-direction perturbations; they preserve the gripper exactly and are rejected if action projection occurs. Paired randomness ensures that Equation (2) measures the candidate effect rather than a different prefix or environment draw.

Task admissibility is evaluated independently from behavior utility. A branch must preserve local goal progress, endpoint error, joint margin, and local success under frozen tolerances. An outcome-aware Oracle is used only to test whether the candidate basis contains a solution. Deployment controls

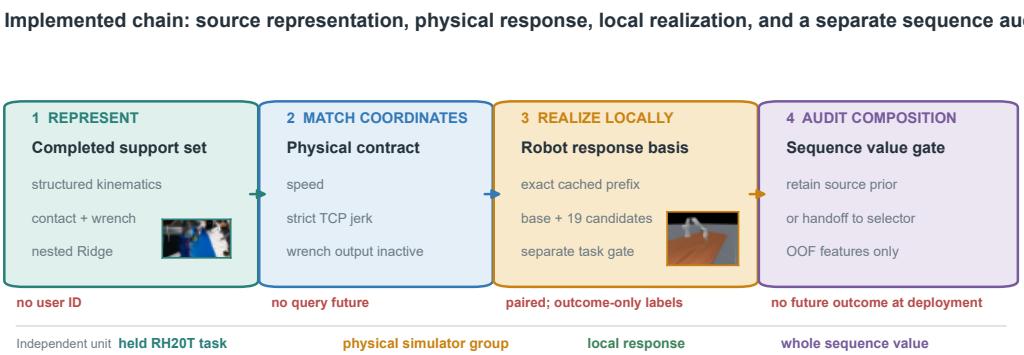

Figure 2: **HEPT recovers a source prior, matches physical coordinates, and realizes or audits robot behavior.** The representation uses completed support demonstrations but withholds source ID and the query future. Robot responses are measured by exact-prefix branches around a frozen VLA controller. Local realization and sequence composition are distinct decisions with separate controls.

include a fold-local candidate prototype, a direct response predictor, ID-only and action-contrast selectors, and exact-base continuation.

4.4 Source-specific local realization and sequence audit

Two frozen source profiles are induced by the independently estimated RH20T coordinates: a faster/rougher profile and a slower/smoother profile. A simple low-variance prior maps these profiles to fixed members of the candidate basis. The mapping was frozen on earlier development data before the fresh confirmation groups were generated. The decisive control swaps the two source mappings while keeping the task, prefix, candidate library, and execution budget unchanged.

To test composition, a later experiment begins with the matched fixed prior and compares retaining it with handing off to a frozen action-contrast selector. The final sequence gate uses 72 deployment-visible features ending before the second intervention and a five-member random-forest ensemble under four-fold physical-group cross-fitting. It defaults to HOLD and hands off only when mean-minus-standard-deviation predicted sequence gain is positive. This is an audit of whether local source realization survives another decision, not a claim that the gate is a successful whole-episode controller.

5 Experimental Design

We organize evaluation by four research questions instead of experiment chronology.

RQ1: Does the source representation add held-out information? The independent cfg3 confirmation contains 168 real RH20T teleoperation demonstrations, 14 tasks, two operators, six matched scenes per operator–task cell, and 28 outer source–task units. Each unit holds out one operator's entire target task. The bootstrap unit is the held-out task, not an individual trajectory. The primary metric is task-macro normalized mean absolute error (NMAE; lower is better).

RQ2: Are the matched coordinates locally reachable? Thirty independent simulator physical groups contain three anchors each. The base plus 19 candidates give 1,710 exact-prefix counterfactual branches. The primary reachability statistic is group-macro task-safe positive gain, bootstrapped over physical groups. The Oracle observes outcomes and is explicitly non-deployable.

RQ3: Does the source change robot behavior specifically? The fresh confirmation uses 20 procedurally selected physical groups, four per family, two inference seeds, and both source profiles. The 440-run study compares matched prior, source swap, ID-only, action contrast, frozen direct response, and Oracle. All mappings and learned controls are frozen before outcomes. The independent unit is the physical group.

RQ4: Does local realization compose over time? Twenty disjoint groups yield 80 source-conditioned sequence contexts. Four-fold cross-fitting holds out one group per family in every fold. We compare hold, always-handoff, a fold-local prototype, the conservative gate, and a post-hoc two-option Oracle that chooses the better observed option per context. The Oracle is a diagnostic ceiling, not a method.

