OFFLINE IMITATION ERROR AS A MODEL-SELECTION PROXY UNDER DISTRIBUTION SHIFT: EVIDENCE FROM REAL-CAMERA ACTIVE PTZ CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Practitioners often screen imitation-learning policies with logged-state imitation error—per-step action error against an expert on logged data—before expensive closed-loop evaluation. We study when this proxy is valid on a real pan-tilt-zoom (PTZ) camera for indoor single-target tracking. **(RQ1)** Does in-distribution error preserve the ordering of held-out imitation error? Not across data-collection regimes: across three optimisation runs the held-out BC-DAgger separation persists, while in-distribution error favours behaviour-cloning (BC) policies in $178/189$ pooled cross-family comparisons; the DAgger reruns share one fixed collection chain, and orderings within the BC family do not survive retraining. **(RQ2)** Controlled retraining shows that access to the DAgger-collected support is sufficient for most of the held-out gain, and the warm-start fine-tuning trajectory is not required to recover it: plain BC from scratch on the collected states recovers it under matched-epoch and approximately matched-update budgets, without DAgger's in-distribution degradation; we do not show that learner-induced states are uniquely responsible. **(RQ3)** For our geometry-defined expert, a pretrained RGB branch raises held-out error $10\text{-}17\times$ even though the sufficient geometric inputs are given explicitly; in separate controls, constant and shuffled images stay much closer to geometry-only, consistent with the cost depending on state-aligned visual content rather than on the presence or size of the image branch, and a protocol-locked external check agrees in direction but is imprecise. Finally, a pre-specified live test of the same ten checkpoints separates *selection* from *ranking*: under free zoom the in-distribution choice has the highest mean telemetry-derived centred-tracking score and more in-frame coverage than the held-out choice in 5/5 blocks, while held-out error orders centred tracking across all ten far better (Spearman $+0.71$ vs. $+0.01$). Offline imitation error is thus not one model-selection signal: its validity depends on where it is evaluated, how the candidates acquired their training support, and whether the goal is to select one policy or rank many. All physical-system results come from one PTZ rig.

1 INTRODUCTION

Imitation-learning policies are often screened with cheap offline metrics before deployment. A common choice is per-step action error against an expert on a held-out log: policies with lower validation error are promoted, while policies with higher error are discarded. This practice is attractive because closed-loop evaluation on real hardware is slow, risky, and difficult to reproduce. However, it relies on a stronger assumption than is usually stated: not merely that offline error is correlated with deployment quality on average, but that it *preserves the ordering* among candidate policies well enough to support model selection.

That assumption is known to be imperfect. In autonomous driving, low offline prediction error need not imply good closed-loop behaviour (Codevilla et al., 2018a); in robot manipulation, checkpoint selection by validation loss can incur substantial online penalties (Mandlekar et al., 2021); and in continuous-control imitation, small expert-distribution errors can compound sharply under policy-induced shift (Simchowitz et al., 2025). Recent work further shows that raw offline metrics can disagree with closed-loop rankings in manipulation and driving (Huang et al., 2026; Wang et al.,

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

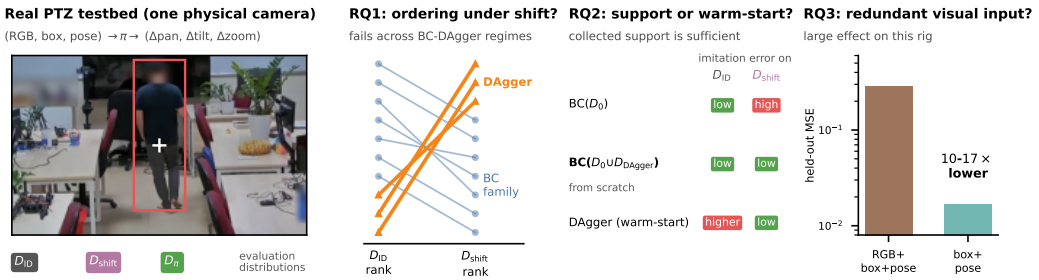

Figure 1: **Overview of the study.** On one physical PTZ camera (left; policy input is RGB, target box, and pose; output is a pan/tilt/zoom increment), policies are scored on three state distributions: logged in-distribution states with expert labels (D_{ID}), held-out actor/motion clips (D_{shift}), and the policy's own induced states (D_π). **RQ1**: the BC-DAgger ordering changes between logged in-distribution and held-out shifted evaluation (Section 4). **RQ2**: BC trained from scratch on the DAgger-collected states has low in-distribution and low held-out error, whereas iterative DAgger has low held-out but higher in-distribution error (Section 5). **RQ3**: dropping the RGB branch cuts held-out error 10-17× on this rig (Section 6). Proxy validity depends on the evaluation distribution, the candidates' training support, and the downstream objective.

2026). We therefore take proxy imperfection as a premise rather than a novelty claim. Our question is more specific: *which properties of the candidate set and evaluation distribution determine whether an offline imitation metric preserves the comparisons needed for model selection?* We separate three factors that are often entangled: the states used to evaluate a policy, the support on which candidates were trained, and the observations given to the learner; closed-loop tests then ask whether offline ranking fidelity carries over to the deployment objective.

We investigate this on active pan-tilt-zoom (PTZ) camera control for indoor single-target tracking. Unlike passive tracking (Milan et al., 2016; Dendorfer et al., 2020), where the camera stream is fixed, a PTZ controller changes the future observations it receives: camera pose is itself part of the decision process, which makes the setting a useful testbed for model selection under covariate shift. Our controller imitates a deterministic center-lock expert that maps the target box and current camera pose to the next pan, tilt, and zoom command; the expert can be queried on arbitrary visited states, which enables both behaviour cloning (BC) and interactive DAgger refinement. Prior learned PTZ controllers are mostly deep-RL policies trained in simulation (Luo et al., 2018; Zhong et al., 2019; Sandha et al., 2023); all our policy training and physical closed-loop evaluation use one real PTZ camera across two indoor rooms, with no simulator and no sim-to-real transfer. We organise the study around three research questions.

RQ1: When does logged-state imitation error preserve policy ordering? Across policies trained under different data-collection regimes, logged in-distribution error and held-out shifted error induce different orderings (Section 4): across three optimisation runs every DAgger checkpoint stays below every BC-family checkpoint in held-out error, while in-distribution error favours the BC family in 178/189 pooled cross-family comparisons (the reruns share one DAgger collection chain); fine-grained ordering within the BC family does not persist. Ranking is by held-out imitation error; deployment is tested separately.

RQ2: Is the BC-DAgger gap due to collected support or the warm-start trajectory? Ordinary BC trained from scratch on the demonstrations plus the DAgger-collected states recovers most of DAgger's held-out fidelity while keeping low in-distribution error, under matched-epoch and approximately matched-update budgets (Section 5). The collected support is sufficient; we do not show that policy-induced states are uniquely beneficial.

RQ3: What is the shift cost of a label-redundant high-dimensional input? For our geometry-defined expert, removing the RGB pathway reduces held-out imitation error by roughly an order of magnitude (10-17×; Section 6). Parameter-matched, shuffled-image, constant-image and frozen-encoder controls indicate that parameter count alone does not explain it; a protocol-locked external check agrees in direction but is imprecise, and the mechanism stays open.

Selection versus ranking, and the lesson. A pre-specified live test of the same ten checkpoints separates two uses of the proxy: under free zoom the in-distribution choice has the highest mean telemetry-derived centred-tracking score, yet held-out error ranks centred tracking across all ten far better (Section 4). Offline imitation error is thus not one model-selection signal: its validity depends on where it is evaluated, how the candidates acquired their training support, and whether the goal is to select one policy or rank many. To answer these questions without conflating policy architecture, training distribution, and evaluation distribution, we next define the testbed, the policy families, and the controlled comparisons used throughout (Figure 1).

2 Related Work

Offline metrics versus closed-loop behaviour. That a cheap offline metric imperfectly predicts closed-loop behaviour is established, and we take it as premise. Codevilla et al. (2018a) showed offline prediction error need not correlate with driving quality; Mandlekar et al. (2021) report that the best-validation-loss manipulation checkpoint can be 50-100% worse than the best rollout checkpoint, and real-robot evaluation is costly enough to motivate autonomous evaluation systems (Zhou et al., 2025). The same lesson drove driving from open-loop metrics to closed-loop benchmarks (Caesar et al., 2021; Dauner et al., 2024) after open-loop scores proved satisfiable without perception (Li et al., 2024); in model-based RL it is *objective mismatch* (Lambert et al., 2020). Theory agrees: beyond compounding-error bounds (Ross & Bagnell, 2010; Ross et al., 2011), Simchowitz et al. (2025) show a vanishing expert-distribution error is compatible with exponentially larger closed-loop error. Two concurrent studies probe the proxy on structured policy sets: Huang et al. (2026) find offline MSE an imperfect but *correct-direction* predictor of manipulation rollout success ($\rho{=}-0.61$, improved to -0.87 by task-critical reweighting, and report a raw-MSE inversion subgroup), and Wang et al. (2026) report open-versus-closed-loop ranking inversions in driving. Our contribution is therefore not a first demonstration of proxy failure but a controlled real-camera active-perception characterisation of *when* ordering fails and *why*; we adopt offline-policy-selection reporting (rank correlation and regret (Fu et al., 2021)); offline policy ranking from expert data is treated directly by Da et al. (2024); in the worst case, offline policy selection is as hard as off-policy evaluation (Liu et al., 2026).

Interactive imitation and distribution support. BC suffers compounding error off the logged collection distribution (Ross & Bagnell, 2010); DAgger (Ross et al., 2011) relabels the policy's own visited states, with later safety refinements (Kelly et al., 2019; Menda et al., 2019), and has served driving (Codevilla et al., 2018b) and drone servoing (Loquercio et al., 2018; Kaufmann et al., 2020). DAgger changes two things at once—the *states* available for training and, in our implementation, the *warm-start* optimisation trajectory across rounds; RQ2 fixes the collected data and asks whether the latter is required. Cost-sensitive imitation reweights *which* transitions matter (Zhang et al., 2021); we derive the per-frame weight from field-of-view geometry. Extra or task-irrelevant observations can hurt imitation (de Haan et al., 2019; Park et al., 2021; Bai et al., 2025; Dai et al., 2026); in our RQ3, RGB is label-redundant given the supplied geometry by construction, so its shift cost is measured with expert, labels and clips fixed.

Active camera control. Passive multi-object tracking pairs a detector with data association (Wojke et al., 2017; Zhang et al., 2022; Aharon et al., 2022) on fixed-view benchmarks (Milan et al., 2016; Dendorfer et al., 2020; Sun et al., 2022); we use such a tracker (Aharon et al., 2022; Jocher et al., 2023) as perception and treat pose as the decision. Classical PTZ work includes real-time tracking with background mosaicing (Kang et al., 2003), cooperative PTZ camera networks (Micheloni et al., 2010), and learning where a broadcast PTZ camera should look from human operators (Chen & Carr, 2015). Many prior learned PTZ controllers are deep-RL policies trained in simulation (Luo et al., 2018; Zhong et al., 2019; 2021; Sandha et al., 2023; Fang et al., 2022), and a prior imitation-based controller (Kyrkou, 2020) uses synthetic views. Learned control reached real cameras only recently (Yang et al., 2025; Kerr et al., 2025; Wang et al., 2025), with occlusion-aware active visual tracking (Sun et al., 2026) and learned orientation scheduling for commodity PTZ analytics (Kong et al., 2026). These works optimise tracking, perception or task performance; we study the validity of offline metrics for selecting among learned controllers. These lines motivate a controlled study in which we vary policy family, training support, observation design, and evaluation distribution while holding the expert fixed, on a single physical PTZ system.

