ADDITIVITY AND SELECTION VALUE IN DEPTH-INDEXED MODEL UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting multi-step losses for a shared dynamics model requires predicting the value of training on a set of objectives. Does this prediction need interactions between losses? We study the question through finite-update probes that fit additive and quadratic set predictors to a fixed teacher-relative utility. For fixed-length SGD blocks, we derive orders $O(\eta)$, $O(\eta^2)$, and $O(\eta^3)$ for raw singleton, pair, and triple effects at step size η. The leading singleton term is initial-gradient alignment. Experiments on two continuous-control tasks with feedforward and recurrent dynamics exhibit this hierarchy. Ridge fitting nevertheless assigns substantial coefficients to pair features: removing the penalty shifts most larger-set prediction comparisons in favor of the quadratic model but changes no choice within the three-loss family. Additive and quadratic choices agree at 178 of 180 common states. In an optimizer stress test, ten-step Adam blocks from fresh optimizer state produce larger higher-order effects, and both predictors select negative-gain updates despite positive mean oracle gain. Closed-loop controls distinguish useful composition from useful interaction modelling. Selected triples improve lower-tail return over All-five updates at matched total loss weight, but increasing All-five's total weight reverses the ordering; continued inexpensive training also outperforms quadratic selection under the recorded cost budgets. These results identify three requirements for interaction modelling to improve training: nonadditive update effects, decision-relevant prediction, and benefits that justify acquiring the labels.

1 INTRODUCTION

Multi-step losses measure errors at different prediction depths but update shared transition parameters. Latent overshooting makes such supervision explicit (Hafner et al., 2019), while model-based agents use predicted trajectories for planning and behavior learning (Chua et al., 2018; Hafner et al., 2020; 2021; 2025; Hansen et al., 2022; 2024). Choosing losses therefore means choosing a joint update, not independent depth-specific modules.

An objective's current error need not measure its training value. Objective mismatch and value-aware learning motivate evaluating predictions through downstream use (Lambert et al., 2020; Farahmand et al., 2017; Voelcker et al., 2022). For training, that quantity is the benefit of an executable update. Gradient compatibility, learned task weighting, and task-group prediction estimate how objectives work together (Du et al., 2018; Liu et al., 2022; Standley et al., 2020). Yet selecting a useful combination and modelling its interactions are different achievements: an additive predictor may choose the same set as a richer one.

This creates an identification problem. Nonlinear dynamics need not produce strongly nonadditive short updates. Fitted pair coefficients do not directly measure that nonadditivity, because estimation can redistribute signal between correlated features. Acquiring joint values must also pay off through better training outcomes. Analyses of multi-task optimization have similarly found that simple scalarizations can remain competitive once training and regularization are controlled (Kurin et al., 2022; Xin et al., 2022).

We study these levels through finite-update probes. Candidate subsets receive the same optimization block from a common state, and a fixed teacher-relative utility measures gain over the common-objective update. Singleton and pair labels fit additive and quadratic predictors; unseen triples test

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.

extrapolation. Walker-walk and Cheetah-run with MLP and GRU dynamics provide common-state audits and closed-loop controls.

Our contributions are threefold.

1. A finite-block expansion gives a hierarchy of raw SGD effects and identifies gradient alignment as the leading singleton value. A cold-start Adam stress test illustrates how nonlinear gradient transformations can depart from this hierarchy.
2. A closed-form ridge analysis separates additive signal assigned to pair features from actual pair effects. Removing ridge improves most larger-set prediction comparisons yet leaves feasible choices unchanged; a margin diagnostic certifies 130 of 178 additive–quadratic agreements.
3. Closed-loop controls distinguish adaptive composition, total correction weight, and probe cost. Composition improves on tested random and fixed choices, while continued All-five training gives the highest recorded-budget lower-CVaR means among the compared methods.

The practical takeaway is to start with singleton or gradient-alignment selection for short, small-step SGD corrections, and tune total correction weight before acquiring pair labels. Interaction modelling must improve decisions enough to offset the training that its probes replace.

