arXiv:2611.10825v1 [cs.RO] 7 Nov 2026

MoE-DAgger: Transition-Aware Router-Entropy Gating for Selective Human Correction in Phase-Structured Manipulation

Sungjun Heo[1], Sungjun Yang[2], Heedon Jeong[3], Junhyoung Ha[2,†] and Jeong hwan Jeon[1,3,†]

***Abstract*— Selective data acquisition in Dataset Aggregation (DAgger) on robots requires deciding which learner rollouts to correct and when to intervene. In manipulation with ordered phases, router entropy from a phase-supervised soft mixture-of-experts (MoE) is insufficient: the same value can reflect either a normal adjacent-phase handoff or support for nonadjacent or earlier phases. We propose MoE-DAgger, which uses causal router history to distinguish these cases without providing the policy with a phase label, progress variable, or hard phase command during execution. The acquisition rule allows an adjacent transition to settle, but requests correction when high entropy persists outside such a transition, the transition times out, or routing persistently favors an earlier phase. Under matched active human-control budgets, MoE-DAgger achieves 84.0% task success on Green and 78.0% on Red, exceeding phase-unaware filtered entropy by 22 percentage points on Green and 26 points on Red, with no observed collisions in 50 rollouts per task. On 50 shared intervention-free traces per task, 83.3% and 80.0% of its first proposals fall on failed rollouts within predefined failure windows when 15 and 25 traces are selected, compared with 70.0% and 72.0% for temporal action-chunk disagreement and 63.3% and 60.0% for ensemble disagreement. Shadow replay and an active no-veto ablation indicate that suppressing queries at normal adjacent handoffs is a key contributor to the improved allocation.**

Project page: https://SJ824.github.io/MoE-DAgger/

I. Introduction

Visuomotor behavior cloning can reproduce long-horizon manipulation from demonstrations [1], [2], but compounding errors can move a deployed policy beyond the demonstrated state distribution. DAgger addresses this shift by adding expert supervision on learner-visited states [3]; with intermittent human takeover, however, monitoring and corrective control are costly in real-world interactive systems [4]–[7]. Under a fixed budget of active human control, acquisition must choose both which rollout to correct and when to take over. Rollout-level selection alone is insufficient: a proposal after collision onset is too late, whereas one at a benign handoff prevents autonomy from reaching later interaction states. We therefore evaluate selective acquisition by rollout selectivity, failure-mode-specific critical-window alignment, and burden on successful rollouts. Our demonstrations follow a known ordered semantic phase graph and supervise phase-anchored ACT-style action heads within a soft MoE [8], [9]. Each head predicts a complete Action Chunking with Transformers (ACT) [1] chunk rather than a hand-authored skill or finite-state mode. During execution, routing remains soft and observation-driven; phase topology is used only by the acquisition rule. Consequently, the same scalar entropy can describe either a normal current-to-successor transition or support for nonadjacent or earlier phases (Figure 1).

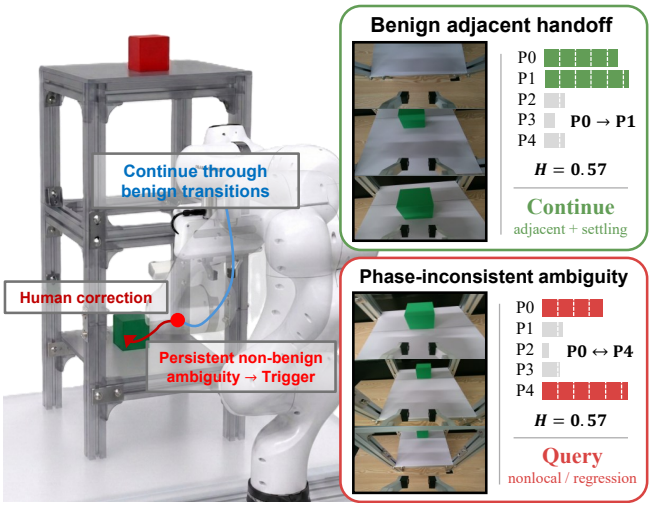


Fig. 1: **Equal router entropy can require different acquisition decisions.** *Green:* The robot remains autonomous while router support shifts from $P0$ to $P1$ and settles within the allowed transition time. *Red:* The same entropy value can arise when the router supports the nonadjacent phases $P0$ and $P4$. The acquisition rule does not issue a query during the settling adjacent transition, but requests human correction if the nonadjacent support persists.

MoE-DAgger uses only causal router history to track legal current-to-successor support and temporal settling. It suppresses queries at bounded compatible handoffs and triggers on persistent non-benign ambiguity, transition timeout, or regression. The phase graph is therefore an acquisition prior, not a runtime action-routing constraint. We evaluate (i) downstream utility from independently acquired correction sets under matched active human-control time and (ii) count-matched first-proposal allocation on the same 50 Green and 50 Red intervention-free traces. The latter credits a failed rollout only when its first proposal lies in the corresponding fixed failure-mode-specific critical window. Phase-unaware filtered-H is the primary same-signal mechanism ablation, while ACT-Temporal-Disagreement and Ensemble-Disagreement are the two principal phase-unaware action-space baselines. Across the two tasks, MoE-DAgger im-

This work was supported by the InnoCORE program of the Ministry of Science and ICT (#N10260008 and UNIST 1.260037.01), by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2022-NR070854), and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST); No. RS-2026-25527734, Development of an industry-specified multimodal hyperscale foundation model).

† Equal Advising, Corresponding authors: {j.ha, jhjeon}@unist.ac.kr

[1]Artificial Intelligence Graduate School, [2]Department of Mechanical Engineering, and [3]Department of Electrical Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea.

proves task success over the same-signal gate by 22/26 percentage points and achieves the highest aligned yield in all four task–budget blocks. Shadow replay and an active no-veto test isolate the adjacent-handoff mechanism.

Contributions. We (1) show that scalar router entropy does not determine transition structure in an unclamped phase-supervised soft MoE; (2) introduce a causal history-only gate that conditions persistent entropy on legal support and settling while retaining timeout and regression triggers; and (3) evaluate downstream utility, critical-window-aligned allocation, and the adjacent-veto mechanism under matched human-control time.

II. Related Work

A. Interactive Imitation and Selective Acquisition

DAgger aggregates expert supervision on learner-induced states [3], [10]. SafeDAgger [11], HG-DAgger [4], EnsembleDAgger [12], and ThriftyDAgger [5] reduce supervision using learned safety, human takeover, disagreement, novelty, or risk, while related systems learn from interventions or mixed corrective/evaluative feedback [6], [7]. Recent work gates on diffusion uncertainty or learned intervention criteria, or changes correction granularity through skill-level queries, residual policies, and force-aware editing [13]–[16]. AutoIntervene compares proposed chunks with a phase-local visual–action memory and calibrates bidirectional policy–operator transfer [17]. Our gate instead uses soft-MoE router history and a known phase graph to distinguish nominal adjacent handoffs from correction-worthy ambiguity.

B. Action-Sequence Policies and Uncertainty

ACT and diffusion policies model long-horizon, multimodal visuomotor behavior [1], [2]. Deep ensembles estimate predictive disagreement [18], and Diff-DAgger derives uncertainty from a deployed diffusion policy [19]. We compare phase-unaware temporal disagreement across overlapping ACT chunks and same-family ensemble disagreement.