3 STUDY DESIGN

3.1 REAL-CAMERA PTZ TESTBED

One Hikvision DS-2DE 25× PTZ camera streams RGB at 15 Hz (1920×1080) and accepts ISAPI commands at ≤3 Hz. At each control tick the agent observes $o_t = (I_t, \mathbf{p}_t, \mathbf{b}_t)$—the frame, the measured pan/tilt/zoom $\mathbf{p}_t$, and the box $\mathbf{b}_t = (c_t^x, c_t^y, w_t, h_t, v_t)$ of the selected target from an off-the-shelf detector/tracker (v_t a validity flag)—and emits a normalised increment $a_t = (\Delta p_t^{az}, \Delta p_t^{el}, \Delta z_t) \in [-1, 1]^3$; the command $\mathbf{p}_{t+1}^{\text{cmd}} = \text{clip}(\mathbf{p}_t + Sa_t)$ with $S = (20°, 10°, 90 \text{ zoom steps})$ per tick is sent as an absolute target pose and dropped by the rate limiter if the previous command is unacknowledged. The goal is to keep a *single* target in view and centred (ISAPI details in Appendix E). Demonstrations, DAgger, and offline checks were recorded in room A; the camera was then remounted in room B, where all closed-loop evaluation runs. Five 5-minute logged clips span distinct behaviours (perimeter walking, sit/stand, desk work, sharp turns, mixed motion), captured from a fixed wide pose so labels are unaffected by recording-time motion: `clip2`/`clip3`/`clip5` train (13,499 frames) and `clip4`/`clip6` form in-distribution validation (8,999). Two further clips form the *held-out-shift* set: `eval1` (a different held-out actor—appearance shift) and `eval2` (novel motion; its actor identity was not retained, so only `eval1` supports an actor-shift reading). Clip statistics are in Appendix A.

3.2 QUERYABLE CENTER-LOCK EXPERT

The expert is a deterministic function $a^\star = f(\mathbf{b}, \mathbf{p})$: sticky target selection (previous track ID, else the most confident person), FOV-based pan/tilt that converts the box-centre offset into an angular correction outside a dead zone, and a box-height zoom loop with a dead band (thresholds in Appendix B). The do-command bit $\texttt{do}_t$ marks frames where any loop issues a correction (85.4% of training frames). Because the expert is cheap and deterministic it can be queried at *any* state a policy visits—the property that DAgger, the on-policy diagnostic, and the geometry-only ablation all rely on. Writing $g_t = (\mathbf{b}_t, \mathbf{p}_t)$, the label is $a_t^\star = f(g_t)$, so two observations with the same geometry share the same label: once g is known, the RGB frame carries no further information about the expert action.

3.3 POLICY FAMILIES

CenterLock-IL maps $(I_t, \{\mathbf{p}_\tau, \mathbf{b}_\tau\}_{\tau=t-L+1}^{t})$ (history L=5 by default) to a TanhNormal over a_t: an ImageNet-pretrained ResNet18 encodes the *full* frame to 512-d (the target enters only through the box history), a two-layer MLP embeds the pose/box history to 64-d, and a two-layer head outputs $(\mu, \log\sigma)$ with deterministic action $\tanh\mu$ (11.27M parameters, 3.5 ms per inference; Appendix E). The policies vary along three axes.

Axis A: training support and optimisation. Flat BC minimises a do-command-masked MSE plus a small entropy bonus,

$$\mathcal{L}_{\text{BC}}(\theta) = \tfrac{1}{\sum_t \texttt{do}_t} \textstyle\sum_t \texttt{do}_t \,\|\tanh\mu_\theta(o_t) - a_t^{\text{exp}}\|_2^2 + \lambda_{\text{ent}}\,\mathbb{E}[-\log\sigma_\theta(o_t)], \tag{1}$$

with λ_{ent}=10^{-3} and $\log\sigma_\theta$ clamped to $[-5, 2]$, so the term is bounded (deployment uses only $\tanh\mu_\theta$); masking keeps the many "hold" frames from pulling μ_θ toward zero. *DAgger* (Ross et al., 2011) then runs three rounds on the physical camera: round i rolls out $\pi_i^\beta = \beta_i\pi^{\text{exp}} + (1-\beta_i)\pi_{i-1}$ for 10 min ($\beta_i \in \{0.5, 0.25, 0.1\}$), logs expert labels on every visited state, appends them to the dataset, and fine-tunes for 5 epochs at lr 10^{-4}, each round starting from the previous checkpoint (conventional iterative DAgger). *Aggregate-from-scratch* BC trains the flat-BC recipe from scratch on the union of the BC clips and the DAgger-collected states, isolating the states from the warm-start trajectory.

Axis B: observation design. The default policy observes RGB together with box+pose history; the *geometry-only* variant drops the RGB branch, keeping objective, clips, and expert.

Axis C: secondary interventions. History $L\in\{1, 15\}$, a no-zoom head (A=2), and a *FOV-risk* weighting that scales each masked BC term by $w(s_t) = 1 + \alpha\, r(s_t)$ (renormalised per batch; α=0

Table 1: Experimental identification strategy: each comparison changes one primary factor relative to flat BC (**bold**); all others are held fixed. $D_{1:k}$: states collected in DAgger rounds $1..k$.

Comparison	Training states	Optimisation	Observation	Scored on	RQ
Flat BC	D_0	from scratch	RGB+geom.	D_{ID}, D_{shift}	1
BC variants (L, A, FOV-risk)	D_0	**loss / architecture**	RGB+geom.	D_{ID}, D_{shift}	1
DAgger-R1-R3	$\mathbf{D_0 \cup D_{1:k}}$	**iterative fine-tune**	RGB+geom.	D_{ID}, D_{shift}	1, 2
Aggregate-from-scratch	$\mathbf{D_0 \cup D_{1:3}}$	from scratch	RGB+geom.	D_{ID}, D_{shift}	2
q-mixtures	$\mathbf{D_0 \cup q\,D_{1:3}}$	from scratch	RGB+geom.	D_{ID}, D_{shift}	2
Geometry-only BC	D_0	from scratch	**geom. only**	D_{ID}, D_{shift}	3
OPED (diagnostic)	—	—	—	$\mathbf{D}_\pi$	supp.

is flat BC), where $r(s_t) = \text{clip}_{[0,1]}(\lambda_e e_t^2 + \lambda_s \varsigma_t + \lambda_\ell(1 - v_t))$, $e_t = 2\max(|c^x - \frac{1}{2}|, |c^y - \frac{1}{2}|)$ measures closeness to the frame border, $\varsigma_t = \text{clip}_{[0,1]}((h_{\text{safe}} - h)/h_{\text{safe}})$ penalises a shrinking target, and $(\lambda_e, \lambda_s, \lambda_\ell, h_{\text{safe}}) = (0.6, 0.3, 0.1, 0.25)$. No commanded frame ($\texttt{do}_t$=1) has a lost target, so the lost-target term contributes only with hold supervision (below), where the same FOV-risk weight also multiplies the $\texttt{do}_t$=0 hold term. Cost-sensitive imitation is established (Zhang et al., 2021); here the cost is derived from field-of-view geometry. For the closed-loop matrix only, we optionally add *hold supervision*, $\lambda_h \,\|\tanh \mu_\theta\|_2^2$ on $\texttt{do}_t$=0 frames, which curbs over-commanding at 2× zoom; all policies in the offline ranking use λ_h=0.

3.4 Controlled Comparisons

Table 1 lists which factor each comparison changes relative to flat BC; every other factor is held fixed. This design lets us ask first whether the cheap in-distribution metric preserves ordering at all (RQ1), then which difference between BC and DAgger produces the gap (RQ2), and whether observation design contributes independently (RQ3).

3.5 Evaluation Distributions and Metrics

The same policy can look strong or weak depending on where it is scored, so we separate three state distributions: D_{ID}, the logged in-distribution validation clips; D_{shift}, the held-out-shift clips; and D_π, the states a policy itself induces on the camera. The primary offline quantity is the do-masked imitation error $E(\pi; D) = \mathbb{E}_D\big[\texttt{do} \,\|\pi(o) - a^\star\|_2^2\big]/\mathbb{E}_D[\texttt{do}]$, with $\pi(o) = \tanh \mu_\theta(o)$, over pan and tilt (zoom excluded so A=2 is commensurable), pooled over a distribution's clips. For a policy set Π we define *ranking validity* from D_a to D_b and the *selection regret* of screening on D_a (Fu et al., 2021):

$$\begin{aligned} R(D_a, D_b; \Pi) &= \rho_{\text{S}}\big(\{E(\pi; D_a)\}_{\pi\in\Pi}, \{E(\pi; D_b)\}_{\pi\in\Pi}\big), \\ \text{Reg}(D_a \to D_b) &= E(\hat{\pi}_a; D_b) - \min_{\pi\in\Pi} E(\pi; D_b), \end{aligned} \tag{2}$$

with ρ_{S} the Spearman correlation and $\hat{\pi}_a = \arg\min_{\pi\in\Pi} E(\pi; D_a)$. For a pair (π_i, π_j) let $\Delta_{ij}^D = E(\pi_i; D) - E(\pi_j; D)$ and let $s(\pi) = E(\pi; D_{\text{shift}}) - E(\pi; D_{\text{ID}})$ be the policy's *shift penalty*. Then, identically,

$$\Delta_{ij}^{\text{shift}} = \Delta_{ij}^{\text{ID}} + \big[s(\pi_i) - s(\pi_j)\big], \tag{3}$$

so a pair reverses ($\Delta_{ij}^{\text{ID}}\Delta_{ij}^{\text{shift}} < 0$) exactly when the difference in shift penalties opposes the in-distribution margin and exceeds it in magnitude: an in-distribution ordering is only as safe as its margin is large relative to how differently the two policies degrade under shift. The identity is descriptive: without shifted data $s(\pi)$ is unknown, so it accounts for reversals rather than predicting them. RQ1 measures R within the BC family and across the BC and DAgger families. Live, the policy runs in the same rate-limited loop used for demonstrations (RTSP capture, YOLOv8+BoT-SORT, ISAPI writes $\leq$3 Hz), with recovery re-home *disabled* in all reported evaluation. We define reproducible scenarios in which the actor repeatedly leaves the field, run policies interleaved in rotating order from a common start pose (actor blind to the controller), and report HOTA (Luiten et al., 2021), IDF1 (Ristani et al., 2016), MOTA (Bernardin & Stiefelhagen, 2008), coverage (in-frame ratio), and centring error (CErr).