2 Related work

Decision-aware model learning addresses the mismatch between prediction likelihood and downstream control performance (Lambert et al., 2020). Value-aware losses incorporate value-function structure, and iterative value-aware learning links estimation to the planning procedure (Farahmand et al., 2017; Farahmand, 2018). Value equivalence characterizes models through their Bellman updates, whereas VaGraM weights errors by state-space value gradients (Grimm et al., 2020; Voelcker et al., 2022). ALM aligns representation, model, and policy learning through a common objective (Ghugare et al., 2023). We evaluate finite updates of existing loss subsets through decision agreement, rather than introduce a new value-aware dynamics objective.

PlaNet's latent overshooting supervises predictions across multiple steps (Hafner et al., 2019). Dreamer uses latent rollouts for behavior learning, and TD-MPC combines task-oriented model learning with planning (Hafner et al., 2020; 2021; Hansen et al., 2022; 2024). Our losses similarly supervise open-loop predictions along recorded actions. Selection changes the objective, not the prediction horizon or planner; the depth losses themselves are not new.

For shared-task optimization, GradNorm balances gradient magnitudes; PCGrad and CAGrad modify update directions; multi-objective approaches seek common descent or bargaining solutions (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021; Sener & Koltun, 2018; Navon et al., 2022). Gradient similarity offers a low-cost signal for auxiliary-task usefulness (Du et al., 2018). Task grouping estimates which objectives should share a network (Standley et al., 2020; Fifty et al., 2021); the former already uses two-task measurements to approximate larger groups. We test whether short-block value estimates transfer to unlabelled triples and change selection. Evaluations of multi-task optimizers also emphasize the strength of scalarized training and the influence of regularization (Kurin et al., 2022; Xin et al., 2022). Our analysis isolates how ridge assigns additive update value to correlated pair features. This motivates checking raw joint effects before interpreting fitted coefficients as task interactions.

Update valuation also appears in validation-guided sample weighting, meta-learned task weights, and subset selection (Ren et al., 2018; Shu et al., 2019; Liu et al., 2022; Killamsetty et al., 2021). AuxiLearn uses implicit differentiation to learn auxiliary objectives and nonlinear combinations of losses (Navon et al., 2021). Influence functions provide another way to reason about training-data effects (Koh & Liang, 2017). Datamodels fit subset indicators to predict training outcomes, demonstrating the usefulness of linear surrogates despite nonlinear underlying training (Ilyas et al., 2022). Where Datamodels predict outcomes from training-data subsets, we study loss subsets within a finite update block. We derive a local step-size hierarchy and identify nonlinear gradient transformations as a mechanism that can break it; label acquisition enters the cost comparison. Unlike coordinate

descent and Gauss–Southwell rules, the selected depths do not index disjoint parameter blocks (Nesterov, 2012; Nutini et al., 2015). Quadratic scores are standard pseudo-Boolean models (Boros & Hammer, 2002); our contribution is the separation of raw finite-update effects, their fitted representation, and the resulting selection value in shared-depth model updates.

3 SELECTING DEPTH-INDEXED CORRECTIONS

We evaluate subset selection over losses on shared transition parameters. Let $\mathcal{D} = \{2, 3, 4, 6, 8\}$ denote five prediction depths within a fixed horizon $H = 8$. The feasible family $\mathcal{S}_B = \{S \subseteq \mathcal{D} : |S| \le B\}$ includes the empty set; $B = 3$ gives 26 candidates. Three is the smallest budget that tests extrapolation beyond singleton and pair labels. It constrains cardinality rather than hardware cost: the losses share an eight-step rollout. All-five controls use all depths at matched per-depth or total weight. Figure 1 distinguishes discarded probe updates from retained training commits.

