Supplementary Methods: Viewport selection for automatic StarCraft observation with learned scene interest and switching costs

Ho-Taek Joo and Jongwon Kim

State representation and interest regression

The primary controller combines supervised interest regression with stateful camera selection. Its supervised regressor receives game-state summaries and candidate coordinates. The previous camera enters the switching cost, not the learned feature vector. Table S1 summarizes this distinction.

Table S1: Components of the sequential decision formulation.

Component	Operational meaning
Game state G_t	Current nine-channel map; independent of camera action.
Optional history H_t	Maps 2 and 4 s earlier, not a recurrent memory.
Previous camera c_{t-1}	The controller's own preceding output; cost input only.
Action a_t	One of 840 absolute viewport positions.
Control reward	Human-view snapshot overlap minus a per-change cost.
Transition	Replay advances; camera memory becomes $c_t = a_t$.

Index i maps to viewport top-left coordinate $(4(i \bmod 28), 4\lfloor i/28 \rfloor)$. Initialization is index 434, $(56, 60)$. Exact score ties choose the smallest index, with no stay-first override. Earlier unavailable maps repeat the earliest available features. Local and global channel sums use $\log(1+n)/\log(241)$ and $\log(1+n)/\log(16385)$; these fixed area-based scales are not object-count upper bounds. Historical differences follow this transformation.

For candidate top-left coordinates $a = (x, y)$ in tiles, the coordinate features are $\boldsymbol{p}(a) = (x/108,\ y/116)$, normalizing each axis by its maximum valid top-left coordinate.

Table S2: Shared LightGBM regression settings.

Parameter	Value
Objective	Huber
Huber threshold (alpha)	0.9
Trees / learning rate	450 / 0.035
Leaves / maximum depth	31 / 7
Minimum child samples	80
Row fraction / resampling frequency	0.85 / 1
Feature fraction / ℓ_2 regularization	0.90 / 5.0
CPU threads	8
Deterministic / column-wise fitting	Enabled

No early stopping was used for these regressors. The implemented objective is Huber with threshold $\alpha = 0.9$. For an absolute residual of at most 0.9, the loss is proportional to squared

error; above that threshold it is linear. With interest labels in $[0, 1]$, we treat this as a shared regression setting, without attributing a separate robustness benefit to the loss. Other unspecified parameters retain library defaults. The reproduction environment used Python 3.12.7, LightGBM 4.6.0, NumPy 1.26.4 and pandas 2.2.2.

Match split and fitting order

The table below gives the test and validation assignments. Inner fitting excludes both; after validation selects the settings, refitting uses every non-test match. All 3 selector seeds and all 3 detector seeds share this split. No observer identity is a model input, and different matches need not contain the same five people.

Table S3: Outer test groups and inner validation matches. All other matches form inner training.

Round	Test matches	Validation match
1	275, 3529, 1725	36
2	1559, 36, 2351	438
3	11251, 438, 6219	522
4	4520, 522, 4664, 212	7950
5	1628, 7950, 1660, 9105	275

Each selector configuration was trained and evaluated with 3 random seeds: 20260907, 20260908 and 20260909. For seed s and fold f, sampling uses $s+100f$, the first regressor uses $s+10000+100f$, and the second regressor uses one greater. A NumPy SeedSequence combines the sampling seed and match identifier. Thus a shared training match receives the same action samples during inner and outer fitting.

Primary-controller scoring time

We measured the 20D primary controller on an AMD Ryzen 9 9950X3D CPU using eight threads, Windows 11, Python 3.12.7 and LightGBM 4.6.0. The five frozen outer-fold models from seed 20260907 scored their respective held-out matches. For each match, 50 evenly spaced cached decisions were evaluated three times after ten warm-up calls, giving 2,550 timed calls. Each call predicted all 840 candidate interest scores from a preloaded, contiguous float32 feature array; timing included the Python prediction interface.

The equal-match mean was 1.251 ms per 840-candidate call. The pooled median and 95th percentile of individual calls were 1.240 and 1.396 ms. These are single-decision scoring measurements, not batch time divided by the number of decisions. The mean is 0.063% of the 2-s decision interval. State extraction, feature construction, model loading, action selection and camera actuation were outside the timed region.

The accompanying README identifies the individual timings, per-match summaries and hardware records.

Reproducibility scope

The accompanying reproducibility package provides match-level records, predictions and analysis code for the reported tables. Its README contains the file index, method identifiers and SHA-256 checksums, and distinguishes summary reproduction from model inference and training. Summary statistics can be checked using the supplied metrics without original labels. Reconstructing features, human-view targets or frame-level overlaps requires the original game arrays and observer trajectories, which are not redistributed in this package. The source game-state arrays are available from the corresponding author on reasonable request.

Scene-interest targets measure human-view overlap; control reward additionally subtracts switching cost. The README maps the manuscript terminology to identifiers and fields in the supplied records.

The overlap tables use fixed-discount selector comparisons and the 3-seed detector mean. Match-level scores, paired-analysis inputs, component contrasts and candidate-error diagnostics are indexed in the README. Selector seed summaries appear under Fixed-discount repetitions and sampling diagnostics; detector repeats appear under Mask R-CNN comparator.

Selected value corrections, analysis-set eligibility and baseline motion are described under Fixed-budget sampling and value controls, Analysis-set eligibility and Baseline camera-motion statistics. The package includes 51 resampled baseline coordinate files (17 for detector seed 20260907 and 17 for each rule observer) and switching-cost trajectories and diagnostics; their locations are listed in the README.

Fixed-budget sampling and value controls

Rewards can be calculated for all 840 actions, but the matched comparison fits 81 unique actions per training state. Uniform sampling takes the first 81 indices of a random permutation. Target-aware sampling combines the stable top 12 under union interest, the stable top 12 under aligned interest, and up to five nearest-grid human-associated actions. Duplicate indices are removed and remaining positions are filled from the same permutation. Indices are sorted before fitting. These rows are fixed for a run, not resampled each tree or epoch. Reward and feature conditions share row draws and estimator seeds.