Table 2: Offline imitation-error rankings on the in-distribution and held-out clips: pan+tilt MSE (lower is better) and induced rank for the ten learned policies (seed 0, the checkpoints used throughout). Fixed-wide and sweep are non-learned references, excluded from correlations.

	in-distribution		held-out clips		
Policy	MSE	rank	MSE	rank	Δrank
fixed-wide	0.7014	—	0.5699	—	—
sweep	0.7873	—	0.6618	—	—
L=1	**0.0036**	1	0.1533	4	+3
FOV-risk α=3	0.0036	2	0.1769	5	+3
FOV-risk α=1	0.0042	3	0.2650	8	+5
A=2 (no zoom)	0.0046	4	0.2529	7	+3
FOV-risk α=2	0.0047	5	0.2324	6	+1
L=15	0.0058	6	0.2775	9	+3
BC (α=0)	0.0065	7	0.3128	10	+3
DAgger-R1	0.0141	8	0.1027	3	−5
DAgger-R2	0.0208	9	**0.0728**	1	−8
DAgger-R3	0.0227	10	0.0980	2	−8

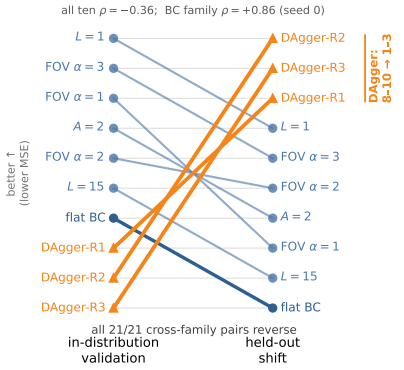

Figure 2: **Rank flow under held-out shift.** Each line joins a policy's in-distribution rank (left) to its held-out rank (right). DAgger moves from ranks 8-10 to 1-3; all $21/21$ cross-family pairs reverse. The BC-family order is seed 0 and does not survive retraining (Figure 3); the aggregate $\rho = -0.36$ is descriptive (CI $[-0.79, +0.44]$).

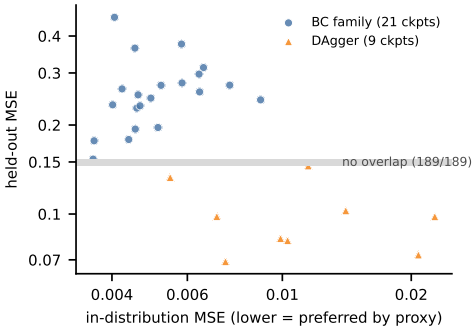

Figure 3: **The held-out cross-regime separation persists across optimisation seeds.** All 30 checkpoints (ten policies, three optimisation seeds; DAgger data fixed): every DAgger checkpoint has lower held-out error than every BC-family one (grey band; $189/189$ pairs) while in-distribution error favours BC ($178/189$). Within the BC family, held-out error is not ordered by in-distribution error.

4 RQ1: When Does Offline Error Preserve Policy Ordering?

Scored on D_{ID}, flat BC beats a fixed wide camera $36/210/7\times$ on pan/tilt/zoom; on D_{shift} only $1.1/2.8/3.4\times$—on pan barely better than holding still (Table 6, Appendix C). A practitioner, however, uses the proxy to *rank*, so we score every policy on both distributions (Table 2).

The two orderings are *negatively* correlated ($R(D_{\text{ID}}, D_{\text{shift}}) = -0.358$ over all ten, Eq. 2; Kendall $\tau = -0.156$; 26 of 45 pairs discordant, 21 of them crossing the BC-DAgger boundary). In-distribution, BC has $3.5\times$ lower error than DAgger-R3 and the proxy ranks the DAgger chain last; held-out, the chain takes the top three and BC falls to last (Figure 2). In selection terms, choosing the lowest-in-distribution policy incurs $2.1\times$ the best held-out error ($\text{Reg}(D_{\text{ID}} \to D_{\text{shift}}) = 0.081$), top-3 recall is $0/3$, and pairwise order agreement is 0.42. The aggregate alone is uncertain (bootstrap 95% CI $[-0.79, +0.44]$, permutation p=0.31), so we rest the claim on the two analyses below.

Seed robustness. Retraining every policy with two further seeds (30 checkpoints; DAgger seeds vary the optimisation on the fixed collected data; Appendix H) separates what is robust from what is not (Figure 3). The held-out separation persists in every seed: *every* DAgger checkpoint has lower held-out error than every BC-family checkpoint (all 189 cross-family comparisons; worst DAgger 0.146 vs. best BC-family 0.153) and higher in-distribution error in $178/189$, and the aggregate ρ is negative for 99.9% of the 3^{10} one-seed-per-policy combinations (mean -0.41). The within-family order is not: its ρ is $+0.86$, $+0.54$, and -0.21 across seeds, because run-to-run variation within the

Table 3: Same-checkpoint live test (five blocks, mean per policy) with the offline errors of Table 2. Coverage = telemetry in-frame ratio; $J_{0.25}$ = fraction of evaluation ticks in frame and centred; $<2\times$ = fraction of ticks below $2\times$ zoom. †Selected by in-distribution error; ‡selected by held-out error.

Policy	MSE in-dist.	MSE held-out	Coverage ↑	$J_{0.25}$ ↑	$<2\times$
flat BC	0.0065	0.313	0.936	0.090	0.99
L=1†	**0.0036**	0.153	0.907	**0.727**	0.99
FOV-risk α=2	0.0047	0.232	0.826	0.615	0.99
FOV-risk α=1	0.0042	0.265	0.807	0.122	0.99
DAgger-R2‡	0.0208	**0.073**	0.640	0.549	0.87
DAgger-R1	0.0141	0.103	0.619	0.485	0.99
A=2 (no zoom axis)	0.0046	0.253	0.611	0.052	0.00
DAgger-R3	0.0227	0.098	0.571	0.496	0.85
FOV-risk α=3	0.0036	0.177	0.464	0.125	0.99
L=15	0.0058	0.278	0.054	0.029	0.83

BC family (held-out s.d. up to 0.16) is as large as the differences between variants. All pairs share one DAgger collection chain, so they are not independent replicates.

Why the order reverses. By Equation 3, cross-family in-distribution margins (0.008-0.019) are an order of magnitude smaller than the DAgger policies' shift-penalty advantage (0.06-0.25), so all $21/21$ cross-family pairs reverse; within the BC family margins are tiny (median 0.001) against shift-penalty differences (median 0.058), so that order is unstable. Per clip, the in-distribution penalty sits on the near-stationary clip (Appendix J).

Same-checkpoint live test (pre-specified). We deployed the ten Table 2 checkpoints unchanged (five Williams-ordered blocks of ten 75 s rollouts, one scripted actor, initialised at $2\times$ with policy-controlled zoom; rules locked before data; Table 3, Appendix N). The primary endpoint, telemetry coverage for DAgger-R2 (held-out choice) minus L=1 (in-distribution choice), falls in the locked category *no mis-selection detected*: L=1 kept the target in frame longer in 5/5 blocks (-0.27, 95% CI $[-0.42, -0.15]$); a post-hoc single-rater check of offline-detector boxes confirms it (-0.20 $[-0.38, -0.05]$, 5/5) but not the $J_{0.25}$ difference. Selection and ranking then diverge. *Top-1 selection*: the in-distribution choice L=1 has the highest mean telemetry $J_{0.25}$ of the ten and is within 0.03 of the best coverage (regret 0.00/0.03 on $J_{0.25}$/coverage, vs. 0.18/0.30 for the held-out choice). *Ranking*: over all ten, held-out error orders $J_{0.25}$ well (Spearman $+0.71$ vs. $+0.01$), and neither proxy orders coverage ($+0.15$, -0.14). Zoom-controlling policies spent 83-99% of ticks below $2\times$ (mostly $1\times$), so coverage rewards zooming out (flat BC: coverage 0.94, $J_{0.25}$ 0.09; zoom-floor follow-ups: App. O, P; sensitivity: App. Q).

Answer to RQ1. Logged-state imitation error is not a stable ordering proxy across the training regimes in our candidate set. The robust result is not the aggregate correlation over ten checkpoints but the repeated BC-DAgger regime separation, while ordering within the BC family is comparable to optimisation noise; by Equation 3, rankings change when shift-penalty differences exceed the in-distribution margins. Live, the in-distribution choice had the highest mean telemetry $J_{0.25}$ under free zoom, whereas held-out error gave far better rank agreement with $J_{0.25}$ across the ten: offline fidelity, top-1 selection and global ranking are different requirements. We establish a failure mode of logged-state ranking across collection regimes, not a universally superior selector. DAgger differs from flat BC both in its training states and in its warm-start optimisation, which we isolate next.

5 RQ2: Is the Warm-Start Trajectory Needed Given the Collected Data?

Is the warm-start schedule necessary? Training plain BC *from scratch* on the union of the BC clips and all three rounds of DAgger-collected states ($D_0 \cup D_{1:3}$; flat-BC objective, epochs, and architecture; ten seeds) recovers DAgger-R3's held-out fidelity—median held-out MSE 0.05 (mean $0.07_{\pm.04}$; Appendix L), comparable to DAgger-R3 ($0.087_{\pm.009}$ over three optimisation seeds) and well below flat BC ($0.325_{\pm.035}$)—while keeping in-distribution error at the BC-family level (0.004, vs. DAgger-R3's 0.023). The collected states are therefore *sufficient* to recover most of the held-out gain under this recipe; the warm-start trajectory is not required for that recovery.

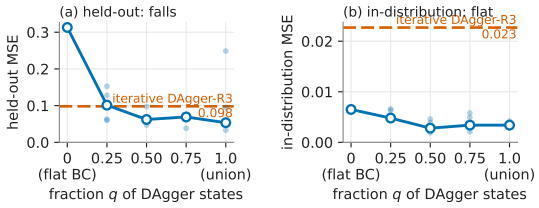

Figure 4: **Collected support recovers most of the held-out gain without warm-start** (matched 20 epochs, five seeds for $q{>}0$, median line; $q{=}0$ and DAgger-R3 are seed-0). Adding 25% of the DAgger-collected states cuts held-out imitation error 0.313 (seed 0) $\rightarrow$ median 0.101 *(a)* while in-distribution error stays $\sim$0.003-0.006 *(b)*; warm-start DAgger-R3 (dashed) sits at in-distribution 0.023. Sufficiency, not uniqueness (Section 5).

How much collected data suffices? Let $\theta_q = \mathrm{Train}_{20}\big(D_0 \cup S_q(D_{\mathrm{Dag}})\big)$ be a policy trained from scratch for the *same 20-epoch budget as flat BC* on D_0 plus a fraction q of the DAgger-collected states $D_{\mathrm{Dag}}{=}D_{1:3}$ (θ_0 is flat BC), with *shift gain* $G(q) = E(\pi_{\theta_0}; D_{\mathrm{shift}}) - E(\pi_{\theta_q}; D_{\mathrm{shift}})$ and *in-distribution cost* $C(q) = E(\pi_{\theta_q}; D_{\mathrm{ID}}) - E(\pi_{\theta_0}; D_{\mathrm{ID}})$. At $q{=}0.25$ held-out MSE falls from 0.313 (seed 0) to a median of 0.101, so $G{=}0.21$, while $C{\approx}{-}0.002$ (no cost); by $q{=}0.5$ held-out error is $\sim$0.06 and in-distribution error stays $\sim$0.003-0.006 throughout (Figure 4). At a fixed $\approx$8.6k-update budget (Appendix L), training longer on D_0 alone lowers held-out error $0.313 \rightarrow 0.173$, and adding 25% of the collected states lowers it further to 0.060.

Answer to RQ2. Given the fixed collected dataset, DAgger's held-out advantage does not require the warm-start optimisation trajectory: ordinary BC recovers most of the shifted-fidelity gain while keeping flat-BC in-distribution error, under both matched-epoch and approximately matched-update budgets. The supported causal statement is limited: access to the collected support is *sufficient*; without an equal-budget control of comparably diverse learner-independent states, we do not claim that policy-induced support is necessary or superior.