C. Phase-Structured Policies and Expert Routing

MoE models learn specialized computation through sparse or soft routing [8], [20]. State-gated experts combine supervision sources in contact-rich manipulation [21], while phase-supervised MoEs use semantic phases to specialize ACT-style action prediction [9]. Maneuver-conditioned hierarchical control [22] and stage-conditioned imitation [23] provide the policy with a runtime stage or mode command; our acquisition gate uses phase topology only to interpret router ambiguity.

III. Method

A. Problem Setup and Phase-Supervised Soft MoE

For task $d \in \mathcal{T}$, we train a separate policy π_θ^d and annotate demonstrations with the ordered semantic graph $0 \to 1 \to \cdots \to S-1$. At runtime, the phase graph is used only by the acquisition gate, not to command the policy: no phase/progress variable or hard mode command is provided to the policy, and the router remains free to blend or revisit heads. Starting from $\mathcal{D}_0^d$, round i executes the learner, proposes takeover from policy-output history, records an accepted suffix $\mathcal{C}_i^d$, and retrains on $\mathcal{D}_{i+1}^d = \mathcal{D}_i^d \uplus \mathcal{C}_i^d$. Proposal timing is part of acquisition because an accepted takeover replaces the remainder of the autonomous rollout with human control, changing which later learner-induced states are visited and corrected. We therefore evaluate proposal timing against synchronized physical task semantics rather than treating the router's head index as ground-truth phase. At policy step t, $\mathbf{o}_t = (\mathbf{I}_{t-L:t}^{1:V}, \mathbf{q}_t, \dot{\mathbf{q}}_t)$ contains synchronized external/wrist views, task-relevant crops, and proprioception. In both tasks, the $S = 5$ phase anchors are approach, entry/alignment, push, withdrawal/clearance, and home return. A shared encoder produces $\mathbf{h}_t$; each of the $K = S$ phase-anchored heads predicts a complete ACT chunk $\widehat{\mathbf{A}}_t^{(k)} \in \mathbb{R}^{C \times n_a}$, and

$$p_{t,k} = \frac{\exp(r_k(\mathbf{h}_t)/T_r)}{\sum_{j=0}^{K-1} \exp(r_j(\mathbf{h}_t)/T_r)}, \quad k \in \{0, \dots, K-1\},$$
$$\widehat{\mathbf{A}}_t = \sum_{k=0}^{K-1} p_{t,k} \widehat{\mathbf{A}}_t^{(k)}. \tag{1}$$

We temporally ensemble overlapping chunks for execution.

Each supervised frame has a manually reviewed phase label z_t; correction suffixes label only the phases they contain. With label-smoothed target $\widetilde{\mathbf{y}}_t$, training uses

$$\mathcal{L} = \mathcal{L}_{\mathrm{BC}} + \lambda_r \, \mathrm{CE}(\widetilde{\mathbf{y}}_t, \mathbf{p}_t) + \lambda_e \sum_k \widetilde{y}_{t,k} \mathcal{L}_{\mathrm{BC}}^{(k)} + \lambda_b \mathcal{L}_{\mathrm{balance}} + \lambda_s \mathcal{L}_{\mathrm{chunk\text{-}smooth}}. \tag{2}$$

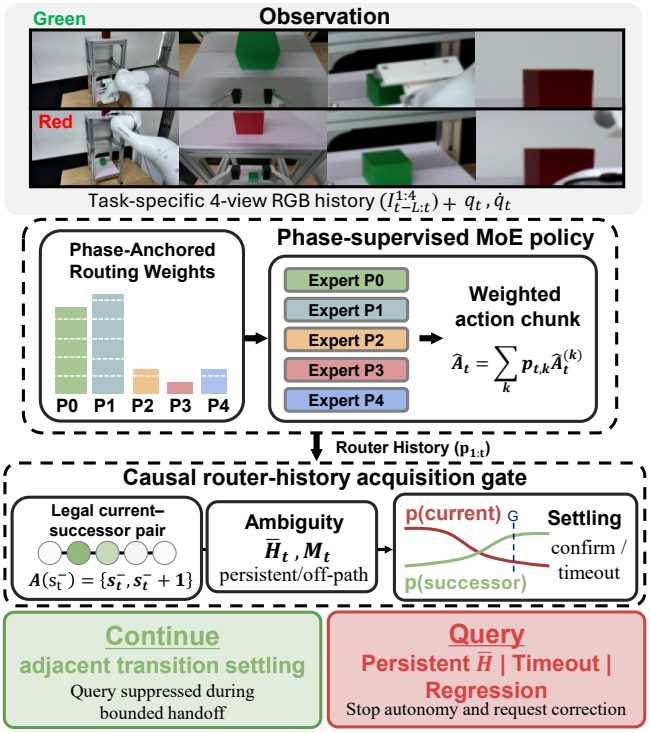


Fig. 2: **Online MoE-DAgger acquisition architecture.** Separate policies for Green and Red map synchronized four-view RGB history and proprioception to a soft mixture of outputs from five phase-anchored action heads. The acquisition module receives only causal router history. It permits a bounded, settling transition between the current phase and its legal successor, but requests human correction when high entropy persists outside that case, when the transition times out, or when routing repeatedly moves backward.

Because $\mathbf{p}_t$ is not projected onto the phase graph, the router remains re-entrant and backward routing remains observable to the acquisition gate.

B. Router Ambiguity and Legal Phase Support

We compute normalized Shannon entropy and its causal W_H-step mean,

$$H_t = -\frac{1}{\log K}\sum_{k=0}^{K-1} p_{t,k}\log p_{t,k}, \quad \bar{H}_t = \frac{1}{W_H}\sum_{\ell=0}^{W_H-1} H_{t-\ell}, \quad (3)$$

with clipping only for numerical evaluation of $\log p_{t,k}$. Entropy is an ambiguity statistic, not a calibrated failure probability, reward, or safety certificate. Let s_t^- be the stabilized phase frozen before the tracker is updated. Legal adjacent support and off-path mass are

$$\mathcal{A}(s_t^-) = \begin{cases} \{s_t^-, s_t^- + 1\}, & s_t^- < S-1, \\ \{S-1\}, & s_t^- = S-1, \end{cases} \quad M_t = 1 - \sum_{k\in\mathcal{A}(s_t^-)} p_{t,k}. \quad (4)$$

Proposition 1 (entropy does not determine transition structure). A router supported on a legal pair satisfies $H_t \le \log 2/\log K$. More generally, if the legal pair is top-2 and $0 \le M_t = m \le \tau_M \le (K-2)/K$, maximizing entropy within the legal and off-path groups gives

$$H_{\max}^{\text{veto}}(m) = \frac{h_2(m) + (1-m)\log 2 + m\log(K-2)}{\log K}, \quad (5)$$

where $h_2(m) = -m\log m - (1-m)\log(1-m)$. Hence entropy combines legal current–successor sharing, uncertainty about being on path, and dispersion among off-path heads. For $K = 5$, $H_{\max}^{\text{veto}}(0) = 0.431$ and $H_{\max}^{\text{veto}}(0.15) = 0.731$; thus a locally compatible handoff can exceed both deployed thresholds once allowed off-path mass is present. Moreover, $\mathbf{p}_{\text{legal}} = \left(.475, .475, \frac{1}{60}, \frac{1}{60}, \frac{1}{60}\right)$ and $\mathbf{p}_{\text{nonlocal}} = \left(.475, \frac{1}{60}, \frac{1}{60}, \frac{1}{60}, .475\right)$ have identical entropy 0.567 for $s_t^- = 0$ but incompatible support. The phase graph therefore affects only acquisition logic: it determines whether entropy-based queries can be temporarily suppressed and defines the legal transitions and phase order used by timeout and regression tests, but never constrains policy routing.