The future-only target is the backward dynamic-programming target minus current interest, centered over all actions before sampled rows are extracted. The discount is fixed at $\gamma = 0.5$. Both future and capacity corrections use

$$\beta \in \{0, 0.25, 0.5, 1\}, \qquad \delta \in \{0.025, 0.05, 0.1\}.$$

Counting the identical $\beta = 0$ policies once leaves ten settings for each correction. Inner validation maximizes mean cached snapshot interest minus the switching cost, including initialization. Stable grid order resolves equal objectives. Selected settings are refitted on all non-test matches.

The capacity control adds a second immediate regressor with the secondary seed. Its correction is the centered difference between the two immediate predictions, with the same mixing and clipping search. The future target uses recorded training futures; no future test map is supplied to either model. A separate all-action sensitivity fits all 840 actions with unchanged regression settings.

These comparisons form a retrospective analysis: method development reused the 17-match cohort, and the discount was fixed at $\gamma = 0.5$ after earlier results were examined.

Selected corrections. Across five folds and 3 seeds, validation selected $\beta = 0$ in five of 15 current-state fits and two of 15 history fits. Nonzero corrections were therefore used in 10 and 13 fits, respectively. Every selected future correction used $\delta = 0.025$. The README identifies the full fold–seed records, including the capacity control, within the package described under Reproducibility scope.

Table S4: Number of selected future-correction settings across 15 fold–seed fits per representation. Discount $\gamma = 0.5$ is fixed. At $\beta = 0$, the clipping value is immaterial; the stored tie-resolved value is 0.025.

β	δ	Current state (20D)	History (92D)
0	0.025	5	2
0.25	0.025	3	6
0.5	0.025	1	5
1	0.025	6	2

Absolute value-control scores. Table S5 retains the absolute scores for both added-estimator conditions alongside their immediate reference. The main-text sampling/history table uses the same reference, and the main paired-effects table reports all value contrasts and intervals.

Table S5: Value-correction and capacity-control overlap. All conditions use evaluation-aligned interest and 81 uniformly sampled training candidates per state. Scores average 3 seeds within each match and then weight all 17 matches equally.

Condition	Current state (20D)	History (92D)
Immediate reference	0.6257	0.6225
Additional current-interest regressor	0.6246	0.6234
Future correction ($\gamma = 0.5$)	0.6261	0.6225

Mask R-CNN comparator

At decision frame t, the detector concatenates nine-channel maps at $t - 24$, $t - 16$, $t - 8$ and t, producing 36 channels. Missing initial maps repeat the earliest available state. Targets are the five archived human masks, with bounding boxes formed from minimum and maximum occupied pixel coordinates, without adding one to the maximum.

The model is torchvision Mask R-CNN ResNet50-FPN initialized from public COCO weights. Its first convolution is replaced by a freshly initialized 36-input, 64-output layer with kernel 7, stride 2, padding 3 and no bias. COCO input weights are not tiled across channels. Box and mask heads are replaced by two-class predictors. Images are resized to 800 pixels; channel means and standard deviations are zero and one. SGD uses learning rate 0.005, momentum 0.9, weight decay 0.0005 and a tenfold learning-rate decrease every three epochs. Batch size is four, with BF16 mixed precision, two data workers and two CPU threads. The original detector run used PyTorch 2.7.1+cu128, torchvision 0.22.1+cu128 and NVIDIA RTX 4090 GPUs.

The detector comparison uses 3 base seeds: 20260907, 20260908 and 20260909, with fold seed $s + 100f$. The runs were independently trained under the same split and training settings. Among epochs 1–12, the highest validation overlap selects the duration; the earliest epoch resolves ties. The model is then reset and refitted on all non-test matches for that duration. For seed 20260907, selected epochs for folds 1–5 were 12, 9, 6, 6 and 8, with validation overlaps 0.6587, 0.6571, 0.7064, 0.5158 and 0.5314. These epoch and hardware records describe seed 20260907; the supplied package includes per-match scores for the other runs. Fold 1 selected epoch 12. An additional training run extended to 20 epochs, but its validation history is unavailable; the epoch selections reported here are those documented in the 1–12 search.

Inference takes the highest-confidence box, floors its top-left to integer tiles and clips x and y to $[0, 108]$ and $[0, 116]$. It is not snapped to the selector's 4-tile grid. The common evaluator uses a fixed 20×12 viewport. If no box is detected, the previous detector camera is retained, initialized at $(56, 60)$; the supplied records report zero fallback for all 3 seeds. Outputs use the

same 48-frame holding schedule as the selectors. Averaging the 3 scores within each match gives an overall detector mean of 0.5753. MO-VS-immediate minus this mean is +0.0504 [0.0267, 0.0791]. Run-specific results appear in Table S15. The detector motion measurements and spatial retention control use only the archived seed-20260907 coordinates; see Baseline camera-motion statistics and Retention matched to detector switching rate.

Numerical reproduction

With the fixed discount, immediate and capacity-control match means agreed with the reference outputs within 10^{-12}. Refitting the first fold under LightGBM 4.6.0 reproduced both input representations' saved predictions exactly. Inference from frozen checkpoints reproduced all 102 immediate and future trajectories at selection cost 0.05, including their frame-level mean overlaps. These comparisons check implementation consistency, not generalization independence.