Across robot studies, actions must be finite and bounded, arm projection must be zero, gripper commands must match exactly, and paired prefixes must pass state/action/TCP/object equality checks. Protocols, groups, features, metrics, thresholds, bootstrap units, and stopping rules are frozen before the corresponding outcomes. Full details and the development chronology are in Sections D and G.

6 RESULTS

6.1 CORRECTLY ASSOCIATED DEMONSTRATIONS PREDICT HELD-OUT EXECUTION

Figure 3 and Table 1 answer RQ1. The frozen structured representation obtains **0.6537 NMAE**, compared with 0.8133 for the task mean, 0.7619 for the physical-prefix Ridge baseline, 0.6978 for three direct trajectory descriptors, and 1.1136 when the same scenes are paired with the other source. Its paired improvements over task mean, physical prefix, source swap, and direct descriptors are all positive: respectively 0.1596 [0.0843, 0.2269], 0.1082 [0.0559, 0.1640], 0.4599 [0.3778, 0.5524], and 0.0441 [0.0131, 0.0707]. Correct support beats swapped support on all 14 held-out tasks. Direction agreement is 75.99% [69.64%, 82.14%].

The axis audit clarifies what transfers. Speed and jerk direction agreement are 86.9% and 86.3% and each beats all required controls. Wrench direction is only 54.8% and its intervals against the references cross zero. The evidence therefore supports a two-dimensional execution prior, not an unrestricted style vector. Contact and wrench features still improve representation, but force is not presented as a transferred goal.

6.2 THE ROBOT ACTION BASIS IS LOCALLY SUFFICIENT

For RQ2, the outcome-aware Oracle finds at least one task-safe positive candidate at all 90/90 anchors for both axes. Group-macro gains are 0.5220 [0.3953, 0.6650] for speed and 0.3467 [0.2787, 0.4208] for strict jerk, with all five family means positive. Thus the fixed bounded library spans useful local responses; inability to select among them cannot be attributed to an empty action basis.

6.3 A MATCHED HUMAN SOURCE PRIOR PRODUCES SOURCE-SPECIFIC ROBOT BEHAVIOR

RQ3 is the key realization result. On 20 fresh groups, the matched fixed prior attains 0.4804 [0.1649, 0.8328] group-macro utility. Both source profiles are positive and the matched selections are 100% locally task admissible. Swapping the source mapping reverses much of the benefit: matched minus swapped is 0.6202 [0.0889, 1.1880]. Matched also beats the frozen Direct reselector by 0.4535 [0.0852, 0.8512]. Against the two parameter-matched learned controls, matched minus ID-only is 0.1401 [-0.2485, 0.5358], and matched minus action contrast is 0.0126 [-0.3674, 0.3861]; these intervals cross zero. We therefore use source swap, rather than generic selector utility, as the decisive source-specificity comparison.

The result is not merely a constant good action. The two source coordinates map to different fixed candidates, the source swap holds every other factor fixed, and the matched-minus-swapped interval is strictly positive. A learned action-contrast selector is also positive after the matched-prior entry (0.4678 [0.2245, 0.7115]) and has positive point estimates in all five families. The fixed prior is slightly negative on push (-0.0447), so we do not claim uniform family-wise dominance. The supported claim is narrower and still substantive: real-demonstration source information causes a positive, source-specific, task-preserving local robot response on new physical groups.

6.4 SEQUENCE COMPOSITION IS DISTINCT FROM LOCAL REALIZATION

RQ4 asks whether a useful local prior should be replaced by a later learned decision. Holding the prior yields 0.1669 [-0.0458, 0.3998]; always handing off yields -0.1029 [-0.2620, 0.0599]. The

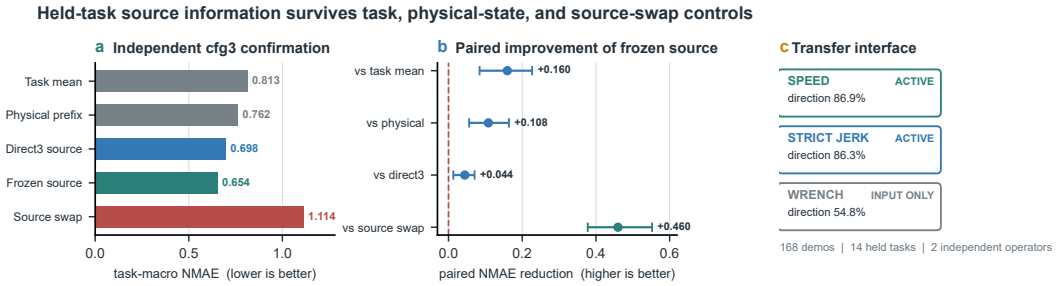

Figure 3: **The frozen source representation beats task, observable-state, direct-descriptor, and source-swap controls.** Results are from the independent cfg3 confirmation; intervals resample held-out tasks. Speed and jerk support the transfer interface, while wrench is retained only as input context.