C. Transition-Aware Query Rule

For stabilized phase $s = s_t^- < S-1$, a current–successor candidate is locally compatible when

$$L_t(s) = \mathbb{I}[\text{Top2}(\mathbf{p}_t) = \{s, s+1\} \wedge M_t \le \tau_M]. \quad (6)$$

An inactive candidate opens when $L_t(s) = 1$ and $p_{t,s+1} \ge \gamma_{\text{start}}$. It confirms when $\arg\max_k p_{t,k} = s+1$, $p_{t,s+1} \ge \gamma_{\text{conf}}$, and $p_{t,s} \le \gamma_{\text{prev}}$; it cancels when the argmax is s and $p_{t,s+1} \le \gamma_{\text{cancel}}$. Confirmation and cancellation require P_{conf} and P_{cancel} consecutive frames, respectively; a cancelled candidate may reopen. Let a_t and g_t be candidate activity and age during gate evaluation. The bounded benign-handoff veto is

$$B_t = \mathbb{I}\left[a_t = 1 \wedge L_t(s_t^-) = 1 \wedge g_t \le G\right]. \quad (7)$$

TABLE I: Active acquisition parameters. G/R denotes Green/Red; all step counts are at 20 Hz. Deployed native-score thresholds are task-specific, frozen at 10/20 trajectory-level proposal incidence on disjoint development sets, and not cross-method comparable. Phase-unaware filtered-H reuses MoE-DAgger entropy-side settings; exact-B audit cutpoints are separate and audit-only.

Quantity	Value	Quantity	Value
Views / experts	4 / 5	Chunk / rate	30 steps / 20 Hz
τ_H (G/R), W_H	0.50/0.46, 5	P_H, $t_{\min}^H$	5/5 steps
τ_M, γ_{start}	0.15, 0.10	$\gamma_{\text{conf}}/\gamma_{\text{prev}}$	0.60, 0.35
$P_{\text{conf}}/\gamma_{\text{cancel}}$	3 steps, 0.05	P_{cancel}/G (G/R)	4 steps, 30/30 steps
τ_{ACT} (G/R)	0.78/0.47	τ_{Ens} (G/R)	1.11/0.91
P_R	3 steps	jump / action query	disabled / disabled

Nonlocal support revokes the veto, whereas a locally compatible candidate that fails to settle remains active beyond G. Entropy and regression persistence are

$$c_t^H = \begin{cases} c_{t-1}^H + 1, & t \ge t_{\min}^H \wedge \bar{H}_t > \tau_H \wedge \neg B_t, \\ 0, & \text{otherwise}, \end{cases}$$
$$c_t^R = \begin{cases} c_{t-1}^R + 1, & \arg\max_k p_{t,k} < s_t^-, \\ 0, & \text{otherwise}. \end{cases} \quad (8)$$

The deployed gate requests correction when $\Gamma_t = 1$, where

$$\Gamma_t = \mathbb{I}[c_t^H \ge P_H] \vee \mathbb{I}[a_t = 1 \wedge g_t > G] \vee \mathbb{I}[c_t^R \ge P_R]. \quad (9)$$

The branches represent persistent non-benign entropy, unresolved-transition timeout, and persistent regression. Off-path mass is not a standalone query branch; through L_t, it governs local transition-candidate and veto eligibility. Support that violates the local-compatibility test revokes the veto, while explicit jump and action-disagreement query branches are disabled. All tests use the pre-update s_t^-, and state updates occur after evaluating Γ_t; confirmation advances the stabilized phase for $t+1$, while confirmation, cancellation, or nonlocal routing closes the candidate.

D. Human Correction, Retraining, and Termination

When $\Gamma_t = 1$, autonomy stops and the robot holds its measured state. Correction uses the initial-data record-and-replay pipeline, base-frame Cartesian-delta convention, schema, and 20 Hz sampling; the full accepted suffix is retained without fixed trimming. Rejected attempts and retries are excluded from training and the matched control-time budget but remain logged. Policies retrain on $\mathcal{D}_i^d \uplus \mathcal{C}_i^d$, never corrections alone, under frozen initialization/warm-start, optimizer updates, sampling, normalization, validation, and checkpoint-selection manifests. The home detector is armed after the end effector leaves a 0.17 m neighborhood of the start and terminates evaluation at $\|\mathbf{x}_t^{\text{EE}} - \mathbf{x}_{\text{home}}^{\text{EE}}\|_2 \le 0.15\,m$. Controller faults, safety/operator stops, and acquisition queries remain distinct termination causes.

IV. Experiments

A. Tasks, Data, and Scope

As illustrated in Figure 3, a Franka Research 3 performs two phase-structured pushing tasks on a rigid two-level fixture. Green uses the lower shelf; Red follows the same approach–enter–push–withdraw–return sequence on the

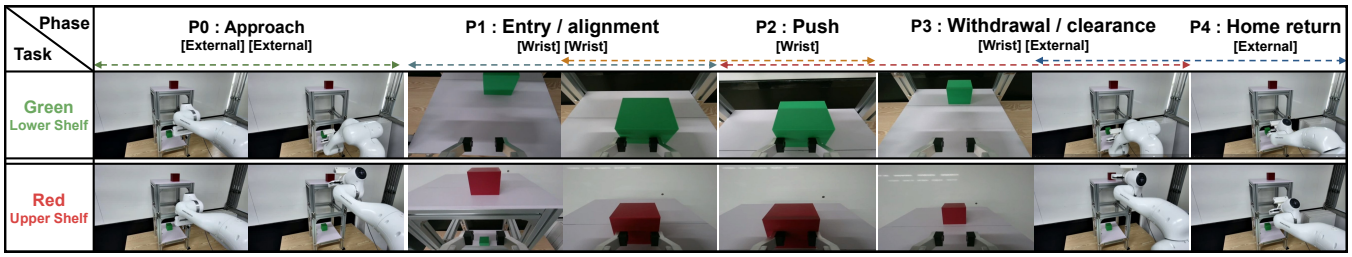


Fig. 3: **Ordered phase structure of the two real-robot pushing tasks.** *Top:* Green operates on the lower shelf; *bottom:* Red operates on the upper shelf. Both tasks follow the semantic sequence P0 approach, P1 entry/alignment, P2 push, P3 withdrawal/clearance, and P4 home return. Representative external and wrist-camera keyframes are shown in execution order. Adjacent phases share boundary keyframes: the last keyframe of one phase is also used as the onset keyframe of the next phase (P1→P2, P2→P3, and P3→P4). The panels are illustrative and not time-scaled.

upper shelf under different geometry, visual context, and contact conditions. In both tasks, $P0$–$P4$ denote approach, entry/alignment, push, withdrawal/clearance, and home return. Deployed operating points are calibrated on a separate development-only set of 20 full-length initial-policy trajectories per task. The 50 Green/50 Red intervention-free initial-policy archive is disjoint from this calibration set and is used for the descriptive audit in Section IV-C; it is also disjoint from downstream policy evaluation. The initial Green/Red datasets contain 30 demonstrations each (9,956/9,410 frames; 968.3 s total at 20 Hz). The retained MoE-DAgger correction sets contain 2,251/2,551 frames, corresponding to 112.55/127.55 s of active human control.