Paired contrasts, multiplicity and metric interpretation

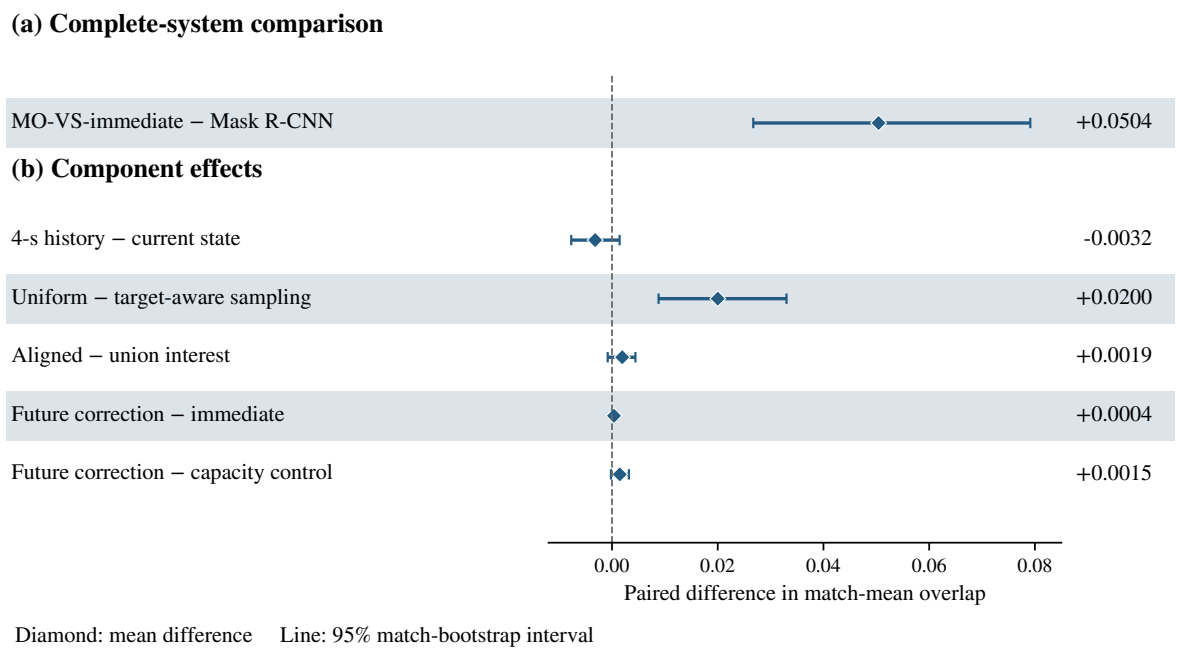

Figure S1: **Selected paired contrasts. (a)** MO-VS-immediate minus the three-seed Mask R-CNN comparator. **(b)** Selected component contrasts, including history versus current-state input. Diamonds show equal-match mean differences and horizontal lines show marginal 95% match-bootstrap intervals; positive differences favor the first condition. Both panels use the same horizontal scale. This figure displays selected contrasts, not the complete component family: all nine component comparisons remain in main Table 3, including the history and capacity-control contrasts.

The nine contrasts in main Table 3 form one exploratory family. A common matrix of 50,000 bootstrap samples of 17 seed-averaged matches uses NumPy seed 20260908. Marginal intervals use percentiles 2.5 and 97.5. The post-hoc Bonferroni-style sensitivity uses tail probability $0.05/(2\times 9)$, giving 99.444% individual intervals. It does not resample method-family selection, fitting or the dependence induced by shared training matches.

Table S6: Multiplicity sensitivity for all nine component contrasts. Intervals are individual 99.444% intervals, not marginal 95% intervals. Differences are computed from unrounded match means.

Contrast	Difference	Adjusted interval
History minus current state	−0.0032	[−0.0095, 0.0033]
20D: uniform minus target-aware	+0.0200	[0.0052, 0.0389]
20D: aligned minus union	+0.0019	[−0.0021, 0.0053]
20D: future minus immediate	+0.0004	[−0.0003, 0.0012]
20D: future minus capacity	+0.0015	[−0.0009, 0.0039]
92D: uniform minus target-aware	+0.0156	[0.0044, 0.0264]
92D: aligned minus union	−0.0008	[−0.0038, 0.0021]
92D: future minus immediate	0.0000	[−0.0005, 0.0004]
92D: future minus capacity	−0.0009	[−0.0039, 0.0017]

Only the two sampling contrasts have adjusted intervals entirely above zero. The result concerns this estimator, candidate budget and target-aware scheme, not uniform sampling in general. The detector comparison resamples paired differences after averaging the 3 training seeds within each match, using bootstrap seed 20260907 and lexicographic match-ID order; rule comparisons use seed 20260908. All-action and additional diagnostic intervals are marginal exploratory intervals outside the nine-comparison family.

For a pixel covered by m human viewports, the four-of-five aggregation has weights $w(0) = 0$, $w(1) = 0.8$ and $w(m) = 1$ for $m \geq 2$. Enumerating all 32 membership patterns verifies this identity. It is neither majority voting nor an estimated human-preference probability; counts above two are not distinguished.

Fixed-discount repetitions and sampling diagnostics

Three independently selected camera trajectories are retained for each condition and match. Seeds are averaged only for reporting, never ensembled during action selection. The per-seed results are:

Table S7: Per-seed match-mean overlap. Future correction fixes the discount at 0.5.

Seed	20D immed.	20D future	All 840	92D immed.	92D future
20260907	0.6267	0.6272	0.6302	0.6232	0.6230
20260908	0.6249	0.6257	0.6297	0.6240	0.6243
20260909	0.6256	0.6254	0.6263	0.6204	0.6203

Candidate error is computed over held-out decision–action pairs. Interest bins are zero, $(0, 0.25]$, $(0.25, 0.75]$ and $(0.75, 1]$. Training-bin shares pool sampled rows across folds and seeds; diagnostic intervals instead resample seed-averaged matches. The high-interest share is 7.54% for target-aware and 0.73% for uniform sampling. These are descriptive shares, not independent inferential samples.

Table S8: Errors across all held-out candidates and own-camera regret. The last column is MAE for interest scores above 0.75.