Table 1: **Source sufficiency on the independent RH20T confirmation.** NMAE is task-macro and lower is better. Improvements compare the frozen source representation with each reference; intervals bootstrap the 14 held-out tasks.

Method	NMAE ↓	Frozen-source improvement
Task mean	0.8133	0.1596 [0.0843, 0.2269]
Physical prefix Ridge	0.7619	0.1082 [0.0559, 0.1640]
Scene-matched source swap	1.1136	0.4599 [0.3778, 0.5524]
Direct trajectory descriptors	0.6978	0.0441 [0.0131, 0.0707]
Frozen structured source	**0.6537**	–

conservative gate hands off only 3/80 times and attains 0.1725 [-0.0404, 0.4041]. It significantly beats always-handoff by 0.2754 [0.1405, 0.4226], showing that abstention detects many poor transitions. It does not improve over hold: 0.0057 [-0.0035, 0.0191].

Could a better classifier solve the same decision? A post-hoc Oracle that sees both realized outcomes and picks the better of hold and handoff reaches 0.2253 [0.0206, 0.4538], improving over hold by 0.0584 [0.0277, 0.0943]. This gap establishes useful decision heterogeneity. However, even that Oracle remains negative in transport and pick-place. The remaining problem is therefore not one scalar threshold: the tested sequence offers an incomplete option set in two families, while deployment-visible prediction must also identify the contexts where handoff helps. The sequence audit is a mechanistic result, not a failed restatement of source transfer; Section E reports the complete comparison and family map.

7 Discussion

What has been learned from humans? The source model learns regularities expressed across completed teleoperation demonstrations. Correct source association matters after task, scene, and early physical state are controlled, and source-swapped robot interventions produce different effects. This supports a source-specific execution prior. It does not establish that the two operators represent a population, that the behavior is a stable personality trait, or that every linguistic notion such as "gentle" is captured by speed and jerk.

Why an explicit execution prior? A direct support model is a strong baseline and indeed performs well. The explicit coordinate adds two capabilities. First, it is auditable across embodiments: speed and physical-time jerk have defined sign, scale, and response windows. Second, the same coordinate can score measured candidate effects while task preservation remains a separate constraint. The gain over direct descriptors in RQ1 and over source swap in RQ3 shows that the structured prior is not only an explanatory ornament. We nevertheless avoid claiming universal superiority over every learned candidate selector.

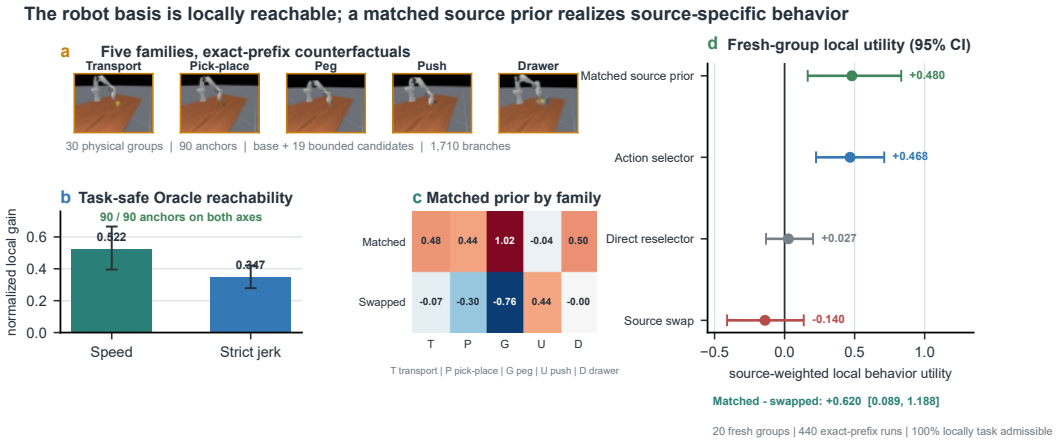

Figure 4: **Matched physical coordinates are locally reachable, and a source-specific prior realizes positive behavior without task loss.** Left: an outcome-aware Oracle finds a task-safe positive candidate at all 90 anchors for both axes. Right: on fresh V23 groups, the matched prior is positive and significantly exceeds source swap and frozen Direct. Oracle coverage and deployable realization are intentionally shown as different evidence.