B. Matched-Budget Methods and Physical Evaluation

Budgets are anchored to the realized MoE-DAgger Green/Red active-control times. Whole suffixes are retained, so the last may cross the target; the stopping rule and $\pm 5\%$ tolerance are fixed before collection. Every interactive method starts from the same task-specific initial BC checkpoint and independently acquires learner-state data.

a) Compared conditions: **BC-only** adds no data; **ExtraDemo-BC** adds full-start demonstrations matched in active-control time. **Human-trigger DAgger** uses score-blind operator takeover when continued autonomy threatens completion, safety, or interaction quality. **Phase-matched Random** samples a frozen phase/progress bin before each rollout and triggers from the online stabilized router phase without future labels. **ACT-Temporal-Disagreement** scores overlapping chunks predicting the same next motion; **Ensemble-Disagreement** uses $N_E = 3$ independently trained same-family policies, one acting and two scoring. **Phase-unaware filtered-H** is the direct same-signal ablation: it reuses MoE-DAgger entropy thresholding, causal averaging, persistence, and minimum-query settings but removes the transition tracker, adjacent veto, timeout, and regression without separate incidence recalibration. **MoE-DAgger (ours)** uses Eq. (9).

For action-space gates, $\widetilde{\mathbf{a}}_t^{(j)}$ is predictor j's predicted next-step six-dimensional motion, normalized by frozen initial-BC statistics; the gripper is excluded. With $\mathcal{P}_t$ the overlapping ACT predictions and $N_E = 3$,

$$D_t^{\mathrm{ACT}} = \frac{1}{|\mathcal{P}_t|} \sum_{j \in \mathcal{P}_t} \|\widetilde{\mathbf{a}}_t^{(j)} - \bar{\mathbf{a}}_t\|_2^2, D_t^{\mathrm{ens}} = \frac{2 \sum_{e<e'} \|\widetilde{\mathbf{a}}_t^{(e)} - \widetilde{\mathbf{a}}_t^{(e')}\|_2^2}{N_E(N_E-1)}. \quad (10)$$

b) Development-only threshold calibration: For each task, we use a separate development-only set of 20 full-length initial-policy trajectories, disjoint from the later 50-trace common-trace archive. Holding each rule's temporal processing and eligibility logic fixed, we sweep the native-score threshold for MoE-DAgger, ACT-Temporal-Disagreement, and Ensemble-Disagreement and select the value whose causal replay produces at least one proposal on 10 of the 20 trajectories, or the closest attainable count. No success/failure balancing or outcome-based subsampling is performed. This calibration equalizes trajectory-level proposal incidence; it does not optimize failure precision, critical-window alignment, or retrained-policy performance. Phase-unaware filtered-H is not separately recalibrated: it reuses the MoE-DAgger entropy-side settings while removing the transition-aware components. All thresholds are frozen before added-data acquisition and are not retuned using physical-evaluation results.

c) Physical evaluation and reproducibility: For each method and task, two independently seeded checkpoints are evaluated in 25 rollouts each, yielding $n = 50$ per method–task cell and 800 rollouts overall. The 7 cm-side cube is sampled 0–5 cm short of the stop line, within ± 4 cm laterally and $\pm 15^\circ$ in yaw; rollouts are randomized and interleaved, and poses are logged. All rollouts share the runtime, controller, sensing/synchronization, temporal ensembling, safety, home-detector, and episode-cap settings; all added-data policies share architecture, retraining, sampling, normalization, validation, and checkpoint selection. Corrections use one logged takeover–hold–control–acceptance–reset protocol. Active human-control time excludes offline annotation and uninstrumented attendance, and only predeclared instrumentation failures are replaced. Each method–task condition uses one acquired dataset, so the two seeds characterize variation within, not across, acquisition sets.

C. Common-Trace Acquisition Audit

Because intervention changes future states, ACT, Ensemble-Disagreement, phase-unaware entropy, and MoE-DAgger are causally replayed on the same 50 Green and 50 Red intervention-free initial-policy traces. For each rule m and task d, sweeping the native-score cutpoint assigns one frozen, outcome-blind rank $r_{m,d}(i)$ to every archived rollout. Thus, $r_{m,d}$ is a method-specific permutation of the same archive rather than a separate trace set. At selected-

rollout budget $B \in \{1, \ldots, 50\}$, $c_{m,d}^{(B)}$ retains exactly the first B rollouts in that ordering, and $q_{m,i}(c)$ is the earliest proposal time under the rule's complete score processing, persistence, and eligibility logic, not the raw-score peak time. The same ordering is used in all panels of Figure 4; Table III reports $B = 15$ and 25, where $B = 15$ matches the 15 accepted MoE-DAgger correction suffixes collected per task. Outcomes and critical windows are joined only after rankings and first-proposal times are frozen. Unlike deployed thresholds calibrated on the disjoint 20-trajectory set, audit-only cutpoints use the held-out 50-trace archive solely for exact-B count matching; they are descriptive and not cross-method comparable. The audit horizon ends before the first offline $P3 \rightarrow P4$ onset.

The archive contains 29/36 mutually exclusive Green/Red failures: collision (3/4), no contact (11/12), and partial push (15/20). Collision is credited only when the first proposal occurs no later than onset and within the onset or preceding phase; no-contact and partial-push windows are $P1$–$P2$ and $P2$–$P3$, respectively. We report aligned failed/proposed yield, successful-rollout proposals, and per-mode coverage from the first retained proposal; retrospective timing credit does not establish recoverability. In Figure 4, panels (a,b) count all aligned failures and panels (e,f) restrict the same count-matched sweep to collision and no contact.

D. Adjacent-Veto Mechanism Tests

a) Shadow component replay: The common traces are replayed with raw entropy, averaging plus persistence, the complete rule without the adjacent veto, and the complete rule under the fixed shadow configuration. All rising edges enter edge and handoff statistics, whereas alignment uses only the first retained edge; proposal counts are unmatched and timeout/regression remain bundled. The complete-rule mask is reused in Table IV and panels (c,d) of Figure 4. Because no correction or retraining occurs, this replay diagnoses allocation rather than downstream utility.

b) Active no-veto test: This prospective ablation removes only Eq. (7); averaging, persistence, timeout, and regression remain active. Full and no-veto gates start from the same initial BC checkpoint, independently collect under matched control time, and use identical retraining and frozen evaluation. Unlike phase-unaware filtered-H, it preserves the transition model; unlike shadow replay, it includes collection and retraining. The full row reuses Table II, and no-veto adds 100 rollouts.