6 RQ3: What Is the Shift Cost of a Label-Redundant High-Dimensional Input?

With the same training clips, expert, and masked-MSE objective, geometry-only BC (box+pose history, no RGB frame) obtains in-distribution pan+tilt MSE $\mathbf{0.0037}{\pm}0.0002$ and held-out $\mathbf{0.0167}{\pm}0.0018$ over three seeds ($\pm$ s.d.); the RGB+geometry baseline, trained on the same seeds and pipeline, obtains held-out $0.289_{\pm.034}$ (in-distribution 0.0065)—$\sim$17$\times$ higher under a paired-seed comparison ($\sim$10$\times$ in a five-seed replication, where a parameter-matched geometry MLP and shuffled or constant images stay near geometry-only; Appendix M). The direction is expected: the expert is a function of box and pose (Appendix B), so the geometry-only control head is appearance-invariant by construction, though its box still comes from an RGB detector (not an end-to-end guarantee). Because RGB is label-redundant given geometry (Section 3.2), what is notable is not that geometry-only is better but the size of the degradation even when the sufficient variables are given explicitly; the controls are consistent with the cost depending on state-aligned visual content rather than on capacity alone, but they do not identify the learned mechanism.

External observation-shift check (protocol-locked, disclosed amendment). On Quad-Track (Luo et al., 2025), a public panoramic tracking dataset, we crop perspective views on a fixed, policy-independent schedule and hold out all 16 annotated sequences once by five-fold cross-fitting (five seeds; protocol, amendment, and disclosure in Appendix K). Geometry-only again has lower error in all five seeds (Table 4), but the interval of the relative reduction exceeds the pre-specified 15-point precision gate, so the external result is classified *imprecise*: directionally positive, not a definitive replication. Corrected shuffled- and constant-image controls again sit near geometry-only (descriptive; Appendix K).

Answer to RQ3. On this rig, a label-redundant RGB observation imposes a large shifted-fidelity cost even though the sufficient geometric variables remain available; parameter-matched and image-content controls indicate that capacity alone does not explain it, and the external check reproduces

Table 4: Pan+tilt imitation MSE on the rig (mean±s.d., 3 seeds) and on QuadTrack (median of 5 seeds, 16 sequences); R = reduction by geometry-only [95% CI]. The QuadTrack R interval fails the pre-specified 15-point precision gate. †Seed-0 checkpoint. ‡Median of 3 seeds, Appendix M. Shuffled/constant rows are descriptive controls.

Observation	Rig in-dist.	Rig held-out	QuadTrack held-out
RGB+geometry	$0.0065^{\dagger}$	$0.289_{\pm.034}$	0.0085
geometry only	$0.0037_{\pm.0002}$	$0.0167_{\pm.0018}$	0.0022
shuffled image+geometry	$0.0064^{\ddagger}$	$0.040^{\ddagger}$	0.0029
constant image+geometry	$0.0049^{\ddagger}$	$0.023^{\ddagger}$	0.0033
R	—	94%	74% [59, 80]

the direction but not a precise magnitude. We read RQ3 as evidence for a finite-sample nuisance-learning failure mode, not that RGB is generally harmful.

7 Practical Diagnostics and Physical Evidence

On-policy expert disagreement (OPED). OPED scores the same do-masked pan+tilt error on each policy's own induced states, $E_{\text{OPED}}(\pi) = E(\pi; d_\pi)$ (target-lost ticks excluded). It aligns with the held-out ordering ($\rho = +0.79$, n=8) where the in-distribution proxy does not (-0.10); it is a diagnostic, not a validated selector.

Physical closed-loop evidence. A five-round $2\times$ matrix of four *retrained* hold-supervised policies (HOTA against manually corrected boxes; Table 10, Appendix I) tests active control, not ranking: every learned policy beats a static wide view on HOTA (0.76-0.81 vs. 0.59), and FOV-risk lowers telemetry centring error over flat BC in 5/5 rounds at a coverage cost.

8 Discussion and Limitations

Our experiments suggest three distinctions for using offline imitation error in model selection, consistent with the distribution-shift account (Ross et al., 2011; Simchowitz et al., 2025). **Evaluation distribution:** low logged-state error does not imply that margins between candidates survive a shift; the BC-DAgger ranking changes because shift penalties differ far more than the small in-distribution margins. **Candidate training support:** comparisons are fragile when candidates were trained on different state supports, and adding the stronger RQ2/RQ3 recipes turns the offline correlation positive (post hoc, Appendix Q); most of the DAgger-family held-out advantage returns without the warm-start trajectory once the collected states are available. **Selection objective:** preserving a ranking and selecting the best deployment policy are different requirements; live, held-out error orders centred tracking across the set while the in-distribution proxy picks the policy with the best mean telemetry $J_{0.25}$ under free zoom. An offline metric should thus be validated together with the candidate-generation process, the anticipated shift, and the downstream decision.

Limitations. Free zoom is a deployment confound: most zoom-controlling policies spend most ticks near $1\times$, so raw coverage partly rewards zooming out; exploratory $2\times$-floor follow-ups change several pairwise and ranking patterns but were designed after the confirmatory result with limited replication, so they only motivate a controlled replication. RQ2 separates the final collected support from the warm-start trajectory, not iterative from non-iterative collection. We have one DAgger collection chain, one actor and room in the live test, and one PTZ unit (Appendix G).

9 Conclusion

Offline imitation error is useful for model selection only relative to a candidate set, an evaluation distribution, and a downstream objective. On a real PTZ system, logged-state error does not preserve the BC-DAgger ordering under shift, the collected support suffices for most of that gap, label-redundant RGB raises shifted error, and better global rank agreement need not select the best policy for a given deployment metric.

REPRODUCIBILITY STATEMENT

All physical-system policy training and closed-loop experiments use a single physical PTZ camera, with no simulator or sim-to-real transfer; the public QuadTrack dataset (Luo et al., 2025) is used only for the policy-independent offline observation-shift check in RQ3 (protocol, amendment, and bands in Appendix K). The expert, policy architecture, imitation losses, DAgger schedule, and evaluation metrics are specified in Section 3; hyperparameters, ISAPI control details, the deployment loop, and the offline evaluation procedure are in Appendix E, and clip statistics in Appendix A. The main ranking table uses one checkpoint per policy; Appendix H retrains all ten policies with two further optimisation seeds, where the DAgger reruns reuse the same fixed DAgger-collected datasets and so measure optimisation rather than collection variance. The support-mixture and FOV-risk experiments use five training seeds. The narrow-FOV closed-loop study uses five interleaved rounds with rotating policy order; tracking metrics are computed against target boxes seeded by an independent offline detector and manually corrected frame by frame for all 30 runs (per-round values in Appendix I). The same-checkpoint live test (Appendix N) deploys the ten Table 2 checkpoints byte-identical (SHA-256 checked at every launch) in 50 rollouts over five pre-assigned blocks; its endpoints, decision categories and every protocol addendum were fixed and hashed before the data they govern, and per-run telemetry, run manifests, checkpoint hashes and the analysis script are released with the supplement. Exact pixel-level replay of closed-loop trajectories is impossible because each policy changes its own subsequent view, so we fix scenario definitions, start poses, and lighting, interleave policy order, and report round-level variation rather than treating frames as independent trials.

ETHICS STATEMENT

The physical-camera study involves consenting participants in controlled indoor scenes staged for data collection and evaluation, with one designated target tracked at a time. No formal ethics-board review was conducted; at the time of data collection, the authors' institution had no specific ethics-review procedure applicable to this type of controlled, consent-based recording study. The training and in-distribution validation clips use the same demonstration actor, while `eval1` uses a distinct held-out actor; the identity associated with `eval2` was not retained in the capture metadata, so we make no actor-specific claim for that clip. External public panoramic footage (QuadTrack, CC BY-NC 4.0 per its dataset card) is used only offline for observation-shift evaluation, with no identity inference. No bystanders are intentionally recorded, no biometric identification or identity inference is performed, and the system uses the designated target's image location only for camera framing. The study does not evaluate surveillance of unaware individuals or unconstrained public deployment. We acknowledge the dual-use potential of active visual tracking and restrict all claims to the controlled, single-target, few-participant setting described here.

USE OF LARGE LANGUAGE MODELS

Per the ICLR 2027 policy we disclose our use of AI assistants. Large language models were used (i) for methodology feedback and research ideation, including literature retrieval; (ii) for experimental-design support, including drafting protocol documents that the authors reviewed and locked; (iii) for writing, running and debugging training, evaluation, analysis and live-session tooling scripts; (iv) for help interpreting results; and (v) to draft and polish the writing of this paper. All experimental design, the choice and interpretation of results, every reported number, and all scientific claims were verified and are the responsibility of the authors; the LLMs did not generate synthetic data or proofs, and no result was reported without author verification against the underlying logs.

REFERENCES

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.

Shuanghao Bai, Wanqi Zhou, Pengxiang Ding, Wei Zhao, Donglin Wang, and Badong Chen. Rethinking latent redundancy in behavior cloning: An information bottleneck approach for robot ma-

nipulation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pp. 2560–2580, 2025.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image and Video Processing*, 2008.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, et al. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Jianhui Chen and Peter Carr. Mimicking human camera operators. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 215–222, 2015.

Felipe Codevilla, Antonio M López, Vladlen Koltun, and Alexey Dosovitskiy. On offline evaluation of vision-based driving models. In *European Conference on Computer Vision (ECCV)*, 2018a. arXiv:1809.04843.

Felipe Codevilla, Matthias Müller, Antonio Lopez, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, 2018b.

Longchao Da, Porter Jenkins, Trevor Schwantes, Jeffrey Dotson, and Hua Wei. Probabilistic offline policy ranking with approximate Bayesian computation. In *AAAI Conference on Artificial Intelligence*, 2024.

Yinlong Dai, Robert Ramirez Sanchez, Ryan Jeronimus, Shahabedin Sagheb, Cara M. Nunez, Heramb Nemlekar, and Dylan P. Losey. CIVIL: Causal and intuitive visual imitation learning. *Autonomous Robots*, 2026. doi: 10.1007/s10514-026-10266-3.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2024. arXiv:2406.15349.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11666–11677, 2019.

Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixe. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

Zeyu Fang, Jian Zhao, Mingyu Yang, Wengang Zhou, Zhenbo Lu, and Houqiang Li. Coordinate-aligned multi-camera collaboration for active multi-object tracking. *arXiv preprint arXiv:2202.10881*, 2022.

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, et al. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2103.16596.

Haoxu Huang, Tongsam Zheng, Yifan Chen, Jiacheng You, and Yang Gao. Critical interval MSE: Toward reliable offline validation for robot manipulation policies, 2026. arXiv preprint arXiv:2606.29898.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLOv8 by Ultralytics. `https://github.com/ultralytics/ultralytics`, 2023.

Sangkyu Kang, Joon-Ki Paik, Andreas Koschan, Besma R. Abidi, and Mongi A. Abidi. Real-time video tracking using PTZ cameras. In *Proc. SPIE 5132, Sixth International Conference on Quality Control by Artificial Vision*, pp. 103–111, 2003.

Elia Kaufmann, Antonio Loquercio, Rene Ranftl, Matthias Mueller, Vladlen Koltun, and Davide Scaramuzza. Deep drone acrobatics. In *RSS*, 2020.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *ICRA*, 2019.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a BC-RL perception-action loop. In *CoRL*, 2025. arXiv preprint arXiv:2506.10968.

Z. Jonny Kong, Sibendu Paul, and Y. Charlie Hu. WideEye: Achieving wide field-of-view traffic video analytics with dynamic orientation adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Findings*, 2026.

Christos Kyrkou. Imitation-based active camera control with deep convolutional neural network. In *IEEE Int. Conf. on Image Processing Applications and Systems (IPAS)*, 2020.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. *Learning for Dynamics and Control (L4DC)*, 2020. arXiv:2002.04523.

Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. arXiv:2312.03031.

Vincent Liu, Prabhat Nagarajan, Andrew Patterson, and Martha White. When is offline policy selection sample efficient for reinforcement learning? In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026.

Antonio Loquercio, Ana I. Maqueda, Carlos R. Del-Blanco, and Davide Scaramuzza. DroNet: Learning to fly by driving. In *IROS*, 2018.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *IJCV*, 2021.

Kai Luo, Hao Shi, Sheng Wu, Fei Teng, Mengfei Duan, Chang Huang, Yuhang Wang, Kaiwei Wang, and Kailun Yang. Omnidirectional multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21959–21969, 2025.

Wenhan Luo, Peng Sun, Fangwei Zhong, Wei Liu, Tong Zhang, and Yizhou Wang. End-to-end active object tracking via reinforcement learning. In *ICML*, 2018.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021. arXiv:2108.03298.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J. Kochenderfer. EnsembleDAgger: A bayesian approach to safe imitation learning. In *IROS*, 2019.

Christian Micheloni, Bernhard Rinner, and Gian Luca Foresti. Video analysis in pan-tilt-zoom camera networks. *IEEE Signal Processing Magazine*, 27(5):78–90, 2010.

Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

Jongjin Park, Younggyo Seo, Chang Liu, Li Zhao, Tao Qin, Jinwoo Shin, and Tie-Yan Liu. Object-aware regularization for addressing causal confusion in imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target multi-camera tracking. In *ECCV Workshops*, 2016.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *AISTATS*, 2010.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Sandeep Singh Sandha, Bharathan Balaji, Luis Garcia, and Mani Srivastava. Eagle: End-to-end deep reinforcement learning based autonomous control of PTZ cameras. In *ACM/IEEE Int. Conf. on Internet of Things Design and Implementation (IoTDI)*, 2023.

Max Simchowitz, Daniel Pfrommer, and Ali Jadbabaie. The pitfalls of imitation learning when actions are continuous. *arXiv preprint arXiv:2503.09722*, 2025.

Haowei Sun, Kai Zhou, Hao Gao, Shiteng Zhang, Jinwu Hu, Xutao Wen, Qixiang Ye, and Mingkui Tan. Instance-level visual active tracking with occlusion-aware planning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 2022.

Shaoan Wang, Jiazhao Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. TrackVLA: Embodied visual tracking in the wild. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *PMLR*, pp. 4139–4164, 2025.

Yiru Wang, Anqing Jiang, Shuo Wang, Yuwen Heng, Hai Yang, Yang Chen, and Hao Sun. Do open-loop metrics predict closed-loop driving? a cross-benchmark correlation study of NAVSIM and Bench2Drive, 2026. arXiv preprint arXiv:2605.00066.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017.

Jiashu Yang, Yifan Han, Yucheng Xie, Ning Guo, and Wenzhao Lian. Look, zoom, understand: The robotic eyeball for embodied perception, 2025. arXiv preprint arXiv:2511.15279.

Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-aware imitation learning from demonstrations with varying optimality. In *NeurIPS*, 2021.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022.

Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. AD-VAT: An asymmetric dueling mechanism for learning visual active tracking. In *ICLR*, 2019.

Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Towards distraction-robust active visual tracking. In *ICML*, 2021.

Zhiyuan Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, and Sergey Levine. AutoEval: Autonomous evaluation of generalist robot manipulation policies in the real world. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *PMLR*, pp. 1997–2017, 2025.

A Training-Clip Statistics

Table 5: Logged clip statistics from the center-lock expert (three clips train, two validation). Lost-frame percentage reflects detector dropouts under heavy motion or partial occlusion; `clip6` is the hardest of the five. `clip4` and `clip6` are held out for validation.

clip	frames	target visible (%)	do-cmd (%)	lost (%)	$\sigma(\Delta p^{az})$
clip2_perimeter_walk	4,498	74.9	74.9	3.6	0.48
clip3_sit_stand	4,499	98.4	98.1	0.0	0.18
clip4_desk_dwell	4,499	99.7	99.7	0.0	0.00
clip5_quick_turns	4,502	81.4	81.4	3.5	0.48
clip6_mixed_natural	4,500	73.1	73.1	11.4	0.48
total	22,498	85.5	85.4	3.7	0.39

B CENTER-LOCK EXPERT: EXACT PARAMETERS

Sticky target selection. At each tick the expert prefers the previous frame's tracked ID; if that ID is missing it picks the highest-confidence person detection above a 0.35 confidence floor. The sticky preference survives $T_{\text{stick}} = 25$ frames of absence (a grace period during which the camera holds pose rather than commit to a new target); after a separate $T_{\text{lost}} = 25$ frames of *any* target being available, the `lost` flag is raised for telemetry but the camera still holds.

FOV-based pan/tilt mapping. The horizontal field of view is approximated as $\theta_{HFOV} = 55^\circ / z_{\text{factor}}$ with $z_{\text{factor}} = z/10$ clamped to $[1, 25]$, and the vertical field of view follows the image aspect ratio, $\theta_{VFOV} = \theta_{HFOV}\, H/W$. A pixel offset $(\Delta x, \Delta y)$ between the box centroid and image centre maps to $(\Delta p^{az}, \Delta p^{el}) = (\theta_{HFOV}\Delta x/W,\ \theta_{VFOV}\Delta y/H)$, applied only when $|\Delta x|/W$ or $|\Delta y|/H$ exceeds a 0.15 dead zone.

Bounding-box height zoom controller. The expert targets a normalised box height $h/H = 0.35$ with a ± 0.10 dead band. Outside the band the per-tick zoom command is $\Delta z = \text{clip}(\lfloor 60\,(0.35 - h/H) \rfloor, -90, +90)$ in ISAPI step units, with the absolute target clamped to $z \in [1, 1000]$. The 60-step coefficient was tuned empirically so that a person filling the centre of the frame at $1\times$ zoom approaches the target height in three to four ticks rather than overshooting.

C PER-CLIP OFFLINE IMITATION MSE

Table 6 aggregates within regime; the per-clip breakdown is below. Note that `clip4_desk_dwell` is near-stationary, so the zero-action fixed-wide baseline is near-optimal there on *pan* (not tilt; see the table) and the in-distribution pan contrast is correspondingly flattered.

Table 6: Offline imitation MSE per axis, aggregated within regime (lower is better). *Top:* validation clips (held out by clip; same actor and scene as training). *Bottom:* held-out evaluation clips (`eval1`: held-out actor; `eval2`: novel motion, unscripted). In-distribution BC stays at or below 1×10^{-2} on every axis, but that margin collapses on the held-out clips, shrinking against fixed-wide to $\approx$1.1×/$\approx$2.8×/$\approx$3.4× on pan/tilt/zoom (on pan, barely better than holding still). Note that this comparison flatters BC by construction—BC is fit to minimise precisely this quantity, while the analytic baselines never optimised it—which is itself the reason a cross-class ranking read off this metric cannot be trusted (Section 4).

Regime	Policy	MSE(Δp^{az})	MSE(Δp^{el})	MSE(Δz)
in-distribution	fixed-wide	0.135	0.566	0.009
	sweep	0.221	0.566	0.009
	BC (ours)	**0.004**	**0.003**	**0.001**
held-out	fixed-wide	0.186	0.384	0.009
	sweep	0.278	0.384	0.009
	BC (ours)	**0.176**	**0.137**	**0.003**

D FOV-RISK SWEEP (FULL)

Retraining with $\alpha \in \{0, 1, 2, 3, 4, 6, 8\}$ (five seeds each) lowers mean held-out high-risk error below flat BC by $\sim$19-50% (non-monotonic, small-sample); we use α=1 in the closed-loop matrix (Table 10).

Full sweep: aggregate and high-risk-bucket offline imitation MSE (summed over axes) on the held-out and in-distribution clips, for risk gain α ($\alpha = 0$ is flat BC), mean$_{\pm\text{std}}$ over five seeds per α (35 training runs). Every $\alpha \geq 1$ reduces held-out high-risk error; the reduction is not monotonic in α, and in-distribution error is flat within seed noise.

Table 7: Offline imitation MSE per axis, per clip (lower is better).

Clip	Policy	MSE(Δp^{az})	MSE(Δp^{el})	MSE(Δz)
`clip4_desk_dwell`	fixed-wide	0.000	0.633	0.010
	sweep	0.090	0.633	0.010
	BC (ours)	**0.001**	**0.001**	**0.001**
`clip6_mixed_natural`	fixed-wide	0.320	0.475	0.007
	sweep	0.400	0.475	0.007
	BC (ours)	**0.008**	**0.004**	**0.001**
`eval1`	fixed-wide	0.221	0.330	0.008
	sweep	0.311	0.330	0.008
	BC (ours)	**0.208**	**0.161**	**0.002**
`eval2`	fixed-wide	0.150	0.439	0.011
	sweep	0.244	0.439	0.011
	BC (ours)	**0.144**	**0.112**	**0.003**

	held-out (eval1/eval2)		in-distribution (val)	
α	all	high-risk	all	high-risk
0 (flat BC)	$0.278_{\pm.051}$	$0.138_{\pm.070}$	$0.0062_{\pm.0006}$	$0.0172_{\pm.0031}$
1	$\mathbf{0.222}_{\pm.037}$	$\mathbf{0.080}_{\pm.015}$	$0.0060_{\pm.0014}$	$\mathbf{0.0132}_{\pm.0023}$
2	$0.229_{\pm.032}$	$\mathbf{0.076}_{\pm.020}$	$0.0059_{\pm.0008}$	$0.0163_{\pm.0032}$
3	$0.257_{\pm.013}$	$0.112_{\pm.040}$	$0.0066_{\pm.0012}$	$0.0162_{\pm.0019}$
4	$0.236_{\pm.029}$	$0.081_{\pm.033}$	$0.0056_{\pm.0015}$	$0.0152_{\pm.0033}$
6	$0.288_{\pm.058}$	$0.079_{\pm.014}$	$0.0060_{\pm.0015}$	$0.0141_{\pm.0017}$
8	$0.235_{\pm.068}$	$\mathbf{0.069}_{\pm.031}$	$0.0057_{\pm.0014}$	$0.0167_{\pm.0048}$

E Implementation and Deployment Details

Camera and control interface. All experiments use a single Hikvision DS-2DE 25× PTZ camera. Frames are pulled over RTSP at 15 Hz, 1920×1080; camera commands are issued over the manufacturer's ISAPI HTTP interface, which the firmware rate-limits to $\leq$3 Hz. Pose and action use distinct normalisations, shared across the dataset builder, trainer and deploy-time wrapper. The measured pose enters the policy normalised with fixed divisors (pan 3600; tilt offset 150, divisor 1050; zoom offset 1, divisor 999). The policy output is a normalised increment $a_t \in [-1, 1]^3$, scaled by $S = (20°, 10°, 90 \text{ zoom steps})$ and added to the measured pose, $\mathbf{p}^{\text{cmd}}_{t+1} = \text{clip}(\mathbf{p}_t + Sa_t)$; this absolute target pose is sent over ISAPI, and a command is dropped if the previous one has not been acknowledged.