Condition	MAE	Bias	Regret	High MAE
20D target-aware	0.0640	0.0348	0.2873	0.2390
20D uniform 81	0.0477	0.0010	0.2698	0.4213
20D all 840	0.0478	0.0011	0.2664	0.4232
92D target-aware	0.0626	0.0342	0.2783	0.2283
92D uniform 81	0.0465	0.0008	0.2647	0.4077

For previous camera c, exact immediate control regret is

$$L_t(c) = \max_a \{\widetilde{C}_t(a) - 0.05\mathbb{I}[a \neq c]\} - \{\widetilde{C}_t(\hat{a}_t(c)) - 0.05\mathbb{I}[\hat{a}_t(c) \neq c]\}.$$

The own-camera diagnostic conditions on each controller's own previous camera. Uniform-minus-target-aware regret is -0.0175 [-0.0293,-0.0074]. The common-camera diagnostic supplies the uniform controller's previous coordinate to both scoring models at each snapshot, without rolling a counterfactual second camera forward. Mean regrets are 0.2698 and 0.2819, respectively; the difference is -0.0122 [-0.0179,-0.0072]. Both use cached training-coordinate snapshot labels on held-out matches, not the frame-level primary endpoint. Uniform sampling reduces selected-action regret while increasing high-interest candidate MAE, so it does not improve every type of candidate prediction.

The all-action sensitivity trained the current-state model on all 840 candidates with unchanged regression settings and seeds. Overlap was 0.6287 versus 0.6257 with 81 uniform samples, a paired difference of $+0.0030$ [-0.0019, 0.0081]. More than ten times as many training rows produced a small observed increase. The README identifies the supplied match-level scores.

Illustrative camera decisions

Main Figure 4 uses the first seed (20260907) and the immediate controller at cost 0.05. Eligible decisions occur after 180 s, have a strictly positive highest-interest-alternative minus previous-camera margin, and a unique highest adjusted score. Across all held-out matches, decisions are separated into stay (margin below 0.05) and switch (margin above 0.05) classes. The median margin case in each class is selected post hoc; numeric match and frame break ordering ties. This selection is explanatory rather than an unbiased sample of viewing quality.

The stay example is match 1559 at frame 34752 (1448 s): previous viewport top-left $(104, 52)$ has predicted interest 0.636948, whereas the best alternative $(100, 56)$ has predicted interest 0.653797 and post-cost score 0.603797. The switch example is match 4520 at frame 12912 (538 s): previous $(96, 84)$ has predicted interest 0.304651, whereas alternative $(52, 12)$ has predicted interest 0.462112 and post-cost score 0.412112. The point maps are occupied tiles from the corresponding archived nine-channel state; human rectangles use the cached training-mask geometry. The figure is regenerated from the saved case manifest and derived tile locations.

Selection-cost, time and coordinate audits

The two-cost contrast below complements Switching-cost sensitivity and control diagnostics, which evaluates six selection costs with the same frozen models. The switching cost of 0.05 is fixed as the common operating point. The comparison below changes the inference-time cost with the models fixed; it does not select an optimal cost. Both conditions in the fixed selection-cost set $\{0, 0.05\}$ are reported. Frozen 20D interest models, future models and validation-selected mixing/clipping parameters are used. The future models retain training cost 0.05 and discount 0.5; only the cost at final action selection changes. The comparison contains $17 \times 3 \times 2 \times 2 = 204$ closed-loop trajectories. No model was refitted or primary method reselected.

For each trajectory, switches count any absolute-coordinate change after initialization. Duration uses the common evaluated frame count and a nominal rate of 24 frames per second. Median dwell is taken over observed constant-coordinate runs, including first and last runs and truncating at the evaluation endpoint. Jump length is the mean Euclidean distance only among changed coordinates. Each statistic is averaged over seeds within a match, then equally across matches. A common 50,000-draw match-bootstrap matrix uses seed 20260909.

Let E_t be evaluator-coordinate overlap at decision snapshot t, and T the number of decisions

inside common evaluation support. The fixed-reference utility is

$$U_{0.05} = \frac{1}{T}\sum_{t=1}^{T}\left[E_t(a_t) - 0.05\mathbb{I}(a_t \neq c_{t-1})\right].$$

Initialization is charged in this utility. The reference cost remains 0.05 even for a trajectory selected with zero cost. Thus the units are reward per decision, not overlap per frame minus switches per minute.

Table S9: Paired selection-cost 0.05 minus 0 differences with marginal 95% match intervals.

Controller	Metric	Difference	95% interval
immediate	Frame overlap	−0.0047	[−0.0078, −0.0016]
immediate	Changes/min	−6.5270	[−7.0375, −6.0123]
immediate	Median dwell (s)	+1.0441	[0.5490, 1.5539]
immediate	Conditional jump (tiles)	+3.2670	[2.3894, 4.1210]
immediate	$U_{0.05}$	+0.0084	[0.0050, 0.0117]
future	Frame overlap	−0.0047	[−0.0077, −0.0017]
future	Changes/min	−6.4872	[−7.0057, −5.9608]
future	Median dwell (s)	+1.0637	[0.6078, 1.5147]
future	Conditional jump (tiles)	+3.2655	[2.3023, 4.1756]
future	$U_{0.05}$	+0.0084	[0.0051, 0.0116]

The primary overlap falls slightly as cost increases, whereas switching and the common decision-scale utility improve. Conditional jump lengths increase: the cost suppresses changes but does not penalize longer moves more heavily. Median dwell and jump length do not establish viewer comfort or moving-event tracking quality.

Across all cached decisions, exact final-score ties occur in 1.30% and 0.49% of immediate decisions at costs 0 and 0.05, respectively, and 0.75% and 0.29% of future decisions. These are seed-within-match then equal-match averages, conditioned on each policy's own previous camera. The smallest-index rule applies to every exact tie.

Time and coordinate conventions

Temporal support. Rule trajectories are sampled at the common 48-frame decision times and held, rather than evaluated at their native cadence. Frame-level evaluation and snapshot-based training labels use different temporal aggregations.