Table 2: **The evidence hierarchy separates reachability, source specificity, and sequential composition.** Utilities use the frozen source coordinate and physical-group bootstrap. Outcome-aware, deployable, and diagnostic rows are labeled explicitly rather than treated as a common leaderboard.

Evidence	Scope	Result and interpretation
Candidate-basis Oracle	Local; outcome-aware	On 30 groups / 90 anchors, speed is 0.5220 [0.3953, 0.6650] and jerk is 0.3467 [0.2787, 0.4208]; a task-safe positive candidate exists at every anchor.
Matched source prior	Local; frozen; fresh groups	On 20 groups / 440 runs, utility is 0.4804 [0.1649, 0.8328] and matched–swap is 0.6202 [0.0889, 1.1880]; all matched selections are task admissible.
Learned selector controls	Local; frozen; fresh groups	ID-only is 0.3403 [0.1561, 0.5320], action contrast is 0.4678 [0.2245, 0.7115], and Direct is 0.0269 [-0.1329, 0.2033]. Matched is source-specific but not significantly above ID-only or action contrast.
Sequence audit	Two decisions; cross-fitted	The conservative gate is tied with hold: 0.0057 [-0.0035, 0.0191]. The full option-set and Oracle analysis is reported in Section E.

Local behavior is not a whole-episode controller. Exact-prefix counterfactuals provide strong causal identification of local action effects, but long-horizon realization must decide when to intervene, continue, switch, or exit as the state distribution changes. Our sequence audit shows that naively composing locally useful decisions can remove their benefit. This distinction is especially important for action-chunk policies, where temporal coherence and reactivity trade off (Liu et al., 2025). A promising next step is a richer set of source-consistent sequence primitives, evaluated on new groups, rather than another threshold search over the same hold/handoff pair.

Scope and limitations. RH20T supplies real robot teleoperation trajectories, whereas causal realization is tested in simulation around a frozen VLA policy. We do not claim hardware deployment, safety certification, subjective human preference recovery, or unrestricted execution style. The source confirmation contains two operators; the robot tests contain five manipulation families and local horizons. Oracle results diagnose reachability and option coverage but are not deployable. These boundaries are explicit because they separate the established contribution from the next scientific problem.

8 Conclusion

Human demonstrations contain execution structure that task success alone does not describe. HEPT treats this structure as a physically grounded execution prior. By separating source sufficiency, physical coordinate matching, and robot realization, we show that this structure can be measured on held-out real demonstrations and carried into source-specific, task-preserving local behavior around a frozen robot policy. The same decomposition reveals why local success does not automatically compose: temporal handoff and sequence-option coverage remain distinct learning problems. The resulting object—a physically grounded human-demonstration execution prior—provides an inspectable interface for studying how robots should act, not only whether they finish the task.

Reproducibility Statement

The complete evaluation contracts, group-aware splits, feature schemas, candidate definitions, hashes, random seeds, bootstrap procedures, stopping rules, and numerical audits are documented in Sections D, E and G and in the anonymous artifact package. RH20T episode identifiers and hashes are provided without redistributing access-controlled raw data; selected simulator evaluation outputs, analysis tables, and exact figure-generation code are included directly.

AI Use Statement

Generative AI tools assisted with experiment planning, software implementation, data-processing utilities, result interpretation, literature organization, figure construction, and manuscript drafting and editing. The authors inspected the underlying experimental artifacts, independently checked all reported numbers against machine-readable results, reviewed AI-assisted text and code, and take responsibility for the final claims and content.

Ethics Statement

No new human participants or subjective labels were collected. RH20T data are used under the dataset's existing terms, and source identifiers are used only to construct matched controls. They are not model inputs or biometric claims. Simulation task admissibility is not a hardware-safety guarantee; further force, collision, latency, and operator-consent validation is required before physical deployment.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *IEEE International Conference on Robotics and Automation*, pp. 653–660, 2024.

Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot visual imitation learning via meta-learning. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 357–368, 2017.

Kanishk Gandhi, Siddharth Karamcheti, Madeline Liao, and Dorsa Sadigh. Eliciting compatible demonstrations for multi-human imitation learning. In *Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1981–1991, 2023.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yu, Soroush Chen, et al.

DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Pannag 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. *arXiv preprint arXiv:2406.09246*, 2024.

Sunwoo Kim, Jeongjun Kim, and Joseph Lim. Time optimal execution of action chunk policies beyond demonstration speed. In *International Conference on Learning Representations*, 2026.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020.