Deployment loop. At deploy time the policy is wrapped by the same rate-limited controller used to record the demonstrations. Three threads run concurrently: an RTSP frame thread pulls 15 Hz video; a pose-reader thread polls the camera at 10 Hz so the proprioceptive history is never more than 100 ms stale; and a control thread runs YOLOv8 + BoT-SORT detection, queries the policy, and writes commands at $\leq$ 3 Hz. DAgger rollouts reuse this loop: at every tick the center-lock expert is queried in parallel for the supervision label, and a per-tick coin flip with probability β chooses whether the expert or the BC policy issues the command. Decoupling labelling from execution means both always see the same observation, even when the BC policy is driving the camera. Because the expert merely holds pose when the target is lost, a single tracking failure can leave the camera fixated away from the scene; the controller therefore offers an optional recovery that returns the camera to the wide home preset after a sustained absence. This was set to 2.5 s in an early version of the study and is **disabled (0 s) for every closed-loop number reported here**: after the tilt-sign correction it degraded performance, since re-homing on a brief dropout moves the camera away from the actor's true position and prevents reacquisition, whereas coasting on the last commanded pose recovers far more reliably. When enabled, recovery ticks are lost frames with $\texttt{do}_t = 0$ and are masked out of the imitation loss, so recovery never contributes a spurious training target.

Optimisation. BC trains for 20 epochs of AdamW ($\eta = 3 \times 10^{-4}$, $\lambda_{wd} = 10^{-4}$), batch size 64, gradient clipping at 1.0, entropy bonus $\lambda_{\text{ent}} = 10^{-3}$. Each DAgger round fine-tunes for 5

epochs at $\eta = 10^{-4}$ on the aggregated dataset, with $\beta \in \{0.5, 0.25, 0.1\}$ over rounds 1-3. The FOV-risk weights are $(\lambda_e, \lambda_s, \lambda_\ell) = (0.6, 0.3, 0.1)$ with $h_{\text{safe}} = 0.25$, and the per-frame weight $w(s_t) = 1 + \alpha\, r(s_t)$ is renormalised per batch so $\alpha = 0$ recovers flat BC exactly. The policy has 11.27M trainable parameters and runs in 3.5 ms per inference (batch 1, single NVIDIA Quadro RTX 8000, mean over 200 calls), nearly two orders of magnitude under the 3 Hz control budget.

Evaluation. Offline imitation MSE replays each recorded frame through the policy in inference mode and compares $\tanh \mu_\theta(o_t)$ against the stored expert action, masked by $\texttt{do}_t$. Closed-loop HOTA (Luiten et al., 2021), IDF1 (Ristani et al., 2016), and MOTA (Bernardin & Stiefelhagen, 2008) are computed against target boxes seeded by a stronger offline detector (YOLOv8x re-run on the recorded live video) and then manually corrected frame by frame for all 30 runs; action rate and in-frame ratio are read from the controller's telemetry. The 2× learned-policy numbers are means across the five interleaved rounds.

F Wide-FOV Scenario-1 Pilot

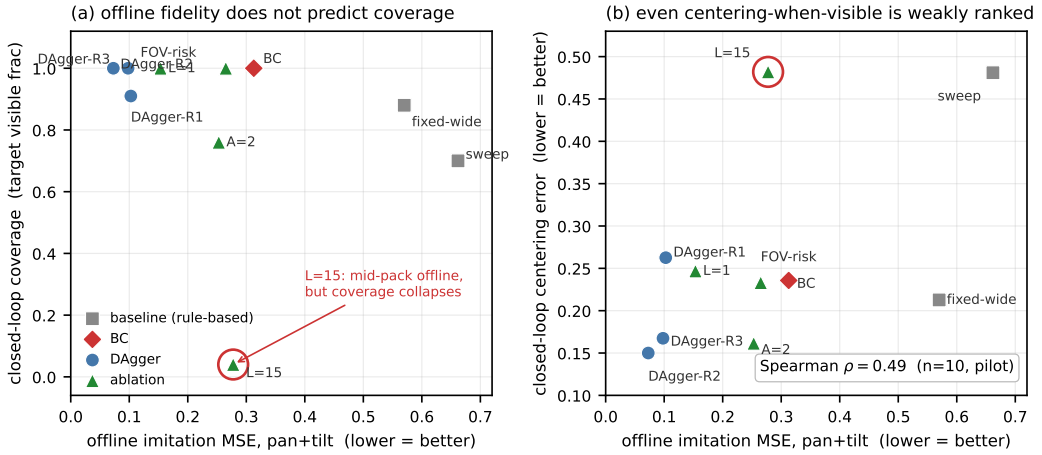

Figure 5: **Wide-FOV Scenario 1 pilot: one anecdotal same-checkpoint case of misranking.** Each point is one controller; the horizontal axis is the held-out offline imitation MSE of Table 2 (pan+tilt), the quantity a practitioner would screen on. *(a)* Closed-loop coverage (in-frame ratio) is flat across a ten-fold span of offline MSE and then collapses for L=15: the proxy ranks L=15 *above* flat BC (0.278 vs 0.313), yet live it holds the target on 4% of frames against BC's 100%. *(b)* Among frames where the target is visible, centring error is only weakly related to offline MSE (Spearman ρ=0.49, n=10, not significant) and is by construction blind to the coverage collapse that dominates deployment quality. Pilot data: one scenario, one interleaved round, ∼30-60 s rollouts. At the wide zoom IFR saturates for all but the outright failures, which motivates the narrow-FOV (2×) matrix of Table 10.

G Threats to Validity

H Seed Robustness of the Ranking Study

Every Table 2 policy was retrained with two further seeds under an identical recipe (seed 0 is the Table 2 checkpoint), and all 30 checkpoints were scored with the same pipeline (re-scoring seed 0 reproduces Table 2 within 10^{-4}). DAgger seeds re-run the R1→R2→R3 fine-tuning chain from the same-seed flat-BC checkpoint on the *fixed* collected data, so they capture optimisation variance only, not collection variance. The RGB+geometry value in Section 6 ($0.289_{\pm.034}$, three seeds paired with the geometry-only runs) and flat BC here ($0.325_{\pm.035}$: the Table 2 checkpoint plus two sweep reruns) are separate training runs of the same recipe. Re-evaluating the saved seed-0 checkpoints reproduces Table 2 within 10^{-4}; retraining from scratch with the same nominal seed is not bitwise reproducible (nondeterministic CUDA/cuDNN operations), and one such flat-BC rerun gives held-

Table 8: Threats to validity and how each is addressed. The ranking claim is deliberately scoped: it concerns the *unweighted* per-step proxy that standard practice screens on, is carried by the paired BC-versus-DAgger inversion and the OPED analysis rather than the aggregate correlation alone.

Threat	How we address it
Proxy favours BC by construction (its training objective, evaluated on logged in-distribution states)	Treated as part of the distribution-mismatch explanation, not rebutted: BC is optimised for logged-state mimicry, DAgger also trains on policy-induced states; RQ2 separates the final collected support from the warm-start optimisation trajectory (Figure 4, Section 5).
Cherry-picked metric—only plain MSE misranks	The inversion is specific to *unweighted* MSE ($\rho{=}-0.36$); a task-critical FOV-risk-weighted proxy gives $\rho{=}+0.18$ (Section 4); OPED restores $\rho{=}+0.79$ (Section 7).
Underpowered ranking ($n{=}10$, permutation $p{=}0.31$)	Primary claim is the large, directional, paired BC-vs-DAgger inversion; the aggregate ρ is reported as support with a bootstrap CI $[-0.79, +0.44]$; a post-hoc family$\times$regime interaction test gives exact permutation $p{=}1/120$ (Appendix J), and over three optimisation seeds the cross-family gap holds in $189/189$ held-out pairs while the within-family order does not replicate (Section 4, Appendix H).
Clip-choice sensitivity (the sign flips per clip)	Both clips reported; outright reversal appears on one of the two in-distribution clips (the near-stationary one), so the claim we rest on is the held-out gap between regimes, not a sign reversal (Section 4).
Closed-loop GT quality	The $2\times$ matrix (Table 10) is scored against manually corrected boxes seeded by an independent stronger detector; the FOV-risk>BC centring result uses controller telemetry, not GT, and holds in all five rounds.
Single camera / scene / off-the-shelf detector	Absolute numbers are rig-specific; all policies share the same detector/tracker, which reduces but does not eliminate perception confounds; cross-camera and detector transfer are not established.

out 0.243 rather than 0.313—still above the worst DAgger checkpoint (0.146). We therefore treat complete training runs, not nominal seed identifiers, as the unit of optimisation replication.

Table 9: Pan+tilt imitation MSE over three seeds (mean $\pm$ s.d.). Pooled over the 30 checkpoints, every DAgger checkpoint beats every BC-family checkpoint held-out ($189/189$ pairs) and is worse in-distribution in $178/189$; these checkpoint-level counts are descriptive, since checkpoints within a family share data.

Policy	in-dist. MSE	held-out MSE
$L{=}1$	$0.0041_{\pm.0005}$	$0.281_{\pm.162}$
FOV-risk $\alpha{=}3$	$0.0043_{\pm.0008}$	$0.228_{\pm.048}$
FOV-risk $\alpha{=}1$	$0.0054_{\pm.0018}$	$0.244_{\pm.044}$
$A{=}2$	$0.0051_{\pm.0011}$	$0.230_{\pm.045}$
FOV-risk $\alpha{=}2$	$0.0049_{\pm.0002}$	$0.225_{\pm.026}$
$L{=}15$	$0.0068_{\pm.0018}$	$0.299_{\pm.069}$
BC ($\alpha{=}0$)	$0.0058_{\pm.0011}$	$0.325_{\pm.035}$
DAgger-R1	$0.0110_{\pm.0034}$	$0.106_{\pm.038}$
DAgger-R2	$0.0111_{\pm.0084}$	$0.101_{\pm.030}$
DAgger-R3	$0.0143_{\pm.0073}$	$0.087_{\pm.009}$
Spearman ρ (ID vs. held-out), seeds 0/1/2: all ten $-0.36/-0.13/-0.62$; BC family $+0.86/+0.54/-0.21$.		
Over all 3^{10} one-seed-per-policy combinations: all-ten ρ mean -0.41 (negative in 99.9%); BC-family mean $+0.43$, range $[-0.75, +1.00]$.		

I PER-ROUND CLOSED-LOOP RESULTS

The four learned policies of Table 10 were retrained on $2\times$ demonstrations (8 chunks, 13,895 frames; chunks 7-8 held out as their $2\times$ validation set, on which all four reach $\approx 1.3 \times 10^{-3}$) with hold

Table 10: Narrow-FOV ($2\times$) closed-loop matrix of *retrained* hold-supervised policies (five interleaved rounds, mean). HOTA/IDF1/MOTA against manually corrected boxes; coverage = in-frame ratio; CErr = visible-frame centring error. Expert HOTA is dragged by one acquisition failure.

Policy	HOTA ↑	IDF1 ↑	MOTA ↑	Coverage ↑	CErr ↓
Non-learned reference					
fixed-wide	0.586	0.781	0.617	0.540	0.504
center-lock (expert)	0.662	0.811	0.737	0.640	0.358
Learned (retrained with hold supervision)					
+ FOV-risk (α=1)	0.790	0.942	0.891	0.79	**0.260**
DAgger-R2	0.769	0.931	0.872	0.76	0.262
flat BC	**0.805**	**0.954**	**0.912**	**0.87**	0.290
DAgger-R1	0.764	0.925	0.859	0.60	0.271

supervision λ_h=0.3. Scored on the wide-FOV clips of Table 2, which lie outside their training FOV, they reach pan+tilt MSE 0.41-0.58 in-distribution and 0.51-0.63 held-out (flat BC of Table 2: 0.0065 and 0.313 under the same pipeline), so no shared offline score links Table 10 to Table 2. OPED (Section 7) uses eight Table 2 checkpoints (flat BC, FOV-risk α=1, DAgger-R1-R3, L=1, L=15, A=2), each scored on its own states from two live wide-FOV rounds; because target-lost ticks are excluded, OPED does not capture failures to keep the target.