All 17 caches start at frame zero with spacing 48. Training labels describe the corresponding snapshot. Evaluation uses every integer frame before the shortest final human timestamp, excluding that timestamp. Missing intermediate observations are forward-filled. Policy outputs hold until the next decision or the last available evaluation frame. Cache decisions after the common endpoint are ignored by frame evaluation. In two matches (1559 and 1660), the final available camera holds for 89 and 60 frames; the actual evaluated duration is used. The README identifies the full per-match support records.

Coordinate conventions. The original mask-preparation function divides positive raw pixel coordinates by the tile size 32 and truncates to integers. At all common cached snapshots, these coordinates reproduce the archived human-mask sets exactly. The source evaluator instead uses horizontal and vertical pixel constants 3456 and 3720, normalizes to the valid viewport range, and rounds. The replicated evaluator preserves this conversion. Matched human coordinates differ by at most one tile per axis. Snapping to a nearest action index is a third operation used only by target-aware sampling; it does not alter either mask convention. The source constants are reproduced, not replaced by geometric estimates. The README identifies the source functions, their replica and match-level coordinate comparisons.

The raw-state scatter routine counts co-located objects, so the first eight channels need not be binary. The separate vision parser supplies the ninth tile-wise vision channel. Interest-target alignment in this study matches observer aggregation, not coordinate quantization or averaging over the complete holding interval. These comparisons use frozen models and unmodified source labels.

Exact continuation and action-score gaps

A seven-step, six-action brute-force recursion was compared with the efficient maximum implementation for discounts 0, 0.5 and 0.9 and costs 0 and 0.05. Maximum absolute disagreement was 4.77×10^{-7} (tolerance 5×10^{-7}). The zero-cost centered future target vanishes within that tolerance, the continuation range respects $\gamma\lambda$, and setting $\beta = 0$ reproduces immediate actions exactly.

At cost 0.05 and discount 0.5, the immediate control-score gap between the best two candidates is strictly below 0.025 in 37.71% of decisions within frame-evaluation support, above it in 62.29%, and never exactly equal. The reference is the uniform current-state controller's own previous camera, initialized at viewport top-left $(56, 60)$. Scores use all 840 cached interest targets, not fitted predictions, with a 0.05 penalty for leaving that reference camera. The cached labels are binary16 values; integer arithmetic on a scale of 25×2^{24} represents both those labels and the decimal cost/bound exactly. The same calculation over all cached decisions, including decisions beyond the common human-trajectory endpoint, gives 37.72% below the bound. Results average the 3 seeds within each match and weight matches equally. A gap strictly above the bound rules out reranking by exact continuation at that reference state; it is not a bound on fitted-policy differences.

Descriptive score references

These post-hoc references place the primary overlap scale in context without changing any model, sampling budget or selected setting. All match summaries use the same 500,997 evaluation frames and equal match weights.

For the random reference, an action is drawn uniformly from the existing 840 candidates at each existing decision and held to the next decision, including the evaluator's final-boundary convention. We report its exact expected overlap, not one random trajectory. If $E_f(a)$ denotes frame-level evaluator overlap, linearity gives

$$\mathbb{E}[E_f(A)] = \frac{1}{840} \sum_{a \in \mathcal{A}} E_f(a).$$

Holding affects correlations between frames but not this expectation. The calculation integrates the probability that a uniformly drawn viewport covers each pixel against the evaluator's 0, 0.8 and 1 attention weights. An independent direct evaluation of all 840 rectangles at five deterministic frames in each match agreed to within 10^{-12} (85 checked frames). No training interest targets or model predictions were used. The fixed-center reference retains the controller's original initial viewport top-left $(56, 60)$ throughout; this position was not selected using evaluation scores.

For human agreement, each human viewport is compared with the union of the other four:

$$H_f = \frac{1}{5} \sum_{j=1}^{5} \frac{\left| \mathrm{VH}_{f,j} \cap \bigcup_{i \neq j} \mathrm{VH}_{f,i} \right|}{|\mathrm{VH}_{f,j}|}.$$

The scored observer never appears in its own reference union. We retain the original recorded human trajectories, so they need not lie on the action grid or change only at the controller's decision times. This is a descriptive agreement reference, not a same-constraint controller score or an upper bound.

Table S10: Contextual overlap references, averaged equally across matches. Human agreement is separated because it evaluates a human against only the other four, under its original recorded timing.

Reference	Mean score
Uniform random candidate (exact expectation)	0.0425
Fixed initial central viewport	0.0274
Human versus other-four agreement	0.5146

A selector can score above the human reference by choosing areas common to several observers, even when each human chooses a different valid scene. The metric alone does not establish superhuman observation, editorial quality or viewer preference. Per-match scores and the exact-expectation check are included in the reference-score audit.

Public source-data inventory

The snapshot of the public source-data repository is identified in the accompanying README. It contains five human files for 14 of the 17 analyzed matches, replay files for 11 analyzed matches, and no preprocessed state arrays. Human trajectories for matches 3529, 7950 and 9105 are absent from that snapshot. On the common evaluation frames, all 70 available human trajectories reproduce the corresponding local raw-coordinate sequences after sorting, resolving duplicate timestamps and forward filling. The comparisons allow a permutation of observer file names within each match; they do not claim byte-identical files or equality outside the common evaluated support.

The README provides the repository URL, full commit identifier and locations of coordinate-comparison records. The snapshot contains no license file, preprocessed state arrays or human trajectories for the three matches listed above. Raw human files were read for correspondence checks and are not redistributed in the reproduction package.