Sergey Levine and Vladlen Koltun. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Chenliang Li, Siliang Zeng, Zeyi Liao, Jiaxiang Li, Dongyeop Kang, Alfredo Garcia, and Mingyi Hong. Joint reward and policy learning with demonstrations and human feedback improves alignment. In *International Conference on Learning Representations*, 2025.

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, and Chelsea Finn. Bidirectional decoding: Improving action chunking via guided test-time sampling. In *International Conference on Learning Representations*, 2025.

Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Inference-time enhancement of generative robot policies via predictive world modeling. *IEEE Robotics and Automation Letters*, 2025.

Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. In *International Conference on Learning Representations*, 2025.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, Simon Alibert, Matthieu Cord, Thomas Wolf, and Remi Cadene. SmolVLA: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Varshith Sreeramdass, Rohan Paleja, Letian Chen, Sanne van Waveren, and Matthew Gombolay. Generalized behavior learning from diverse demonstrations. In *International Conference on Learning Representations*, 2025.

Entong Su, Tyler Westenbroek, Anusha Nagabandi, and Abhishek Gupta. RFS: Reinforcement learning with residual flow steering for dexterous manipulation. In *International Conference on Learning Representations*, 2026.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: Gpu parallelized robotics simulation and rendering for generalizable embodied AI. *arXiv preprint arXiv:2410.00425*, 2024.

Ran Tian, Yilin Wu, Chenfeng Xu, Masayoshi Tomizuka, Jitendra Malik, and Andrea Bajcsy. Maximizing alignment with minimal feedback: Efficiently learning rewards for visuomotor robot policy alignment. *arXiv preprint arXiv:2412.04835*, 2024a.

Ran Tian, Chenfeng Xu, Masayoshi Tomizuka, Jitendra Malik, and Andrea Bajcsy. What matters to you? towards visual representation alignment for robot learning. In *International Conference on Learning Representations*, 2024b.

Jun-Jie Yang, Qian-You Zhang, Chia-Heng Hsu, Xi Liu, and Ping-Chun Hsieh. Learning from diverse experts: Behavior alignment through multi-objective inverse reinforcement learning. In *ICLR Workshop on Bidirectional Human–AI Alignment*, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

Allan Zhou and Anca D. Dragan. Cost functions for robot motion style. *arXiv preprint arXiv:1809.00092*, 2018.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

A Evidence Map and Claim Boundaries

The manuscript uses *execution prior* for a behavioral coordinate inferred from completed demonstrations. It does not use *human preference* as a synonym for a subjective judgment. The operator field defines support association and matched controls but is never an input feature. "Fresh physical group" denotes a newly generated simulator identity, not a previously unseen semantic task category.

B RH20T Source Representation

B.1 Cohort and observables

The independent confirmation uses RH20T cfg3, operators 10 and 11, 14 tasks, and six matched scenes for each operator–task cell, totaling 168 real teleoperation demonstrations. The development cohort used a different configuration and different operators. In each of 28 outer cells, one target operator's complete task is held out as query; fitting uses both operators on the remaining 13 tasks. Test support consists of 78 completed demonstrations from the target operator on non-query tasks. The task prior uses the alternate operator's six episodes of the query task. Correct and mismatched supports use the same task/scene structure and differ only in source association.

All candidate files are screened for valid TCP, joint, gripper, wrench, and camera timestamp streams before hash-based selection. The first version of this confirmation stopped before scoring because of one numeric stream discontinuity. V11.1 screened the complete candidate pool first and reapplied the original hash rule; it did not change tasks, models, metrics, or gates and did not reuse partially extracted files.

B.2 Descriptor contract

For a timestamped TCP position $p(t)$, direct speed is the log of mean squared $\|\dot{p}(t)\|_2$. Direct jerk is the log of mean squared $\|d^3p(t)/dt^3\|_2$ using the declared physical sampling interval. Wrench uses the zeroed translational wrist signal. The structured block contains robust moments and quantiles of translational speed, acceleration, jerk, angular speed, joint speed, force and torque; path and timing summaries; four phase bins; and contact duty/onset summaries. The frozen transfer slice includes indices 0–48. Gripper coordinates 49–56 and visual coordinates 57–95 are audit-only.

The physical-prefix control has 38 numeric and 39 visual summaries computed strictly from the first second. Every interpolation query is restricted to raw timestamps at or before the cutoff; the last available sample is held at the boundary. Query endpoint, duration, later samples, ratings, and task outcomes are prohibited.