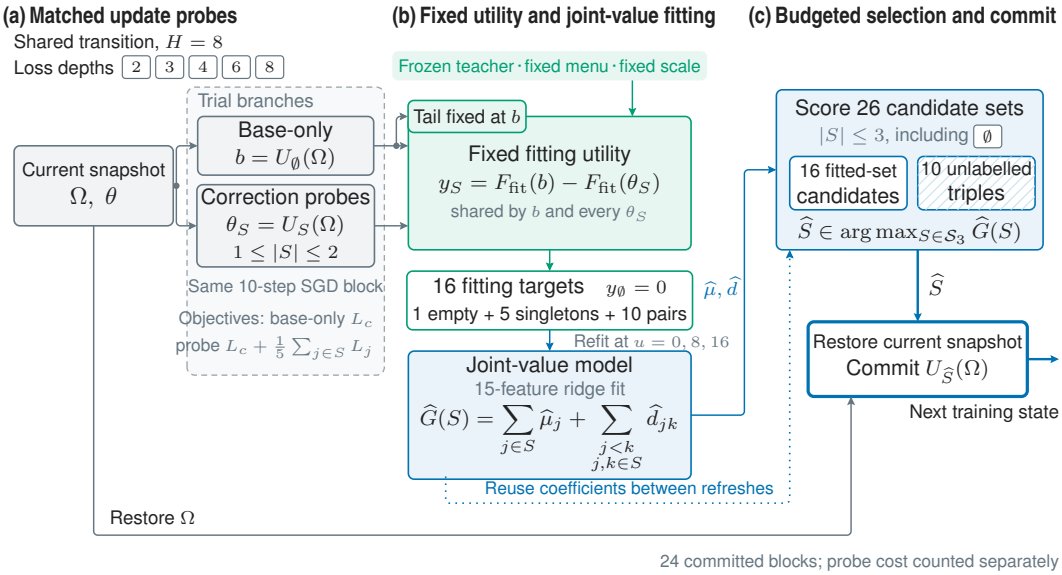

Figure 1: Joint-value correction selection. A common snapshot supplies the base-only branch and singleton/pair probes. Fixed reference targets, menus, scales, and base-only tail members define 16 fitting labels. A ridge quadratic model scores 26 feasible sets, including ten unlabelled triples. The selected ten-step SGD block is recomputed from the restored snapshot. Refreshes occur after 0, 8, and 16 committed blocks.

3.1 COMMON UPDATE, DIFFERENT DEPTH LOSSES

Standardized features $z_t \in \mathbb{R}^p$ use frozen training-set statistics. A residual transition predicts $\widehat{z}_{h+1} = \widehat{z}_h + f_\theta(\widehat{z}_h, a_h)$ from $\widehat{z}_0 = z_0$ along recorded actions. The base loss and depth losses are

$$L_0 = \frac{1}{p}\mathbb{E}_{\text{tr}}\|\widehat{z}_1 - z_1\|^2, \qquad L_j = \frac{1}{p}\mathbb{E}_{\text{tr}}\|\widehat{z}_j - z_j\|^2, \quad j \in \mathcal{D}. \tag{1}$$

The expectation averages trajectory windows, and targets are detached. Depth one is excluded because it duplicates L_0. The transition is either a two-hidden-layer SiLU MLP of width 256 or a single-layer GRU-256; both output state increments. GRU branches use up to four recorded steps of burn-in (Appendix B).

Base training optimizes L_0. During correction, a common objective $L_c = L_0 + 0.1L_{\text{rew}} + 0.1L_{\text{inv}}$ includes reward consistency and inverse dynamics along predicted states. Their trainable auxiliary heads are separate from the frozen reward readout used for utility. Both auxiliaries are enabled in every comparison reported here. We include their parameters in θ.

An update state Ω contains parameters, a fixed minibatch of 256 trajectory windows, historical inputs, and random state. For a set S, U_S takes ten plain-SGD steps of size $\eta = 10^{-3}$ on $L_c + \sum_{j \in S} \lambda_j L_j$, with $\lambda_j = 1/5$. The minibatch stays fixed within the block. Branches clone the same state and use no momentum, weight decay, or clipping. Let $\theta_S = U_S(\Omega)$ and $b = \theta_\varnothing$. Thus the empty branch still trains the common objective. Weights are not divided by $|S|$: changing cardinality changes the total correction weight as well as its composition.