V. Metrics and Statistical Analysis

A. Downstream and Budget Metrics

We distinguish collision-free execution of the prescribed physical sequence from geometric task completion:

$$\begin{aligned} E_{\text{seq}} &= I_{\text{safe}} I_{\text{complete}} I_{\text{time}}, \\ S_{\text{task}} &= E_{\text{seq}} I_{\text{line}}, \qquad I_{\text{line}} = \mathbb{I}[m_{\text{push}} \le 0]. \end{aligned} \tag{11}$$

Here, $I_{\text{safe}} = 1$ when the rollout completes without collision or a collision-related protective stop, $I_{\text{complete}} = 1$ when the prescribed physical sequence $P0 \rightarrow P1 \rightarrow P2 \rightarrow P3 \rightarrow P4$ is completed, including withdrawal and return home, and $I_{\text{time}} = 1$ when completion occurs before the episode cap. These indicators are determined from the synchronized physical rollout outcome, not from router-head transitions. Thus E_{seq} does not require stop-line crossing, whereas S_{task} additionally requires the cube to cross the line. With unit push direction $\mathbf{d}_{\text{push}}$, $m_{\text{push}} = \mathbf{d}_{\text{push}}^{\top}(\mathbf{p}_{\text{goal}} - \mathbf{p}_{\text{cube,rear}}(T))$. Negative values indicate stop-line crossing and positive values the remaining shortfall; only the sign enters S_{task}, while the magnitude is descriptive. Active human-control time is the matched supervision budget, and all non-instrumentation failures remain policy outcomes.

B. Failure-Mode Alignment

Using the frozen ordering and audit cutpoints in Section IV-C, let $q_{m,i}^{(B)} = q_{m,i}(c_{m,d}^{(B)})$ be method m's earliest retained proposal on rollout i, with $q_{m,i}^{(B)} = \varnothing$ when none is retained. Let $y_i = 1$ denote eventual failure, f_i its frozen mode, and $\mathcal{W}_{d,f_i}(i)$ the corresponding critical window. With $N_{d,f}$ archived failures of mode f, we report

$$\begin{aligned} a_{m,i}^{(B)} &= \mathbb{I}\left[y_i = 1 \wedge q_{m,i}^{(B)} \in \mathcal{W}_{d,f_i}(i)\right], \\ Y_{m,d}^{(B)} &= \frac{1}{B} \sum_i a_{m,i}^{(B)}, \\ N_{m,d}^{\text{succ}}(B) &= \sum_i \mathbb{I}\left[y_i = 0 \wedge q_{m,i}^{(B)} \neq \varnothing\right], \\ C_{m,d,f}^{(B)} &= \frac{1}{N_{d,f}} \sum_{i:f_i=f} \mathbb{I}\left[q_{m,i}^{(B)} \in \mathcal{W}_{d,f}(i)\right]. \end{aligned} \tag{12}$$

$Y_{m,d}^{(B)}$ is the critical-window-aligned failed/proposed yield, $N_{m,d}^{\text{succ}}(B)$ counts successful traces among the exactly B proposal-bearing rollouts, and $C_{m,d,f}^{(B)}$ is per-mode coverage. Consequently, the aligned and successful counts need not sum to B. All alignment metrics use only the first retained proposal; later proposals do not repair a mistimed first edge. Higher Y and C, and lower N^{succ}, indicate better count-matched allocation on this fixed archive but do not measure intervention recovery. Panels (c,d) of Figure 4 apply a fixed-shadow proposal mask to the same ordering $r_{m,d}(i)$. Let $\widetilde{q}_{m,i}$ be the first retained proposal under the fixed shadow configuration, with $\widetilde{q}_{m,i} = \varnothing$ when no proposal is produced. We plot

$$\widetilde{N}_{m,d}^{\text{succ}}(B) = \sum_i \mathbb{I}[y_i = 0 \wedge r_{m,d}(i) \le B \wedge \widetilde{q}_{m,i} \neq \varnothing]. \tag{13}$$

Unlike $N_{m,d}^{\text{succ}}(B)$, this quantity is not count-matched: B denotes the common rank-prefix size, while the number of fixed-shadow proposals within that prefix may be smaller than B. At $B = 50$, it equals the total number of successful traces that actually receive a proposal under the fixed shadow rule. The complete-rule row of Table IV uses the same $\widetilde{q}_{m,i}$. There, "Prop." is the number of rollouts with any retained rising edge, and "fail/prop." credits a failed rollout only when its first retained edge lies in $\mathcal{W}_{d,f_i}(i)$. Handoff overlap uses all retained edges and is a transition-confounding diagnostic rather than a calibrated false-positive rate.

TABLE II: **Frozen downstream performance under runtime-identical, matched-supervision evaluation.** Each method uses two checkpoints from one fixed acquisition set and 25 rollouts per checkpoint and task ($n = 50$). Cells show point estimates and 95% CIs: Wilson for binary outcomes and pose-stratified rollout bootstrap for signed margin. Sequence execution does not require stop-line crossing; task success does. Margin is the all-rollout mean ($m_{\text{push}} < 0$ indicates crossing; more negative is deeper). Bold/underlining mark best/second-best binary estimates; ties are retained.

	Human-control (s)		Seq. execution (%)		Task success (%)		Signed margin (cm)		Collision (%)	
Method	G	R	G	R	G	R	G	R	G	R
BC-only	0	0	88.0 [76.2,94.4]	82.0 [69.2,90.2]	52.0 [38.5,65.2]	48.0 [34.8,61.5]	1.22 [0.981, 1.468]	1.29 [1.022, 1.612]	12.0 [5.6,23.8]	18.0 [9.8,30.8]
ExtraDemo-BC	114.85	128.50	90.0 [78.6,95.7]	82.0 [69.2,90.2]	58.0 [44.2,70.6]	46.0 [33.0,59.6]	0.97 [0.664, 1.397]	1.27 [1.017, 1.599]	10.0 [4.3,21.4]	18.0 [9.8,30.8]
Human-trigger DAgger	114.45	127.95	98.0 [89.5,99.6]	92.0 [81.2,96.8]	72.0 [58.3,82.5]	68.0 [54.2,79.2]	−0.18 [−0.398, 0.122]	0.21 [−0.009, 0.397]	**0.0** [0.0,7.1]	4.0 [1.1,13.5]
Phase-matched Random	116.20	126.45	92.0 [81.2,96.8]	90.0 [78.6,95.7]	62.0 [48.2,74.1]	60.0 [46.2,72.4]	−0.02 [−0.132, 0.105]	0.63 [0.317, 0.821]	8.0 [3.2,18.8]	10.0 [4.3,21.4]
Phase-unaware filtered-H	114.85	128.35	86.0 [73.8,93.0]	84.0 [71.5,91.7]	62.0 [48.2,74.1]	52.0 [38.5,65.2]	1.05 [0.814, 1.221]	1.54 [1.129, 1.910]	14.0 [7.0,26.2]	16.0 [8.3,28.5]
ACT-Temporal-Disagreement	115.25	128.45	**100.0** [92.9,100.0]	96.0 [86.5,98.9]	82.0 [69.2,90.2]	72.0 [58.3,82.5]	−1.67 [−2.173, −0.991]	−0.81 [−1.215, −0.542]	**0.0** [0.0,7.1]	**0.0** [0.0,7.1]
Ensemble-Disagreement	115.85	127.65	98.0 [89.5,99.6]	96.0 [86.5,98.9]	**84.0** [71.5,91.7]	68.0 [54.2,79.2]	−1.80 [−2.190, −1.586]	−0.41 [−0.928, −0.007]	**0.0** [0.0,7.1]	2.0 [0.4,10.5]
MoE-DAgger	112.55	127.55	**100.0** [92.9,100.0]	**98.0** [89.5,99.6]	**84.0** [71.5,91.7]	**78.0** [64.8,87.2]	−1.98 [−2.280, −1.521]	−1.13 [−1.417, −0.985]	**0.0** [0.0,7.1]	**0.0** [0.0,7.1]

TABLE III: **Common-trace failure-aligned allocation** on the same 50 Green/50 Red traces; outcome-blind audit cutpoints retain exactly $B = 15$ or 25 rollouts per method and task (Section IV-C). "Align.": first proposals inside mode-specific critical windows on failed rollouts; "Succ.": first proposals on successful traces (21 Green/14 Red); other columns: hits/all archived failures per mode. Bold/underline: best/second-best; lower is better only for "Succ."