B.3 Nested prediction and scaling

For every outer cell, all feature scaling, target normalization, Ridge fitting, and model selection use only the outer training bank. The direct model uses the three direct descriptors. The structured residual uses coordinates 0–48. The convex residual strength $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$ and Ridge penalty $\alpha \in \{10, 10^2, 10^3, 10^4, 10^5, 10^6\}$ are selected by leave-one-task validation inside the outer bank. Ties prefer stronger regularization and smaller residual strength. All 28 outer cells selected $\gamma = 1$, and no selected penalty touched the frozen upper boundary.

Task-macro NMAE first normalizes absolute error by the training-bank target scale and then averages queries within task and tasks equally. The 95% intervals resample the 14 held-out tasks 10,000 times with seed 20261221. The direction score records whether predicted and observed deviations from the available-bank task center have the same sign.

C Physical Coordinate Contract

The active bridge has two coordinates.

1. **Speed**: mean-squared translational TCP velocity over the declared window, log transformed for source prediction and standardized within domain for directional transfer.

Table 3: Evidence-to-claim map. No row is used to support a stronger scope than its design permits.

Evidence block	Independent unit	Positive claim	Explicit boundary
V11.1 RH20T	14 held-out tasks	Correct source support improves held-out speed/jerk prediction beyond task, state, direct, and source-swap controls	Two operators; observational source association; no subjective label or robot intervention
V14 response basis	30 physical groups	The 19-candidate basis contains task-safe positive local effects for both active axes in all five families	Outcome-aware Oracle; development groups; not deployable selection
V17 direct response	20 physical groups	Deployment-visible inputs can select positive local response for both profiles and all five families	Local anchors; frozen control inside a larger study whose primary stage-table gate did not pass
V23 matched prior	20 fresh physical groups	A frozen matched source prior has positive, source-specific, locally task-admissible effect	One family has a slightly negative point estimate; no whole-episode claim
V24–V25 sequence audit	20 fresh groups per study	Holding the prior is stronger than the tested learned handoff; abstention rejects harmful handoffs; option coverage is limited	Cross-fitted development/audit evidence; no deployment checkpoint or successful sequence controller

Table 4: Axis-localized V11.1 results. Lower NMAE is better. The axis audit was performed after the frozen aggregate decision to determine the permitted cross-embodiment interface; it does not replace the official evaluation.

Axis	Task mean	Physical	Direct	Frozen source	Source swap
Speed	0.8139	0.7434	0.6420	**0.5723**	1.2670
Jerk	0.8364	0.7653	0.6661	**0.6029**	1.3027
Wrench	0.7896	0.7768	0.7853	0.7859	0.7710

2. **Strict jerk**: mean-squared third physical-time derivative of translational TCP position. Both source and robot use a nine-point cubic local-polynomial estimator at 10 Hz and a 0.9 s evaluation window.

The contract fixes derivative order, time units, sampling rate, window, sign, and scale before behavior outcomes. Historical labels called "smoothness" or second position difference are not used. Source and simulator coordinates are matched by directional standardized dose; raw force units are not equated across embodiments. Wrench is allowed as a context/safety observation but not as a desired output because its source gate does not pass.

D Counterfactual Robot Response Protocols

D.1 Task families and base policy

Robot experiments use five manipulation families: transport, pick-place, peg insertion, pushing, and drawer actuation. Groups vary task definition, initial state, physical level, and inference seed. The base policy is a frozen SmolVLA controller receiving fixed-camera RGB, Panda state, and the task instruction. The policy produces action chunks; no waypoint controller or scripted fallback is used in the counterfactual evaluations reported here.

The 19 non-base candidates are positive and negative perturbations of seven joint directions, positive/negative scale changes, and bridge/anti-bridge chunk reshaping. Each intervention lasts ten 50 ms control steps and is followed by ten fixed base recovery steps. Arm actions are bounded; any projection fails the integrity audit. Gripper commands are copied byte-for-byte from the base chunk.

D.2 CACHED-PREFIX CAUSAL INTEGRITY

For every anchor, the environment, policy, and recurrent state are materialized once. Child branches restore the exact parent state rather than separately replaying stochastic prefixes. Audits compare parent and child action tapes, Panda state, joint positions, TCP, and object state. The maximum accepted prefix error is 3×10^{-6} in earlier replay protocols and exact equality for the later cached-parent collector. Results are counted only when cardinality, hash, horizon, finite/bounded action, projection, gripper, intervention-effect, and prefix checks all pass.