3.2 A FIXED TEACHER-RELATIVE UTILITY

Each task uses a frozen GRU-512 residual teacher and a separate frozen reward readout $\bar{r}$, a width-256 two-hidden-layer SiLU MLP with a scalar sigmoid output. A utility case consists of an initial state, its available recorded history, and a fixed menu of 16 length-eight action sequences: eight teacher-CEM elites and eight uniform samples. The selection criterion compares learned-model predictions without querying simulator returns for these candidate sequences. For branch ϑ, define energies and choice probabilities by

$$E_\vartheta(i,k) = -\sum_{h=0}^{H-1} \gamma^h \bar{r}(\widehat{z}_{i,h}^{\vartheta,k}, a_{i,h}^{(k)}), \qquad p_\vartheta(k \mid i) = \frac{e^{-E_\vartheta(i,k)/(Ts_i)}}{\sum_l e^{-E_\vartheta(i,l)/(Ts_i)}}. \tag{2}$$

Here $\gamma = 0.99$, $T = 1$, and $s_i = \max\{\mathrm{sd}_k E_{\mathrm{ref}}(i,k), 10^{-3}\}$. The energy uses the unaveraged discounted sum. The reference supplies the target $k_i^* = \arg\min_k E_{\mathrm{ref}}(i,k)$ and the scale, not the candidate branch. Its disagreement loss is $\ell_i(\vartheta) = -\log p_\vartheta(k_i^* \mid i)$.

For a 128-case batch, retain the 13 highest losses at b and assign equal weights to these cases. With this tail $\mathcal{T}_\Omega$ fixed,

$$F_\Omega(\vartheta) = \frac{1}{13} \sum_{i \in \mathcal{T}_\Omega} \ell_i(\vartheta), \qquad G_\Omega(S) = F_\Omega(b) - F_\Omega(\theta_S). \tag{3}$$

Reference targets, menus, scales, and base-only tail members remain fixed across candidates. Since $G_\Omega(\varnothing) = 0$, the exact utility maximizer has nonnegative gain. The actual utility gain of a predicted maximizer can be negative. Fitting and audit batches use disjoint trajectories and determine their tails separately. Closed-loop control is evaluated independently of this teacher-relative criterion.

3.3 FITTING AND COMMITTING JOINT VALUES

Singleton gains and finite-difference pair effects define a quadratic set representation:

$$\mu_j = G(\{j\}), \quad d_{jk} = G(\{j,k\}) - \mu_j - \mu_k, \qquad \widehat{G}_P(S) = \sum_{j \in S} \widehat{\mu}_j + \sum_{j<k \in S} \widehat{d}_{jk}. \tag{4}$$

We call the fitted predictor Quadratic to distinguish its coefficients from the raw pair effects. Positive pair effects indicate complementarity relative to singleton gains; negative effects indicate interference. At each refresh, the fitting family $\mathcal{S}_2$ supplies 16 labels, including the empty set. For feature vector $\varphi(S)$, we minimize

$$\frac{1}{n} \sum_{S \in \mathcal{L}} [y_S - \varphi(S)^\top \beta]^2 + 10^{-3} \|\beta\|_2^2, \qquad y_S = G_\Omega^{\mathrm{fit}}(S). \tag{5}$$

Quadratic uses 15 singleton–pair indicators with $(\mathcal{L}, n) = (\mathcal{S}_2, 16)$. Additive-16 uses the same labels and five singleton indicators. Additive-6 uses only the empty and five singleton labels, with $n = 6$. All fits have no intercept and zero empty score. Zero-penalty refits on the same labels separate representation from regularization.

We enumerate the 26 feasible sets and select the highest fitted score. Ties prefer smaller sets, then lexicographic depth order. We refresh after 0, 8, and 16 completed blocks and reuse each set for eight blocks, giving 24 commits per run. Methods follow their own training trajectories; common-state comparisons fit both predictors from one snapshot.

Quadratic and Additive-16 acquire 48 probe blocks and commit 24 blocks, for 720 SGD steps in total. Additive-6 acquires 18 probes and commits 24, for 420 steps. All-five commits 240 steps without fitting probes. These counts exclude menu generation, utility evaluation, copying, fitting, and search, which enter the cost analysis below.

4 RAW EFFECTS, FITTED COEFFICIENTS, AND SELECTION

A nonlinear model can have nearly additive update values. To separate raw effects from fitted coefficients and decisions, define the omitted effect of a quadratic representation as

$$\kappa(S) = G(S) - \sum_{j \in S} \mu_j - \sum_{j<k \in S} d_{jk}. \tag{6}$$

It vanishes on $\mathcal{S}_2$ by definition. For a triple, it is the third finite difference of the set function, not a fitted prediction error.