Table 11 gives the five interleaved rounds behind Table 10. HOTA uses the manually corrected target boxes; coverage and centring error come from the controller's telemetry and do not depend on GT. Round 1 of the expert is the acquisition failure noted in Table 10.

Table 11: Per-round results (rounds 1/2/3/4/5) of the narrow-FOV matrix.

Policy	HOTA	Coverage	CErr
fixed-wide	0.73 / 0.70 / 0.58 / 0.47 / 0.45	0.55 / 0.66 / 0.51 / 0.48 / 0.51	0.46 / 0.53 / 0.54 / 0.58 / 0.41
center-lock (expert)	0.17 / 0.85 / 0.78 / 0.73 / 0.77	0.02 / 0.95 / 0.86 / 0.48 / 0.88	0.80 / 0.23 / 0.28 / 0.23 / 0.25
FOV-risk (α=1)	0.76 / 0.87 / 0.80 / 0.72 / 0.80	0.81 / 0.95 / 0.90 / 0.54 / 0.78	0.25 / 0.26 / 0.27 / 0.27 / 0.25
DAgger-R2	0.72 / 0.76 / 0.85 / 0.79 / 0.73	0.60 / 0.90 / 0.91 / 0.82 / 0.59	0.28 / 0.27 / 0.25 / 0.28 / 0.24
flat BC	0.80 / 0.83 / 0.83 / 0.76 / 0.80	0.90 / 0.95 / 0.94 / 0.62 / 0.92	0.30 / 0.26 / 0.28 / 0.31 / 0.30
DAgger-R1	0.73 / 0.83 / 0.73 / 0.76 / 0.77	0.77 / 0.70 / 0.31 / 0.41 / 0.81	0.26 / 0.26 / 0.31 / 0.27 / 0.25

J Post-hoc Family×Regime Interaction

Relative to the BC family, the seed-0 DAgger rounds have $4.1\times$ higher in-distribution but $2.6\times$ lower held-out error (geometric means); of all $\binom{10}{3}$=120 three-policy "families", the DAgger split is the most extreme (exact permutation p=1/120). The in-distribution penalty is concentrated on the near-stationary clip ($17\times$; the families tie on the other), so outright reversal appears on one of two clips (per-clip ρ -0.37/$+0.44$), while the held-out gap holds on both. Scoring only top-tercile FOV-risk states moves ρ from -0.36 to $+0.18$, consistent with critical-interval weighting (Huang et al., 2026). This permutation treats the ten policies as exchangeable, which the three related DAgger rounds (one collection chain) do not strictly satisfy, so the p-value is exploratory.

K External QuadTrack Observation-Shift Check

Protocol. QuadTrack (Luo et al., 2025) provides panoramic ($360^\circ\times70^\circ$) sequences with persistent person tracks; only its training split carries ground truth. For each annotated person track we generate perspective 224×224 crops along a fixed schedule of six conditions (centred, four offsets, wide) that no policy influences, so changing the evaluated model never changes the crops. Labels and geometry use the *visible* (clipped) box; frames enter only if at least 70% of the box is visible. The expert and the four variants (RGB+geometry, geometry-only, shuffled-RGB, constant-RGB) use the rig's architecture and objective. Per (variant, seed, sequence) we compute do-masked pan+tilt MSE over at least 20 commanded frames, average sequences with equal weight per seed, and take the median over five seeds; uncertainty is a paired 10k sequence bootstrap. Primary: all frames with visibility $\geq$0.70. Gates, fixed before the run: at least 10 eligible sequences, then a confidence-interval width of R of at most 15 points, then the category (clear / modest / uncertain / null / reversal); failing the width gate classifies the result as imprecise.

Amendment and disclosure. The original fixed split left at most 7 held-out sequences because QuadTrack's test split has no ground truth, which failed the sequence gate. Before any model of the reported run was trained, we amended the split to sequence-level five-fold cross-fitting over all sequences with eligible frames (16; fold = sorted-id rank mod 5), leaving everything else unchanged. A preliminary run under the old split had been seen (fully-visible band only: geometry $0.036 <$ RGB $0.040 <$ shuffled 0.048); it is not used here.

Result. All gates up to precision pass (16 sequences; R=74%, 95% CI $[59, 80]$%; $\Delta = E_{\text{RGB}} - E_{\text{geom}} = +0.0063$ $[+0.0027, +0.0098]$; 5/5 seeds), but the R interval is 21 points wide, so the locked classification is *imprecise*. Table 12 reports all bands descriptively; secondary bands cannot change the primary classification.

Table 12: QuadTrack pan+tilt MSE by visibility band (median over seeds of equal-sequence means) and the paired effect $\Delta = E_{\text{RGB}} - E_{\text{geom}}$ [95% CI].

Band	N_{seq}	geometry	RGB+geometry	Δ [95% CI]
All $\geq$70% (primary)	16	0.0022	0.0085	$+0.0063$ $[+0.0027, +0.0098]$
Fully visible	15	0.0025	0.0099	$+0.0075$ $[+0.0034, +0.0124]$
70-80% visible	15	0.0019	0.0060	$+0.0041$ $[+0.0017, +0.0066]$
Near-edge, fully visible	15	0.0029	0.0113	$+0.0084$ $[+0.0038, +0.0138]$
Interior, fully visible	15	0.0020	0.0092	$+0.0072$ $[+0.0031, +0.0123]$

Controls (corrected). In the first run, the shuffled- and constant-RGB variants were trained on shuffled or constant images but, through an implementation defect, evaluated on real crops. The correction was documented before any corrected result was seen: at evaluation the constant model receives the same constant image and the shuffled model a cross-sequence permutation of held-out crops built by the training rule; only these two variants were retrained (5 folds $\times$ 5 seeds), with folds, seeds, epochs, bands and gates unchanged. The primary result is unchanged (R=74.2% $[58.7, 80.1]$, imprecise). On the primary band, shuffled images give 0.0029 and constant images 0.0033 (geometry-only 0.0022, RGB+geometry 0.0085); both remain below RGB+geometry in every band (paired Δ CIs exclude 0). As locked, these rows are descriptive: they are consistent with the RGB cost depending on state-aligned image content rather than on the image branch itself, but they do not identify a mechanism.

L Fixed-Update Support Control

The matched-epoch curve (Figure 4) gives larger mixtures more optimiser updates. We therefore also trained each mixture $D_0 \cup q\, D_{1:3}$ from scratch for a fixed $\approx$8,600 updates (epochs $41/32/27/23/20$ for $q = 0/0.25/0.5/0.75/1$ on $13{,}499/16{,}997/20{,}495/23{,}993/27{,}491$ training states at batch size 64, i.e. $8{,}651/8{,}512/8{,}667/8{,}625/8{,}600$ actual updates, within ± 1%; same objective and architecture, five seeds). Each q uses one nested subset that takes the same fraction of every collection round and is fixed across seeds, so seeds vary only initialisation and mini-batch order. The q=0 row is the baseline that trains longer on D_0 alone. Per-seed held-out values: q=0: .166/.173/.192/.201/.167; 0.25: .070/.038/.083/.055/.060; 0.5: .097/.050/.047/.063/.057; 0.75: .038/.040/.038/.048/.069; 1: .095/.033/.054/.249/.049.

Table 13: Fixed-update control: pan+tilt imitation MSE (median and mean$\pm$s.d. over five seeds; the q=1 mean includes one outlier seed at 0.249).

q	0	0.25	0.5	0.75	1
held-out	0.173	0.060	0.057	0.040	0.054
held-out mean$\pm$s.d.	$0.180_{\pm.014}$	$0.061_{\pm.015}$	$0.063_{\pm.018}$	$0.047_{\pm.012}$	$0.096_{\pm.079}$
in-distribution	0.003	0.003	0.003	0.003	0.003

Longer training on D_0 alone lowers held-out error (0.313 at 20 epochs to 0.173), but adding a quarter of the DAgger-collected states lowers it about three times further at the same update budget, with in-distribution error unchanged, so the support effect is not an artefact of the extra updates.

Aggregate-from-scratch seeds. The first five seeds give held-out $.050/.182/.051/.074/.051$; five further seeds, added later with the same recipe, give $.069/.055/.033/.059/.040$ (in-distribution me-

dian 0.003). Over ten seeds the median is 0.053 (IQR $[0.050, 0.069]$, mean 0.066 ± 0.040), inside the range of the first five.

M RQ3 OBSERVATION CONTROLS (EXPLORATORY)

These runs were added after the main RQ3 comparison and are exploratory; their reading rules were fixed before any result was viewed. Every variant uses the flat-BC recipe on D_0 (20 epochs, $4{,}220$ updates, best-by-validation checkpoint) and changes only the observation pathway: ***RGB*** (ImageNet-initialised ResNet-18, fine-tuned, 11.27M parameters); *geometry-only* (box+pose MLP, 28k); *geometry, parameter-matched* (three 2,288-wide layers, 11.20M); *frozen encoder* (ImageNet ResNet-18 frozen, 93k trainable); *strong augmentation* (colour jitter, greyscale, blur, random erasing on training images); *constant image* (mid-grey in training and evaluation); *shuffled image* (each state paired with another state's image from the same split, fixed permutation, in training and evaluation). The RGB and geometry-only runs here are a new replication, separate from the Section 6 runs; run-to-run nondeterminism (Appendix H) explains the different RGB level.

Table 14: Held-out (eval1/eval2) and in-distribution pan+tilt MSE of the observation controls (median; per-seed held-out values). Seeds are paired across variants.

variant	ID	held-out	held-out per seed
RGB (5 seeds)	.0048	.196	.196/.163/.183/.208/.254
geometry-only (10 seeds)	.0039	.018	.020/.030/.019/.014/.016/.016/.018/.023/.018/.017
geometry, param.-matched	.0022	.011	.011/.007/.014
frozen encoder	.0044	.041	.047/.041/.034
strong augmentation	.0082	.145	.145/.550/.101
constant image	.0049	.023	.023/.034/.019
shuffled image	.0064	.040	.040/.050/.033

RGB is worse than geometry-only in all five paired seeds (ratios 5.4-$15.6\times$, median $\sim 10\times$), smaller than the $\sim 17\times$ of the original three-seed comparison. Parameter count alone does not explain the gap: the parameter-matched geometry network is the most accurate variant. The RGB branch itself is not harmful when the image carries no state-specific information: constant images match geometry-only (1.0-$1.2\times$) and shuffled images cost only $\sim 2\times$. The large cost therefore appears only when real frames are paired with their states, consistent with the policy fitting appearance that co-varies with the action on the training actor; this does not by itself identify the mechanism. Freezing the ImageNet encoder recovers most of the gap (0.041, below RGB in 3/3 seeds) but stays $\sim 2\times$ above geometry-only; strong appearance augmentation does not reliably help (below RGB in 2/3 seeds, one seed at 0.55). All values are offline imitation error; none of these variants was deployed.