	Allocation		Mode coverage		
Rule	Align.	Succ.	Coll.	No-cont.	Partial
Green, B=15, (30% selected fraction)					
ACT temporal disagr.	11/15 (73.3%)	2/21 (9.5%)	**2/3**	6/11	**3/15**
Ensemble disagr.	10/15 (66.7%)	3/21 (14.3%)	**2/3**	5/11	**3/15**
Phase-unaware H	0/15 (0.0%)	7/21 (33.3%)	0/3	0/11	0/15
MoE-DAgger	**12/15 (80.0%)**	**1/21 (4.8%)**	**2/3**	**7/11**	**3/15**
Red, B=15, (30% selected fraction)					
ACT temporal disagr.	10/15 (66.7%)	1/14 (7.1%)	2/4	5/12	**3/20**
Ensemble disagr.	9/15 (60.0%)	3/14 (21.4%)	2/4	4/12	**3/20**
Phase-unaware H	1/15 (6.7%)	5/14 (35.7%)	1/4	0/12	0/20
MoE-DAgger	**13/15 (86.7%)**	**0/14 (0.0%)**	**3/4**	**7/12**	**3/20**
Green, B=25, (50% selected fraction)					
ACT temporal disagr.	17/25 (68.0%)	3/21 (14.3%)	**3/3**	**10/11**	4/15
Ensemble disagr.	15/25 (60.0%)	5/21 (23.8%)	**3/3**	6/11	**6/15**
Phase-unaware H	0/25 (0.0%)	9/21 (42.9%)	0/3	0/11	0/15
MoE-DAgger	**19/25 (76.0%)**	**2/21 (9.5%)**	**3/3**	**10/11**	**6/15**
Red, B=25, (50% selected fraction)					
ACT temporal disagr.	19/25 (76.0%)	2/14 (14.3%)	3/4	9/12	**7/20**
Ensemble disagr.	15/25 (60.0%)	4/14 (28.6%)	3/4	8/12	4/20
Phase-unaware H	1/25 (4.0%)	8/14 (57.1%)	1/4	0/12	0/20
MoE-DAgger	**21/25 (84.0%)**	**0/14 (0.0%)**	**4/4**	**11/12**	6/20

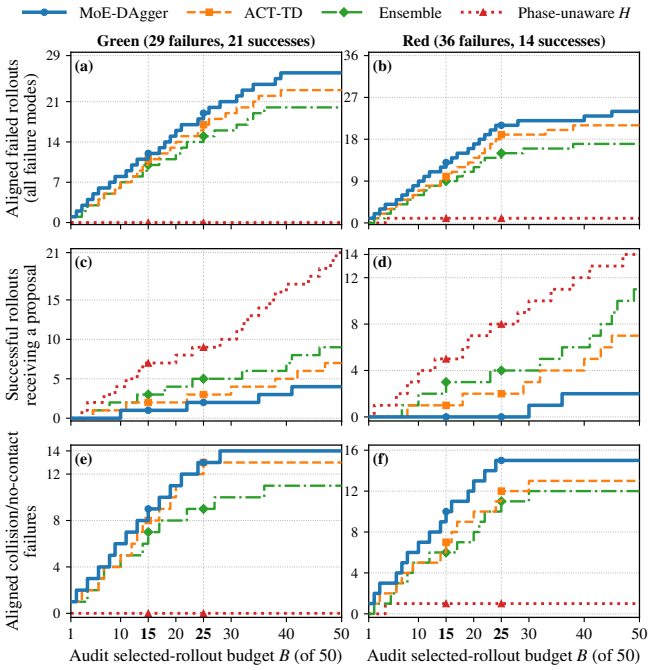


Fig. 4: **Shared-ranking count-matched failure coverage and fixed-shadow successful-rollout burden.** Each method–task pair uses one frozen, outcome-blind ordering in all panels. Panels (a,b) count aligned failures, and panels (e,f) count aligned collision/no-contact failures. Panels (c,d) apply the fixed-shadow mask to the same prefixes and count successful rollouts receiving a proposal, so B is a rank-prefix size rather than the number of shadow proposals. At $B = 50$, they therefore count only successful traces that actually receive a fixed-shadow proposal. Vertical guides mark the $B = 15, 25$ slices in Table III; higher is better in the top/bottom rows and lower in the middle row.

C. Evidence and Statistical Analysis

Rollouts are the observational units; task success is the primary endpoint, with sequence execution and collision complementary and signed margin descriptive. The mechanism-specific comparison is MoE-DAgger versus phase-unaware filtered-H. Per method–task cell ($n = 50$), binary rates use Wilson 95% CIs, pairwise differences unadjusted Newcombe 95% CIs, and margin a rollout-level bootstrap; common-trace values are exact archive counts without CIs.

VI. Results

A. Matched-Budget Downstream Utility

Under nearly identical active human-control time, MoE-DAgger achieves 84.0/78.0% task success, 100.0/98.0% sequence execution, and no observed collisions on Green/Red (Table II). Relative to the direct same-signal phase-unaware gate, this is $+22/+26$ percentage points in task success (95% CI $[4.6, 37.8]/[7.3, 42.4]$), $+14/+14$ in sequence execution ($[4.0, 26.2]/[2.6, 26.6]$), and $-14/-16$ in collision ($[-26.2, -4.0]/[-28.5, -5.5]$). ExtraDemo-BC, Human-trigger DAgger, and Phase-matched Random—controlling for full-trajectory supervision, score-blind human selection, and phase-matched online data—reach 58.0/46.0%, 72.0/68.0%, and 62.0/60.0% task success, respectively. These results indicate that those factors alone do not account for the gain under