D.3 BEHAVIOR AND TASK METRICS

Behavior utility uses the source-oriented dot product of speed and strict-jerk response after division by frozen historical scales. A positive number means that the candidate moves the robot toward the active source coordinate relative to exact base. Task admissibility is separate: progress gain must be at least -0.0015 m, endpoint squared-error gain at least -10^{-4} m^2, joint-margin gain at least -0.005 rad, and local-success loss no worse than the frozen limit. Later prior/sequence studies inherit the progress and endpoint clauses used by their parent protocol. We report both utility and task admissibility; one cannot compensate for the other.

D.4 V14 REACHABILITY AND DEPLOYABLE RESPONSE CONTROLS

V14 contains 30 groups, six per family, three anchors per group, and 19 candidates per anchor (1,710 branches). The Oracle selects the highest measured axis gain among task-safe candidates and is used only for reachability. Five-fold group CV keeps every group's anchors and candidates together. Runtime features include family, physical level, five-step state/TCP/force histories, four-step action history, previous action, base and candidate chunks, and basis coefficient. Prohibited features include script phase, simulator object/goal truth, future trajectory, candidate outcome, success, and sealed-group data.

The V14 learned speed selector passes its complete gate, while learned strict jerk selection does not. This is why the main paper reports V14 primarily as candidate-basis reachability. The separately frozen V17 direct-response control is reported because it runs zero-shot on disjoint groups, is positive for both profiles and all families, and satisfies the declared task constraint. It does not retroactively change the primary V17 stage-table decision.

D.5 V23 SOURCE-PROFILE PRIOR CONFIRMATION

The two source coordinates and fixed mappings are:

$$z_A = (0.4973, -0.5027), \qquad \delta_A = \texttt{anti_bridge},$$
$$z_B = (-0.4958, 0.5042), \qquad \delta_B = \texttt{scale_neg}.$$

The profile-to-candidate mapping was stored in the V22 model freeze before any V23 group existed. Twenty procedurally fresh groups (four per family, at least two task definitions per family) exclude all prior identities. Two inference seeds and both profiles yield 40 roots, 80 parents, and 320 candidate children, or 440 runs. At call 5 the parent executes the matched source prior; after two exact-base blocks, call 8 branches into base, scale-negative, anti-bridge, and joint-0-negative candidates. Controls are exact-base exit, source swap, frozen ID-only and action-contrast selectors, frozen Direct reselect, and task-safe Oracle. Nothing is refit.

Complete-group bootstrap uses 20,000 draws with seed 230120. The fixed matched prior is positive overall and source-specific. Family means are transport 0.4822, pick-place 0.4395, peg insertion 1.0215, push -0.0447, and drawer 0.5034. Hence four of five family point estimates are positive; the manuscript does not state that every family passes.

E Sequence Composition Audit

V24 tests a fixed two-decision sequence on 20 new groups. The first decision is the matched prior; the second either retains that prior or hands off to the unchanged action-contrast selector. All 440 runs pass exact shared-root and prefix checks and are task admissible. Handoff utility is -0.1029 $[-0.2626, 0.0603]$, while retaining the matched prior has point estimate 0.1669. Handoff minus hold is -0.2698 $[-0.4136, -0.1383]$.

V25 uses the opened V24 contexts for one prespecified sequence-level attempt. Features end at step 59, before execution of the call-8 option, and contain 72 deployment-visible values: family, physical difficulty, source coordinate, pre-entry and observed-entry TCP behavior, prefix state/action/force summaries, base chunk summaries, fixed-prior and proposed-candidate action contrasts, and the frozen selector's scores. Goal truth, object future, outcomes, group ID, operator ID, and steps 60 onward are prohibited.

Four-fold cross-fitting leaves one physical group from each family out per fold. Each fold fits five random forests (300 trees, depth 4, minimum leaf 4, feature fraction 0.7) with fixed seeds 2501–2505 and no hyperparameter search. The gate selects handoff only when the ensemble mean minus one standard deviation is positive. The complete-group bootstrap uses 20,000 draws and seed 250120.

The post-hoc two-option Oracle is computed only from stored out-of-fold rows: for each context it selects $\max(U_{\text{hold}}, U_{\text{handoff}})$. Its family utilities are transport -0.0326, pick-place -0.1166, peg insertion 0.3075, push 0.3817, and drawer 0.5864. Thus perfect selection within the tested pair still cannot make two families positive. This analysis is an evaluation ceiling, not a new trained method.

F Statistical Conventions

All confidence intervals are percentile bootstrap intervals and all primary robot intervals resample complete physical groups. RH20T intervals resample held-out tasks. No rollout, anchor, or trajectory is treated as an independent unit when multiple observations share a task or physical group. Paired method differences are recomputed inside each bootstrap draw. Reported intervals are two-sided 95% intervals; no multiple-comparison correction is used because the primary comparisons and stopping gates were declared per protocol. Descriptive subgroup means are not promoted to significance claims without intervals.