Proposition 1 (Set effects of a fixed-length SGD block)**.** *Fix the initial parameters, training data, utility F, and a finite step count m. Suppose F, L_c, and the weighted depth losses are C^3 with bounded derivatives on a neighborhood containing all candidate iterates for sufficiently small η. Put $g_j = \nabla(\lambda_j L_j)(\theta)$, $A_j = \nabla^2(\lambda_j L_j)(\theta)$, $h_\eta = \nabla F(b_\eta)$, $H_\eta = \nabla^2 F(b_\eta)$, and $c_m = m(m-1)/2$, where b_η is the base-only endpoint. Then*

$$\begin{aligned} \mu_j &= m\eta \nabla F(\theta)^\top g_j + O(\eta^2), \\ d_{jk} &= -m^2\eta^2 g_j^\top H_\eta g_k - c_m \eta^2 h_\eta^\top (A_j g_k + A_k g_j) + O(\eta^3), \end{aligned} \tag{7}$$

and $\kappa(S) = O(\eta^3)$ for every triple. The remainders are uniform over the finite candidate family with m fixed.

The second term in Eq. (7) captures nonadditive parameter displacement across steps; at $m = 1$, only the utility-curvature cross-term remains. If the leading singleton RMS is nonzero, raw pair-to-singleton RMS is $O(\eta)$. These are local upper-order statements, not guaranteed nonzero leading coefficients or a limit in which m also grows. Appendix A.5 derives the expansion and shows why nonlinear gradient transformations need not preserve this hierarchy.

Fitted pair coefficients have a different interpretation. For five depths, the nonempty fitting design is $X = [I_5, 0; A, I_{10}]$, where each row of A indicates a pair. Even for exactly additive labels $y = [u; Au]$, ridge with $q = 16 \times 10^{-3}$ gives

$$\widehat{d}_{jk} = \frac{q}{a}\left(u_j + u_k - \frac{2q}{a+5q}\sum_l u_l\right), \qquad a = 1 + 5q + q^2. \tag{8}$$

This predicts fitted pair coefficients from purely additive labels. Positivity requires the pair sum to exceed the threshold, even with positive singleton gains. This fitted signal can be $O(\eta)$ while raw pair effects are $O(\eta^2)$. Appendix A.2 derives the normal equations and design spectrum. More generally, correlated subset features allow ridge to redistribute additive signal into interaction coefficients; Eq. (8) quantifies this effect for our specific fitting design.

Selection adds another requirement. Let $a(S)$ and $p(S)$ be the fitted additive and quadratic scores at the same state. If the unique additive maximizer S_a has margin $\gamma_a = a(S_a) - \max_{S \neq S_a} a(S)$, then:

Remark 1 (Stability of the fitted selection)**.** *If $\varepsilon = \max_{S \in \mathcal{S}_3} |p(S) - a(S)|$ and $\gamma_a > 2\varepsilon$, both fitted models uniquely select S_a.*

Every competitor retains a deficit of at least $\gamma_a - 2\varepsilon > 0$. The condition includes fitted singleton changes and certifies score agreement rather than true utility or probe savings.

Proposition 1 also explains the initial-gradient control: $a_j = \nabla F(\theta)^\top g_j$ is the leading singleton coefficient. At a common state, fresh additive fits preserve nonzero gradient-ranking gaps to first order, including under our ridge design (Appendix A.2). Matching an exactly-three selector also requires three positive fitted coefficients.

5 EXPERIMENTS

5.1 PROTOCOL AND COMPARISONS

We evaluate proprioceptive Walker-walk and Cheetah-run from the DeepMind Control Suite (Tassa et al., 2018), with MLP and GRU dynamics. Per task, 2,000 trajectories combine random actions

and two SAC checkpoint policies. Bases receive 100,000 pretraining updates; the teacher receives 300,000 (Appendix B). Five base checkpoints, three correction runs per base, and three refreshes give 45 common states per setting. Fitting and two audit batches use disjoint trajectories. Audits compare fixed-state updates; separately trained selectors provide closed-loop returns.

Evaluation pairs 64 episode seeds across methods. CEM uses 256 samples, 32 elites, three iterations, and horizon eight; it executes the first action and replans. It scores rollouts with the same frozen reward readout $\bar{r}$ used in utility. Episode score is total reward divided by the 1,000-step episode length. Fractional lower-tail $\mathrm{CVaR}_{0.1}$ emphasizes poorly performing episodes and is computed within each correction run, then averaged over runs and bases (Rockafellar & Uryasev, 2002; Agarwal et al., 2021). We report teacher-relative utility, mean return, and cost separately.