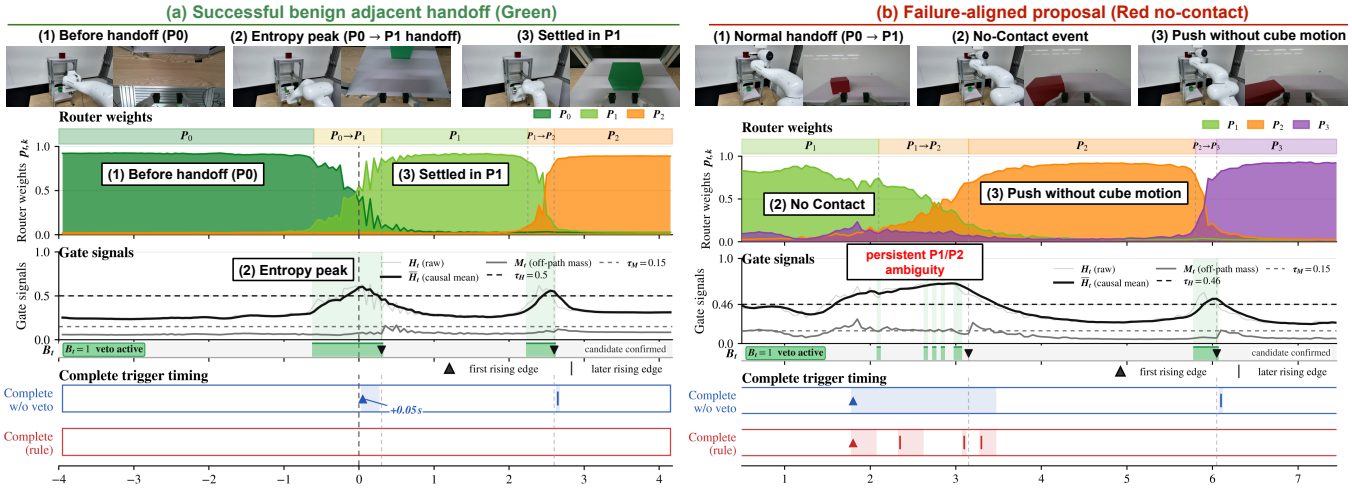


Fig. 5: **Illustrative adjacent-handoff veto behavior under fixed-shadow replay.** (a) On a successful Green trace, no-veto produces its first proposal edge 0.05 s after the offline $P0 \rightarrow P1$ handoff, whereas the full rule suppresses queries during the bounded handoff and produces no retained edge in the shown interval. (b) On a Red no-contact failure, the full rule retains a failure-aligned proposal edge during persistent $P1/P2$ ambiguity. Curves show router weights $p_{t,k}$, raw and causal-mean entropy $H_t, \bar{H}_t$, off-path mass M_t, and veto state B_t; the lower rows show no-veto and full-rule shadow edges. Triangles mark the audit-scored first edge and ticks later edges. Phase bands and frame annotations are offline visualizations, not runtime inputs. Each trace uses its task-specific deployed threshold and the fixed-shadow configuration of Table IV; no audit-only exact-B cutpoint is applied.

TABLE IV: **Shadow gate-component audit.** All variants are replayed on the same 50 Green and 50 Red intervention-free initial-policy traces used for the common-trace audit, using the task-specific deployed entropy threshold frozen on a separate, disjoint 20-trajectory development set for each task. Unlike Figure 4 (panels (a,b,e,f)), no audit-only exact-B cutpoint is imposed; all rising edges within the frozen audit horizon are retained. "Prop." counts rollouts with any retained edge, "aligned fail/prop." is the fraction of proposal-bearing rollouts that eventually fail and whose first retained edge lies in the corresponding mode-specific window defined in Section IV-C, and "handoff ovlp." is the fraction of adjacent handoffs intersected by any retained edge.

Variant	Edges	Prop.	Aligned fail/prop.	Handoff ovlp.
A. Green				
Raw H_t	314	50 / 50 (100.0%)	0 / 50 (0.0%)	147 / 147 (100.0%)
Avg. + persist.	170	50 / 50 (100.0%)	0 / 50 (0.0%)	97 / 147 (66.0%)
Complete w/o veto	170	50 / 50 (100.0%)	0 / 50 (0.0%)	97 / 147 (66.0%)
Complete rule (shadow)	**36**	**32 / 50 (64.0%)**	**24 / 32 (75.0%)**	**0 / 147 (0.0%)**
B. Red				
Raw H_t	250	50 / 50 (100.0%)	0 / 50 (0.0%)	136 / 143 (95.1%)
Avg. + persist.	106	50 / 50 (100.0%)	1 / 50 (2.0%)	71 / 143 (49.7%)
Complete w/o veto	104	50 / 50 (100.0%)	1 / 50 (2.0%)	71 / 143 (49.7%)
Complete rule (shadow)	**34**	**27 / 50 (54.0%)**	**22 / 27 (81.5%)**	**3 / 143 (2.1%)**

matched active-control time. ACT-TD remains competitive at 82.0/72.0% task success with no observed collisions. The task-success differences (MoE-DAgger minus ACT-TD) are $+2/+6$ percentage points, with 95% CIs spanning zero ($[-13.0, 16.9]/[-10.9, 22.5]$).

B. Failure-Mode-Aligned Common-Trace Allocation

Pooled across the two tasks, MoE-DAgger attains critical-window-aligned failed/proposed yields of 83.3/80.0% at $B = 15$ and 25, respectively, compared with 70.0/72.0% for ACT-TD, 63.3/60.0% for the ensemble, and 3.3/2.0% for phase-unaware entropy (Table III). It proposes correction on only 1 and 2 of the 35 successful traces at $B = 15$ and 25, respectively, indicating that transition awareness shifts proposals away from benign successful-rollout handoffs.

Across the full sweep in Figure 4, at $B = 50$ MoE-DAgger aligns 26/29 Green and 24/36 Red failures, including 14/14 and 15/16 collision/no-contact cases. The remaining misses

TABLE V: **Active adjacent-veto ablation.** No-veto removes only Eq. (7); both variants independently collect under matched control time with identical retraining/evaluation. Veto q./rollout denotes the percentage of acquisition rollouts whose query onset meets the counterfactual veto condition. Effects are Full minus No-veto; binary effects are percentage points.

Acquisition gate	Human-control (s)	Veto q. /rollout (%)	Seq. execution (%)	Task success (%)	Collision (%)
No adjacent veto	114.70 / 130.20	92.0 / 50.0	90.0 / 84.0	68.0 / 60.0	10.0 / 16.0
Full MoE-DAgger	112.55 / 127.55	0.0 / 0.0	100.0 / 98.0	84.0 / 78.0	0.0 / 0.0
Δ (Full − No-veto)	$-2.15/-2.65$	$-92.0/-50.0$	$+10.0/+14.0$	$+16.0/+18.0$	$-10.0/-16.0$

are concentrated in partial push, where coverage is 12/15 on Green and 9/20 on Red. Phase-unaware entropy aligns only 1 of 18 and 1 of 33 selected failed rollouts at $B = 15$ and 25, respectively. These counts show that selecting failed rollouts alone does not ensure that first proposals fall within the corresponding critical windows. ACT is stronger on Red partial push at $B = 25$ (7/20 versus 6/20), so the advantage is not uniform across all failure modes.

C. Adjacent-Veto Mechanism Tests

a) Shadow replay: Averaging and persistence reduce raw proposal edges from 314 to 170 on Green and from 250 to 106 on Red, yet without the adjacent veto every trace still receives a proposal, with first edges concentrated near $P0$ or the nominal $P0 \rightarrow P1$ handoff. The complete veto reduces edge counts to 36/34, proposal incidence to 32/27 of 50, and handoff overlap from 66.0/49.7% to 0.0/2.1%, while increasing first-proposal alignment to 24/32 (75.0%) and 22/27 (81.5%) (Table IV). Figure 5 visualizes this contrast: the full rule suppresses a handoff-confounded proposal edge on a successful Green trace but retains a failure-aligned proposal edge during persistent $P1/P2$ ambiguity on a Red no-contact trace. The unmatched aligned counts are lower than the $B = 50$ count-matched counts in Figure 4, even though both analyses use the same traces. The count-matched sweep lowers the audit cutpoint until all 50 archived traces are proposal-bearing, whereas the fixed-shadow complete rule produces proposals on only 32/27 of those same Green/Red traces.