N SAME-CHECKPOINT LIVE TEST

Protocol. The ten Table 2 seed-0 checkpoints were deployed byte-identical (SHA-256 checked at every launch) in the current room with one scripted actor: five blocks of ten 75 s rollouts in pre-assigned Williams orders, initialised at $2\times$ (preset recalled and verified) with policy-controlled zoom, recovery disabled, zoom controlled by the policy. Endpoints, categories and the integration rule were locked before any data (the protocol file records SHA-256 hashes and timestamps of every addendum). A pre-specified feasibility gate (switch the whole study to $1.5\times$ if at least 8/10 policies had warm-up coverage below 0.2) read 3/10 on an unscored warm-up, so the study stayed at $2\times$. Primary: D_b = coverage(DAgger-R2) − coverage(L=1) per block, strong pass for mis-selection if $D_b > 0$ in 5/5 blocks and the paired bootstrap CI of the mean excludes 0. *Definitions* (all from controller telemetry at $\sim$10-14 control ticks/s, not ground truth): coverage is the fraction of ticks on which the on-board tracker detects the designated target with its box centre inside the frame; centring error e_t is the distance of the target-box centre from the image centre divided by the half image diagonal; $J_{0.25}$ is the fraction of *all* ticks with the target in frame and $e_t \leq 0.25$, so target-lost ticks count as 0 (tick-weighted, ε fixed before data); CErr averages e_t over ticks with the target detected. Rank agreement is $\rho_S(E_{\text{offline}}, 1 - M)$ for a metric M where higher is better, so positive values mean lower offline error goes with better live performance. Coverage is controller telemetry (target in frame), not ground truth.

Deviations, all before any scored run. Two warm-up attempts were aborted and archived (a switch to audible step cues, then a cue-timing fix); the audio cues deliver the unchanged timeline. Output moved to local disk for write speed, and the operator adjusted floor marks, after which the whole session restarted from the first warm-up run. All 50 scored rollouts were valid; no exogenous fault or replacement block occurred.

Post-hoc measurement check (Addendum I). After the result was known, the ten primary-pair rollouts were renamed V01-V10 by a seeded shuffle and every fourth control tick (2,321 frames, uniform over each rollout) was labelled: one box on the actor's visible extent, none if not visible. Boxes were seeded by an independent offline YOLOv8x (not the online tracker) and reviewed by one annotator (the operator), who changed 4 of 2,321 frames; this is a single-rater review of detector boxes, not full frame-by-frame ground truth. With these labels, coverage still favours $L{=}1$ in 5/5 blocks (-0.20, 95% CI $[-0.38, -0.05]$, bootstrap over the five paired blocks, not frames; $L{=}1$ is visible on every sampled frame), whereas the $J_{0.25}$ difference is not confirmed (-0.15 $[-0.35, 0.07]$, 1/5 blocks positive). Telemetry on the same frames reproduces the full-rate result (-0.27), and agrees with the labels on 87% of frames for target presence (Cohen's $\kappa{=}0.54$; undefined within $L{=}1$ runs, where presence never varies) and on 90% for the J indicator ($\kappa{=}0.78$); the online tracker under-counts visibility in some DAgger-R2 runs.

Results. Per-block D_b: $-0.56, -0.30, -0.14, -0.19, -0.14$ (mean -0.27, 95% CI $[-0.42, -0.15]$, sign-flip $p{=}0.063$), category *no mis-selection detected*. Secondary (never rescuing the primary): centred-in-frame $J_{0.25}$ pair difference -0.18 $[-0.28, -0.07]$, 0/5 blocks positive; visible-frame centring error favours DAgger-R2 by 0.027 $[-0.059, 0.004]$; DAgger-R2 beats FOV-risk $\alpha{=}3$ on coverage by 0.18 $[0.02, 0.35]$ (4/5). Over all ten policies, rank agreement with live coverage: Spearman $+0.15$ (in-distribution MSE), -0.14 (held-out); top-1 regret 0.03 vs. 0.30. With $J_{0.25}$: $+0.01$ vs. $+0.71$ (point estimates). Human-corrected HOTA for these runs is not computed.

Table 3 (Section 4) gives the per-policy means.

O Exploratory Zoom-Floor Blocks

Motivated by the zoom audit (after the Section 4 result was known), two further blocks ran the same ten checkpoints in Williams row B1 order with the same script, with every policy zoom command clamped to a floor: start and floor $2\times$ (Z2) and start and floor $3\times$ (Z1). Both designs were fixed and hashed before their data (protocol addenda H and J) as exploratory, descriptive-only analyses; each is a single block ($n{=}1$), the clamp changes the deployed actions, and neither can change the pre-specified result above. The floors held: measured zoom never fell below the floor (0% of ticks), and 27-681 commands per rollout were clamped ($A{=}2$, which has no zoom axis, none). Addendum J (the $2\times$ block) was written after the $3\times$ block had been seen. At a $3\times$ floor performance is low for every policy (coverage $\leq$0.41, $J_{0.25}\leq$0.19). At a $2\times$ floor this single block records the reverse pair order: the held-out choice DAgger-R2 exceeds $L{=}1$ on coverage (0.87 vs. 0.75) and $J_{0.25}$ (0.56 vs. 0.25), held-out error ranks the ten well (Spearman $+0.48$ on coverage, $+0.67$ on $J_{0.25}$; top-1 regret 0.10/0.01) and in-distribution error does not (-0.35/-0.59; regret 0.22/0.32). Block-to-block variation is large, however: $A{=}2$, whose zoom never changes, has coverage 0.34-0.89 across the five confirmatory blocks and 0.39 here. Replicated comparisons are needed to attribute any change to the zoom floor.

P Exploratory Waypoint Follow-up at a $2\times$ Zoom Floor

Designed after the confirmatory study and the zoom-floor blocks, and fixed and hashed before its data (protocol Addendum K, following an external handoff plan), one further session ran the same ten checkpoints at a $2\times$ floor (start $530/101/20$, zoom commands clamped to $\geq 2\times$, recovery off, 75 s) on two fixed actor routes: *A*, the confirmatory route, and *B*, a new route through four taped waypoints (near-left, far-right, far-left, near-right, with 3 s stops). Forty runs followed a pre-set order (two repeats per policy and route, the second in reverse order). All 40 slots are valid; one run was invalidated at the time because the actor walked the wrong route and was replaced (1 of 4 allowed). Measured zoom never fell below the floor, and $A{=}2$ (no zoom axis) was never clamped.

Table 15: Coverage and $J_{0.25}$ per policy: confirmatory study (initialised at 2×, free zoom; mean of five blocks) and the exploratory single blocks with a 2× and a 3× zoom floor.

	Free zoom (5 blocks)		Floor 2× (Z2)		Floor 3× (Z1)	
Policy	Cov.	$J_{0.25}$	Cov.	$J_{0.25}$	Cov.	$J_{0.25}$
flat BC	0.94	0.09	0.49	0.06	0.41	0.03
L=1	0.91	0.73	0.75	0.25	0.10	0.00
FOV-risk α=2	0.83	0.62	0.37	0.05	0.03	0.00
FOV-risk α=1	0.81	0.12	0.93	0.20	0.32	0.06
DAgger-R2	0.64	0.55	0.87	0.56	0.06	0.01
DAgger-R1	0.62	0.49	0.97	0.50	0.39	0.19
A=2	0.61	0.05	0.39	0.02	0.33	0.03
DAgger-R3	0.57	0.50	0.79	0.57	0.20	0.11
FOV-risk α=3	0.46	0.12	0.19	0.00	0.13	0.01
L=15	0.05	0.03	0.04	0.01	0.07	0.02

Waypoint distances were not measured. Metrics are the telemetry quantities of Appendix N; with two repeats per route the results are descriptive (no intervals). Both routes use the floor, so this session does not isolate the floor's effect.

Table 16: Waypoint follow-up at a 2× floor: primary pair per route and repeat (Δ=DAgger-R2−L=1; positive favours the held-out choice).

Route	Rep.	Cov. L=1	Cov. D-R2	ΔCov.	J L=1	J D-R2	ΔJ
A	R1	0.73	0.72	-0.02	0.17	0.44	+0.27
A	R2	0.79	0.77	-0.02	0.18	0.49	+0.31
B	R1	0.10	0.75	+0.65	0.01	0.53	+0.52
B	R2	0.07	0.09	+0.02	0.01	0.05	+0.04

Route (mean of 2)	ρ Cov. ID	ρ Cov. HO	ρ J ID	ρ J HO	J regret ID	J regret HO
A	-0.22	+0.44	-0.53	+0.68	0.35	0.06
B	-0.50	-0.01	-0.78	+0.45	0.38	0.10

On route A, coverage of the pair is nearly tied in both repeats (Δ= − 0.02), unlike the single 2×-floor block of Appendix O (+0.12), while $J_{0.25}$ favours the held-out choice in both repeats (+0.27, +0.31). On route B both differences favour the held-out choice in both repeats but with very different size (ΔCov. +0.65 and +0.02; in the second repeat DAgger-R2 lost the target for most of the run). Across the ten, held-out error ranks $J_{0.25}$ positively on both routes and in-distribution error negatively; for coverage the pattern holds on route A only. Under a 2× floor, in this session, the $J_{0.25}$ ordering thus favours the held-out proxy on two routes, while the size of the coverage difference varies strongly between repeats.

Q Post-hoc Sensitivity Analyses

Both analyses below were added after the results were known and use existing checkpoints and logs only.

Extended candidate set (offline). Adding the stronger policies of RQ2 and RQ3 to the ten Table 2 checkpoints changes the offline picture. We report two variants: a *recipe-level* one, where each added recipe is represented by its median or mean over seeds (mixed seed counts), and a *single-run* one, where each added recipe contributes its seed-0 checkpoint. The added recipes were not deployed in the live test, so this is sensitivity analysis, not prospective validation.

In both variants the aggregate correlation moves from negative to positive, and once a candidate is best on both distributions the two proxies agree on top-1: a direct instance of proxy validity depending on the candidate set.

Time weighting (live). Weighting each control tick by its duration instead of counting ticks leaves the live conclusions unchanged: $J_{0.25}$ gives Spearman +0.71 (held-out) and +0.01 (in-distribution) with the same best policy (L=1, held-out-choice regret 0.15), coverage gives −0.28 and +0.28, and the primary difference DAgger-R2−L=1 is negative in 5/5 blocks (mean −0.24).

Table 17: Offline ranking with an extended candidate set: Spearman between in-distribution and held-out error, each proxy's top-1 choice, and the held-out regret of the in-distribution choice (recipe-level; single-run seed-0 values in brackets).

Candidate set	ρ(ID, HO)	ID top-1	HO top-1	HO regret (ID)
Table 2 ten	-0.36	L=1	DAgger-R2	0.081
+ aggregate-from-scratch	-0.10 [-0.02]	L=1 [aggregate]	aggregate	0.100 [0]
+ geometry-only	$+0.08$ [$+0.13$]	L=1 [aggregate]	geometry-only	0.137 [0.030]
+ param.-matched geom.	$+0.27$ [$+0.32$]	param.-matched	param.-matched	0 [0]
+ frozen encoder	$+0.32$ [$+0.35$]	param.-matched	param.-matched	0 [0]

Centring threshold and leave-one-out (live). For the same-checkpoint test, held-out error ranks J_ε positively at every threshold ($\rho = +0.85, +0.78, +0.71, +0.59, +0.30$ for ε=$0.15, 0.20, 0.25, 0.30, 0.35$), weakening as centring is judged more loosely; dropping any one policy keeps it positive (minimum $+0.63$ at ε=0.25, $+0.15$ at 0.35). In-distribution error stays at $\rho \leq +0.01$, and its top-1 choice has J_ε regret $\leq$0.02 at every threshold, so the selection-versus-ranking distinction does not hinge on the pre-specified ε=0.25.