Analysis-set eligibility

The supplied source archive contains 21 match directories with paired NumPy arrays. The analysis requires a $9 \times 128 \times 128$ state and a $5 \times 128 \times 128$ human-mask label at each cached decision, together with five evaluable human trajectories. Seventeen matches satisfy the cache input requirements and have linked trajectories. The initial arrays of matches 7191, 7970, 9301 and 9795 instead contain a two-element label of `None` values, so those directories are excluded. These criteria describe the input requirements of the existing 17-match cohort, not a new performance-based selection. The README identifies the directory counts, initial-array shapes and selected match list. The preparation script checks the required shapes at all decisions used to construct the retained caches.

Baseline camera-motion statistics

The frozen baseline trajectories were placed on the same cached 48-frame decision schedule and held until the next decision or the end of the common human-trajectory support. Mask R-CNN coordinates retain their integer-tile resolution; they are not snapped to the selector's 4-tile candidate grid. A switch is any nonzero change in coordinates after initialization. Dwell includes the first and last constant-coordinate runs, clipped to the evaluated duration, and jump length is the Euclidean distance conditional on a switch. Trajectory medians or means are calculated first,

then given equal match weight. The MO-VS row additionally averages its 3 seeds within each match, as in the main analysis.

The rule-observer comparisons reevaluate stored source trajectories; they are not new executions of the observer programs. All 17 SSCAIT trajectories and 14 AIIDE trajectories come from the source VPD files. The other three AIIDE trajectories come from the source's saved-coordinate records. The README maps these sources to their archive locations. Repository commits identify inspected source snapshots but do not establish which program version generated these archived runs.

Table S11: Descriptive overlap and camera motion on the common decision schedule. Dwell is the match-mean of trajectory median durations; jump is the mean conditional movement distance. Detector overlap and motion in this diagnostic both use archived seed 20260907; the main overlap comparison instead averages 3 seeds. Rule trajectories are fixed.

Method	Overlap	Changes/min	Dwell (s)	Jump (tiles)
MO-VS-immediate	0.6257	9.63	3.04	22.43
Mask R-CNN	0.6078	20.98	2.00	25.16
SSCAIT	0.4943	19.71	2.00	23.18
AIIDE	0.4923	22.48	2.00	45.72

For all 51 baseline match trajectories, recalculated overlap agrees with the original run-level results to within 2.3×10^{-16}. The statistics describe coordinate motion under a shared output cadence: a one-tile detector adjustment and a large scene change each count as one switch. Per-match motion summaries and decision coordinates are included in the package described under Reproducibility scope. The fixed-model cost comparison in the main text isolates the effect of the switching penalty, whereas this table compares complete systems.

Switching-cost sensitivity and control diagnostics

These retrospective fixed-model sensitivity and exploratory robustness analyses keep the existing 17 matches, folds, 3 model seeds and primary operating cost unchanged. The candidate costs $0, 0.01, 0.02, 0.05, 0.10, 0.20$ were fixed before running the analysis. Each immediate and future model is reused without retraining for the cost sweep. Future targets were fitted with $\lambda = 0.05$, $\gamma = 0.5$; selected mixing/clipping values remain frozen. The sweep changes only the selection rule and does not select a replacement primary cost. Recomputed cost-zero and cost-0.05 overlaps match the reference outputs within 10^{-12}.

Motion and aggregation. Predictions retain the original 48-frame decision schedule and common human endpoint. The initial transition is excluded from motion counts; first and last dwell runs are included and clipped to the common support. A jump is the Euclidean distance between different successive viewport top-left coordinates. Seed results are averaged within matches, then matches receive equal weight. Human statistics first average five observers within a match. Native human coordinates use the integer-frame evaluator; held-human coordinates sample these at the existing controller decisions and hold until the next. Their self-excluded agreement is 0.5146 natively and 0.5055 after holding.

Table S12: Full fixed-model sweep. Jump lengths condition on a coordinate change.

Controller	Cost	Overlap	Changes/min	Dwell (s)	Jump (tiles)
Immediate	0.00	0.6304	16.15	2.00	19.16
Immediate	0.01	0.6304	13.58	2.16	20.64
Immediate	0.02	0.6294	12.16	2.41	21.23
Immediate	0.05	0.6257	9.63	3.04	22.43
Immediate	0.10	0.6150	7.38	4.16	23.86
Immediate	0.20	0.5887	4.81	5.73	25.96
Future	0.00	0.6308	16.13	2.00	19.22
Future	0.01	0.6308	13.60	2.12	20.60
Future	0.02	0.6297	12.17	2.37	21.36
Future	0.05	0.6261	9.65	3.06	22.49
Future	0.10	0.6151	7.39	4.22	23.84
Future	0.20	0.5887	4.81	5.71	25.92

Table S13: Human motion under two output cadences; mean of observer summaries within matches, then equal match weighting.

Cadence	Changes/min	Dwell (s)	Jump (tiles)
Native integer frame	111.36	0.15	9.27
48-frame hold	20.58	2.00	34.14