This supports selective suppression of handoff-confounded proposals rather than temporal smoothing alone.

b) Active no-veto test: Relative to no-veto, the full gate reduces the incidence of queries satisfying the veto condition by 92/50 percentage points and improves sequence execution by 10/14 percentage points ($[0.9, 21.4]/[2.6, 26.6]$). The collision-rate differences (full minus no-veto) are $-10/-16$ percentage points ($[-21.4, -0.9]/[-28.5, -5.5]$). Task success improves by 16/18 percentage points, although the intervals for these differences include zero ($[-0.8, 31.8]/[-0.1, 34.6]$); the sequence-execution and collision-rate intervals exclude zero in both tasks.

VII. Discussion

a) What the evidence supports: The active phase-unaware comparison tests downstream utility with fixed entropy-side settings, whereas the common-trace audit tests count-matched first-proposal allocation. Together, consistent with Proposition 1, they show that phase topology changes both rollout selection and proposal timing because equal entropy can represent a settling adjacent handoff or nonlocal/regressive support. The contribution is topology-aware allocation rather than universal dominance over phase-unaware uncertainty.

b) Interpretation and veto tradeoff: The five heads are phase-anchored action predictors rather than independently reusable skills, so the result should be interpreted as an acquisition effect rather than skill discovery or multitask expert reuse. The adjacent veto deliberately trades proposal incidence for reduced interruption at compatible handoffs. Shadow replay isolates this allocation effect, while the active ablation shows its downstream safety and execution consequences under matched control time.

c) Scope and limitations: The study uses one robot and fixture, separate task-specific policies, and two geometric instances of one pushing family; multitask reuse, learned or nonlinear phase graphs, unsupervised skill discovery, other robots, and open-world settings remain untested. Each method–task condition has one acquired dataset, so two training seeds and repeated rollouts do not quantify acquisition-set variability. The matched budget captures active human-control time rather than total operator workload; it excludes offline phase annotation and uninstrumented attendance. Retrospective critical windows do not establish that a credited intervention would recover a rollout, collision coverage contains only 3 Green and 4 Red failures, and partial-push coverage remains limited. One aggregation round establishes neither convergence nor an optimal stopping rule. Extending to larger task- or skill-level MoEs may require hybrid gates: the router may concentrate on one expert and remain low-entropy even when the resulting action predictions exhibit high disagreement, motivating fusion with complementary signals such as ACT temporal disagreement.

VIII. Conclusion

MoE-DAgger conditions persistent soft-MoE router entropy on legal phase support and temporal settling to determine when to request human correction. Under matched active human-control time, it improves task success by 22/26 percentage points over phase-unaware entropy and records no observed collisions in 50 evaluation rollouts per task. Common-trace allocation and adjacent-veto ablations show that this gain is associated with suppressing benign handoff proposals and reallocating correction toward failure-relevant windows. Coverage is strongest for collision and no-contact, while partial-push failures remain the main unresolved mode. These results demonstrate the practical value of known phase topology as an acquisition prior for selective DAgger in structured real-robot manipulation.

References

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[2] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10–11, pp. 1684–1704, 2025.

[3] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[4] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer, "HG-DAgger: Interactive imitation learning with human experts," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[5] R. Hoque, A. Balakrishna, E. Novoseller, A. Wilcox, D. S. Brown, and K. Goldberg, "ThriftyDAgger: Budget-aware novelty and risk gating for interactive imitation learning," in *Conference on Robot Learning (CoRL)*, 2022.

[6] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa, "Learning from interventions: Human–robot interaction as both explicit and implicit feedback," in *Robotics: Science and Systems (RSS)*, 2020.

[7] E. Chisari, T. Welschehold, J. Boedecker, W. Burgard, and A. Valada, "Correct me if I am wrong: Interactive learning for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3695–3702, 2022.

[8] J. Puigcerver, C. Riquelme Ruiz, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *International Conference on Learning Representations (ICLR)*, 2024.

[9] L. Mazza, A. Rodriguez, R. Younis, M. Lelis, O. Hellig, C. Li, S. Bodenstedt, M. Wagner, and S. Speidel, "Supervised mixture-of-experts for surgical grasping and retraction," in *Robotics: Science and Systems (RSS)*, 2026.

[10] E. Kim, H. Jeong, S. Heo, S. Kim, S. Lee, J. Shin, H. Yoo, and J. Jeon, "Deployment-oriented end-to-end autonomous driving: Enhancing closed-loop stability with a lightweight framework," in *IEEE Intelligent Vehicles Symposium (IV)*, 2026.

[11] J. Zhang and K. Cho, "Query-efficient imitation learning for end-to-end simulated driving," in *AAAI Conference on Artificial Intelligence*, 2017.

[12] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer, "EnsembleDAgger: A Bayesian approach to safe imitation learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[13] Z. He, Y. Cao, and M. Ciocarlie, "Uncertainty comes for free: Human-in-the-loop policies with diffusion models," *arXiv preprint arXiv:2503.01876*, 2025.

[14] H. Cai, Z. Peng, and B. Zhou, "Robot-gated interactive imitation learning with adaptive intervention mechanism," in *International Conference on Machine Learning (ICML)*, 2025.

[15] J. Luijkx, Z. Ajanović, L. Ferranti, and J. Kober, "ASkDAgger: Active skill-level data aggregation for interactive imitation learning," *Transactions on Machine Learning Research*, 2025.

[16] Y. Huang, N. Ma, W. Zhao, Z. Liu, J. Sun, Q. Wang, and Y. Chen, "Force-aware residual DAgger via trajectory editing for precision insertion with impedance control," *arXiv preprint arXiv:2603.04038*, 2026.

[17] J. Tang and W. Zhi, "AutoIntervene: Calibrated intervention for action-chunking imitation learning policies," *arXiv preprint arXiv:2608.07065*, 2026.

[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Neural Information Processing Systems (NIPS)*, 2017.

[19] S.-W. Lee, X. Kang, and Y.-L. Kuo, "Diff-DAgger: Uncertainty estimation with diffusion policy for robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[20] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations (ICLR)*, 2017.

[21] V. Surendran, N. Peri, and D. Watkins, "Bridging handheld and teleoperated supervision for contact-rich manipulation via state-gated experts," *arXiv preprint arXiv:2606.26603*, 2026.

[22] S. Heo and J. Jeon, "SHIFT-RL: Sensor-driven hierarchical information fusion transformer for BEV-based maneuvering in dense multi-lane environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026.

[23] Z. Ouyang, K. Wang, J. Liu, H. Lu, and W. Zhang, "SCIL: Stage-conditioned imitation learning for multi-stage manipulation," *IEEE Control Systems Letters*, vol. 9, pp. 913–918, 2025.