Fixed-update comparisons retain 24 blocks. Quadratic and both additive selectors allow at most three depths. A strict-three Quadratic control has the same observed choices and episode outcomes. All-five assigns either 0.12 or 0.20 per depth, giving total weights 0.6 or 1.0. Random, fixed, and initial-gradient alignment controls select exactly three depths. The gradient heuristic evaluates the first-order signal in Proposition 1 at the pre-update point.

Return intervals use 20,000 paired hierarchical bootstrap draws. Four primary families and a separate 12-contrast post-hoc family use family-specific levels, not joint coverage over all 44 comparisons. In each of the four matched-weight contrasts and 12 supplementary contrasts, all five base-paired differences have the same sign. These descriptive patterns coexist with an exact sign-flip resolution of two-sided $p \geq 0.0625$: no contrast rejects at 5% by that test (Appendix B.5).

5.2 Small-step SGD yields a hierarchy of raw set effects

Small-step SGD yields nearly additive update values. At the default ten-step block, raw pair-effect RMS is 0.535–0.722% of singleton-gain RMS on the primary tail audit. Triple residuals are smaller still (Table 10). Figure 2(a) varies η from 3×10^{-4} to 10^{-2} while keeping $m = 10$, the utility anchor, and the state panel fixed. Across the four task–backbone settings, descriptive log–log slopes are 0.969–0.971 for singleton RMS, 1.957–1.974 for pair RMS, and 2.944–3.009 for triple residual RMS, consistent with Proposition 1. Across seven SGD configurations and both targets, setting-mean selected displacements span 0.0016–0.158; pooled pair/singleton RMS ratios stay below 9.1%, with at least 86.7% selector agreement in each comparison (Appendix C.4).

5.3 Ridge changes prediction comparisons without changing selection

Additive shrinkage explains much of the fitted pair-coefficient sign and scale. On the same fitting labels, ridge pair-coefficient RMS is 3.18–3.78 times raw pair-effect RMS; 1,783 of 1,800 fitted coefficients are positive (Figure 2(b)). Applied to the measured singleton gains, Eq. (8) predicts the signs of 1,796/1,800 fitted coefficients and an RMS of 1.56×10^{-4}, compared with the observed 1.30×10^{-4}. The remaining interaction response is nonzero: its RMS is 27.5% of fitted-coefficient RMS (Appendix A.2).

Refitting without the penalty separates this effect from representational capacity. With ridge, Quadratic has higher prediction RMSE than Additive-16 in 30 of 32 four-/five-depth comparisons across settings, audits, and utility targets. Without ridge, Quadratic has lower RMSE in 30 of those 32 cells; its positive relative improvements range from 0.17% to 1.91%. For triples, zero-penalty Quadratic has lower RMSE in 15 of 16 cells. Tail and full-case labels are fitted separately (Appendix E).

Despite those changes, removing ridge changes neither model's $\mathcal{S}_3$ choice at any state, for either target. Quadratic and Additive-16 agree at 178/180 common states, and Remark 1 certifies 130 of those agreements. Along their own training trajectories they agree at 177/180 matched refresh indices; Additive-6 and Additive-16 agree at all 180 and have identical endpoint scores. All choices contain three depths. Better zero-penalty prediction leaves the observed common-state decisions unchanged.