G Development Chronology and Stopping Decisions

The chronology matters for interpretation. Later positive controls do not retroactively change earlier preregistered decisions. Likewise, a stopped sequence route does not negate independently confirmed source sufficiency or local realization. This manuscript reports the final evidence hierarchy rather than selecting only experiments whose full conjunctive gate passed.

H Reproducibility and Artifact Package

Every decisive protocol is represented by a human-readable specification and a machine-readable artifact with file hashes, source/checkpoint hashes, group identities, seeds, metrics, and stopping rules. The anonymous artifact package includes the decisive evidence used to verify the paper's reported numbers:

- the V11.1 frozen protocol, aggregate source-sufficiency evaluation, preflight record, and axis audit;
- the V14 frozen Oracle protocol, collection-integrity report, Oracle evaluation, and grouped response-model summary;
- the V17 aggregate evaluation and the V23 group manifest, aggregate evaluation, and independent audit;
- the V25 model contract, cross-fit prediction array, aggregate evaluation, independent numerical audit, and two-option Oracle ceiling;

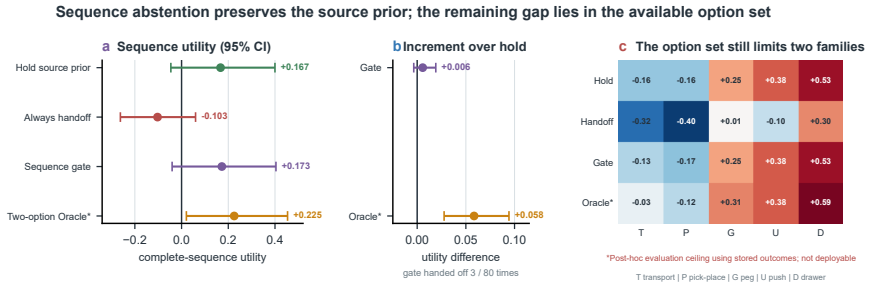

Figure 5: **The complete sequence audit separates conservative abstention from option-set coverage.** The gate significantly exceeds always-handoff and the fold-local prototype but is tied with hold. The post-hoc two-option Oracle improves over hold while remaining negative for transport and pick-place.

Table 5: Condensed chronology. Stopped routes remain evidence and are not silently overwritten by later protocols.

Stage	Question	Decisive observation	Consequence
V10–V11.1	Is source representation sufficient?	Structured kinematic/contact support passes independent cfg3 confirmation	Freeze source interface; activate speed and jerk only
V12–V14	Are matched robot effects reachable?	Legacy smoother and analytic operator lack coverage; 19-candidate basis reaches both axes at 90/90 anchors	Freeze counterfactual basis; model selection separately
V14.1–V16	Can learned ranking/visual/sparse calibration pick jerk candidates?	Speed remains reliable; jerk winner selection is unstable; sparse probes do not beat fixed D-optimal baseline	Stop tuning opened groups; test fixed source prior on fresh groups
V17–V23	Can source information realize local behavior?	Direct response and matched fixed prior are positive, source-specific, and task admissible on new groups	Establish local realization; keep claims local
V24–V25	Does a later learned decision improve the sequence?	Handoff removes fixed-prior benefit; conservative gate is tied with hold; Oracle exposes restricted options	Stop sequence-controller development and report the boundary

- the manuscript result macros, a verifier that checks all 54 reported quantities against those files, an environment manifest, and the exact figure generator with its anonymized visual assets.

RH20T remains subject to its original access terms; the package contains episode identifiers and hashes rather than redistribute restricted raw data. Simulator-generated trajectories and analysis tables can be released directly.

Reference compute and software. The reference workstation uses Windows 11, an NVIDIA RTX 5090 with 32 GB VRAM, an AMD Ryzen 9 9950X3D, 64 GB system memory, and CUDA 12.8. The principal software stack is Python 3.10/3.12, PyTorch 2.7.1+cu128, Transformers 4.53.0, LeRobot 0.3.3, ManiSkill 3.0.0b21, Gymnasium 1.2.0, NumPy 2.2.6, SciPy 1.15.3, scikit-learn 1.7.0, and pandas 2.3.1. The frozen SmolVLA checkpoint hash begins with `115783265d07951e`; complete model, manifest, and result hashes are distributed in the machine-readable protocol files.