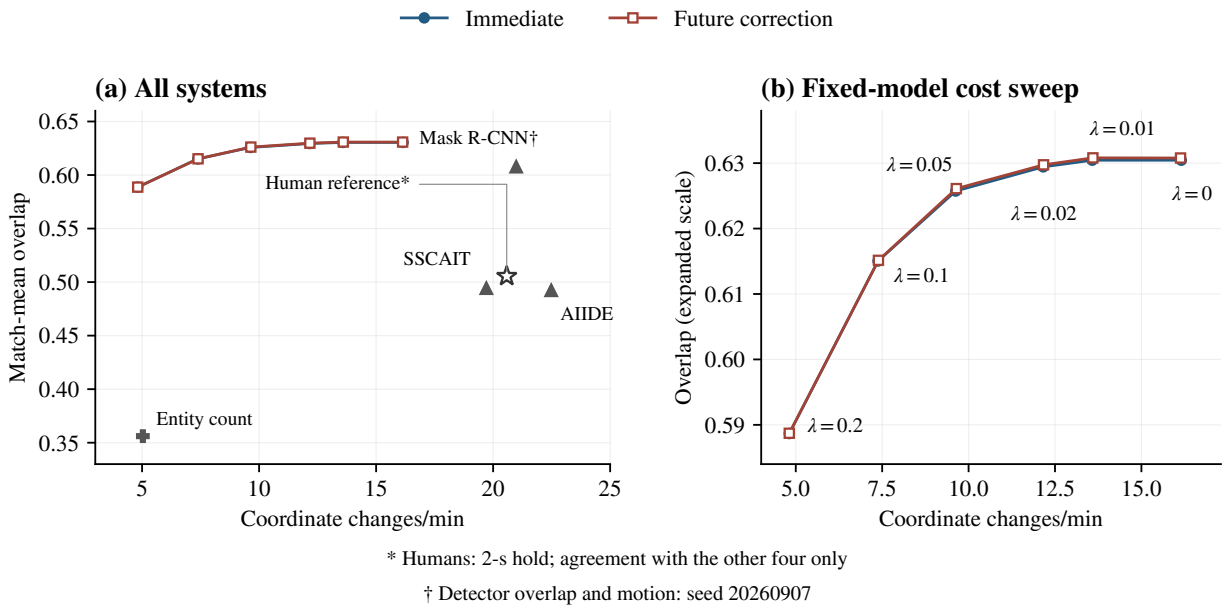

Figure S2: **Overlap and camera-motion references.** (a) Complete-system and descriptive-reference points alongside the fixed-model cost sweep; (b) a closer view of the controller curves. Changes/min counts every nonzero change in viewport coordinates after initialization, the same measure used in main Figure 3. Lines connect the six evaluated selection costs, not a fitted or optimized frontier. The detector point combines overlap 0.6078 and motion from the same archived seed-20260907 run; it is separate from the three-seed detector overlap 0.5753 in main Table 1. Rule-observer outputs use the common 48-frame hold. The human reference also uses that hold and compares each observer with the other four; its overlap 0.5055 differs from native-cadence agreement 0.5146. It is a descriptive agreement reference, not an upper bound or a direct ranking against controllers. Main Figure 3(a) isolates the fixed-model switching-cost curves, whereas Figure 3(b) presents the separate distance-sensitive cost analysis.

Coordinate-removal and heuristic controls. The coordinate-removal model drops feature columns 18 and 19 from the 20D current representation. All 15 fold–seed fits use the same uniform action indices, seed offsets, 450 regression trees, learning rate 0.035 and other regressor settings as the reference model. It is refitted on all non-test matches using those fixed settings. Its 0.6182 mean versus 0.6257 gives removal minus reference −0.0076 [−0.0146, −0.0014]. Differences here and in the retention comparison use unrounded match means.

The entity-count heuristic is an untrained current-state candidate scorer. At each existing 48-frame decision it uses the eight local player-object features for every one of the 840 viewport candidates, reverses their normalization as $n_k(a) = \exp(\ell_k(a) \log 241) - 1$, and chooses the candidate maximizing $\sum_{k=1}^{8} n_k(a)$. It uses no global counts, history, neutral-vision channel, fitted parameters or human-view targets. Archived float16 features introduce numerical quantization into the recovered counts. Exact ties choose the smallest action index, as in the primary selector, and the chosen viewport is held to the next decision under the same evaluation convention. No switching penalty is applied to the count scores. The primary-minus-heuristic overlap difference is +0.2696 [0.2382, 0.2988].

The entity-count heuristic achieved overlap 0.3561, 5.03 changes/min and mean conditional jump distance 51.41 tiles. Removing the primary immediate controller's switching cost gave overlap 0.6304 and a paired difference of +0.0551 [0.0325, 0.0825] from the 3-seed Mask R-CNN mean. The README identifies the inputs and calculations for the latter contrast.

Exact ties. For each frozen model at cost 0.05, the tie fraction counts decisions with multiple exactly equal highest adjusted scores, including the first evaluated decision. Primary frequencies are 0.4964% (immediate) and 0.2921% (future); their cost-zero frequencies are 1.2999% and 0.7539%. Three uniform random tie-resolution trajectories are generated per model seed, with a separate NumPy SeedSequence for each model seed, match, repeat and controller; the fixed seed namespace is documented in the README. Each trajectory updates its own previous camera. Repeats average within seed before seed and match aggregation. Random-minus-deterministic overlap differences are −0.000084 [−0.000241, 0.000055] and +0.000059 [−0.000049, 0.000169], respectively.

Paired intervals in this section use a common 50,000-draw match-bootstrap matrix generated with seed 20260923. They are exploratory marginal intervals and do not extend the original nine-comparison family. Full-precision summaries, random-repeat trajectories, 15 coordinate-removal models and model hashes are included in the reproducibility package.

Retention matched to detector switching rate

This post-hoc diagnostic uses the same 17 matches and the completed detector's highest-confidence coordinates. At each 48-frame decision, let p_t be that proposal and c_{t-1} the retained viewport top-left. Starting at $(56, 60)$, we choose $c_t = p_t$ if $\|p_t - c_{t-1}\|_2 > r$ and otherwise retain c_{t-1}. No detector is refitted, no proposal is snapped to the action grid, and no human coordinate is used in the update. Radius zero reproduces the original detector trajectories exactly.

Before evaluating overlap, radii $r \in \{0, 1, \ldots, 160\}$ tiles are compared by their equal-match mean switching rates. We choose the radius minimizing the absolute difference from the frozen MO-VS-immediate reference rate, with the smaller radius resolving ties. Radius 5 yields 9.694 changes/min versus the target 9.627; the absolute mismatch is 0.067 changes/min. This selects a spatial retention margin from motion alone, without calibrating detection confidence as an overlap prediction. The operating point is matched on the same cohort and is not a held-out validation choice.

Retention-minus-original overlap is −0.0145 [−0.0175, −0.0116]; MO-VS-immediate minus the retained detector is +0.0325 [0.0095, 0.0607]. These exploratory percentile intervals use 50,000 match resamples with seed 20260927 and are conditional on the motion-selected radius; the radius

Table S14: Motion-matched retention on archived detector proposals. Overlap and motion use the common frame support. MO-VS averages 3 seeds within each match; both detector rows use the same archived seed-20260907 outputs. Dwell is the mean of within-trajectory medians; jump distance is conditional on a coordinate change. Differences are computed from unrounded match means.

Controller	Overlap	Changes/min	Dwell (s)	Jump (tiles)
Mask R-CNN	0.6078	20.98	2.00	25.16
Mask R-CNN + retention	0.5933	9.69	3.18	53.02
MO-VS-immediate	0.6257	9.63	3.04	22.43

is not reselected in each resample. These motion-based comparisons are separate from the 3-seed primary overlap comparison and do not estimate variability in motion across detector seeds. The rise in conditional jump distance shows that equal switching frequency does not imply equivalent movement.

The supplementary data provide the protocol, motion-only radius sweep, selection record frozen before overlap evaluation, per-match scores, paired summaries and output coordinates. This diagnostic evaluates spatial retention of detector proposals, not confidence-based rescoring or a trained tracking system.

Sensitivity to matches and training runs

To assess whether particular matches or training runs drive the mean difference, we repeat score aggregation after omitting each match, test fold or detector run in turn. The primary comparison averages the three selector runs and three detector runs within each match and favors the selector in 16 of 17 matches. The mean differences remain positive: +0.0401 to +0.0550 after omitting one match, +0.0315 to +0.0621 after omitting one test fold, and +0.0413 to +0.0667 after omitting one detector run. These sensitivity checks use the frozen scores without retraining.

Table S15: Sensitivity to each completed detector run. Differences compare the three-run selector mean with one detector run, paired by match. Intervals are exploratory marginal 95% match-bootstrap intervals. Differences are computed from unrounded match means.

Detector seed	Overlap	Difference	95% interval
20260907	0.6078	+0.0179	[−0.0058, 0.0476]
20260908	0.5610	+0.0647	[0.0408, 0.0930]
20260909	0.5571	+0.0686	[0.0437, 0.0987]

Mean overlap differences favor the selector in all three detector-run comparisons, while the interval for seed 20260907 includes zero (Table S15). Across the nine selector–detector run pairs, differences range from +0.0171 to +0.0696. Selector run means range from 0.6249 to 0.6267. These run-pair comparisons share trained models. The intervals use 50,000 paired match resamples with seed 20261004, averaging selector runs within each match before resampling.

An exhaustive two-sided sign-flip check enumerates all 2^{17} sign assignments to each frozen match-difference vector. Holm correction is applied only to the three primary baseline contrasts: Mask R-CNN, SSCAIT and AIIDE. The adjusted values are $p = 0.0001831$ for Mask R-CNN and $p = 0.00004578$ for each rule observer. These checks are conditional on exchangeability under paired sign reversal; they do not resample shared training fits or repeated development decisions. Their three-contrast correction does not cover the individual-run diagnostics or replace the separate nine-component multiplicity analysis.

Distance-sensitive switching cost

To distinguish the frequency of camera changes from their spatial extent, we retain the frozen 20D immediate-interest predictor and select

$$a_t = \arg\max_{a \in \mathcal{A}} \left\{ \widehat{C}_t(a) - 0.05\mathbb{I}[a \neq c_{t-1}] - \eta \frac{\|a - c_{t-1}\|_2}{D} \right\}, \qquad D = \sqrt{108^2 + 116^2},$$

where a and c_{t-1} are viewport top-left coordinates in tiles. The original change cost remains $\lambda = 0.05$. The three values $\eta \in \{0, 0.05, 0.10\}$ were fixed in the saved protocol before inference; all three are reported, with no optimization or replacement primary setting. Each trajectory uses its own previous camera, initial action 434 at $(56, 60)$, all 840 candidates and the smallest index for an exact score tie. The existing frame-level evaluator, 48-frame hold and common human endpoint are unchanged.

The calculation uses 15 frozen fold–seed models and produces $17 \times 3 \times 3 = 153$ trajectories without fitting. At $\eta = 0$, all 51 match–seed action sequences exactly reproduce the primary outputs, and the maximum absolute overlap error is zero. For each motion metric, the initial transition is excluded, and the first and last dwell runs are retained. Total travel per minute sums Euclidean distances between successive output coordinates and divides by evaluated duration; unlike conditional jump length, it includes the frequency of movement. Large jumps are changes of at least 32 tiles. Seed summaries are averaged within each match, followed by equal match weighting.

Table S16: All pre-specified distance-cost settings with the original $\lambda = 0.05$. Dwell is the mean of within-trajectory medians; jump is conditional on a coordinate change.

η	Overlap	Changes/min	Dwell (s)	Jump (tiles)	Travel (tiles/min)
0.00	0.6257	9.63	3.04	22.43	215.54
0.05	0.6241	9.37	3.06	19.38	181.60
0.10	0.6227	9.21	3.06	17.67	163.20

Relative to $\eta = 0$, overlap differences are -0.0016 [-0.0043, 0.0011] at $\eta = 0.05$ and -0.0030 [-0.0070, 0.0009] at $\eta = 0.10$. Corresponding travel differences are -33.94 [-41.19, -27.27] and -52.34 [-62.35, -43.21] tiles/min. Large-jump rates are 1.9012, 1.5488 and 1.3380 per minute for the three settings, respectively. These marginal exploratory 95% intervals use 50,000 paired match-bootstrap draws with seed 20261004 after within-match seed averaging; models and settings are held fixed, without multiplicity adjustment. Intervals crossing zero for overlap do not establish equivalence or absence of a loss. The diagnostic demonstrates a distance–overlap trade-off within the frozen selector, not improved viewing comfort or event tracking. The original primary policy is unchanged, and the exact continuation bound derived for a constant per-change cost is not extended to this distance-sensitive rule.