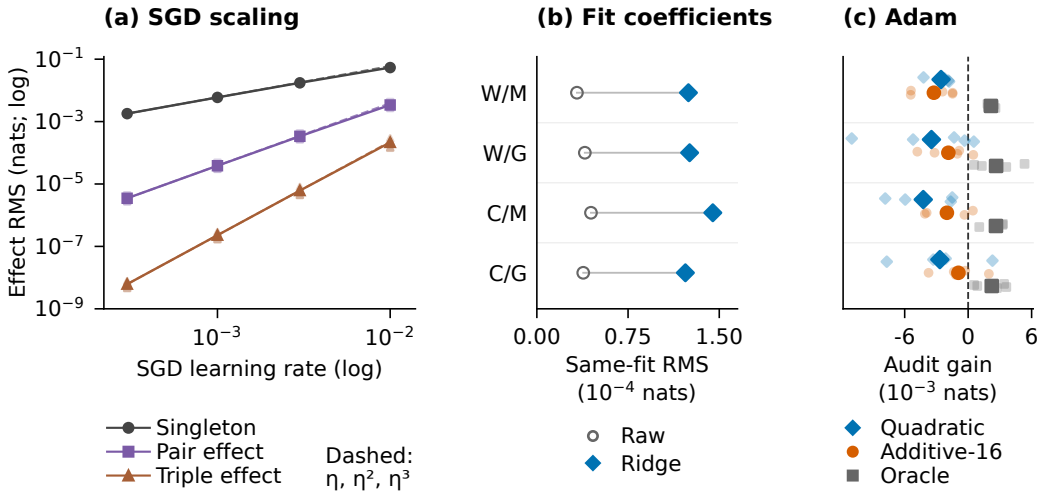

Figure 2: Raw effects, fitted coefficients, and optimizer failure. (a) Primary-audit fixed-tail RMS at four learning rates with ten SGD steps. Each point pools squared raw effects across the four settings; dashed reference slopes are anchored at the smallest step. (b) Same-fitting-split raw pair and ridge-coefficient RMS, separately by setting; these are not the audit labels in (a). (c) Under ten-step cold-start Adam, the 26-set audit oracle has positive mean gain, whereas both fitted selectors have negative mean gain. Open points average nine states within each of five bases; filled markers average the bases. Panels are descriptive, with utility in nats. W/C denote Walker/Cheetah; M/G denote MLP/GRU.

5.4 Stronger interactions can expose selection failure

Cold-start Adam provides a stress test of the nonlinear gradient-map mechanism in Eq. (27), rather than a model of warm-state training (Kingma & Ba, 2017). For ten-step tail-based selections, the ratios of setting-mean displacement norms to default SGD are 0.99–1.19, yet raw triple residuals are over four orders of magnitude larger. Similar mean norms do not isolate a causal mechanism. The first cold-start step approximates a sign update for gradients much larger than Adam's epsilon. A common fixed linear preconditioner would remove its first-order pair term, but warm moments alone need not do so (Appendix A.5). Warm-state Adam is not evaluated.

Useful candidates remain available under cold-start Adam. The 26-set oracle has positive mean tail gain in every setting, from 0.00215 to 0.00264 nats, while Quadratic's mean gain is -0.00256 to -0.00424; Additive-16 is also negative in all four settings (Figure 2(c)). At 83/180 states, oracle gain is strictly positive and Quadratic's selected gain is negative. Some oracle states tie at zero. Full-case audits give the same sign pattern in the four setting means. Appendix C.4 gives the complete fixed-state scan.

5.5 Composition and correction strength have different effects

Adaptive composition improves lower-tail control at matched total loss weight (Table 1; Figure 3). Strict-three Quadratic exceeds All-five at total weight 0.6 by 0.01077–0.01631 lower-CVaR units; all four 98.75% intervals exclude zero. At weight 0.20 per depth, All-five instead exceeds at-most-three Quadratic by 0.00946–0.01262, with all four supplementary 99.5833% intervals excluding zero. Increasing correction strength reverses the return ordering.

Quadratic exceeds random triples by 0.01195–0.01804 and fixed $\{2, 4, 6\}$ by 0.01713–0.02379, with all eight primary intervals excluding zero. For fixed $\{4, 6, 8\}$, the Cheetah/MLP interval includes zero. By contrast, every Quadratic-minus-gradient interval contains or touches zero, with estimates between -0.000027 and 0.000376. The close returns are consistent with the first-order connection in Proposition 1; they do not establish identical rankings or statistical equivalence. Increasing All-five's total weight gives 0.02225–0.02893, larger than the 0.01077–0.01631 composition contrast at weight 0.6. These are distinct interventions, not a decomposition of one effect.

Table 1: Lower-tail return after 24 update blocks, averaged over five bases and three correction runs per base. Additive-6/16 make identical choices and have identical run-level scores; other displayed equalities can reflect rounding.

Method	Total λ	W/MLP	W/GRU	C/MLP	C/GRU
Quadratic	0.6	0.3890	0.3944	0.3208	0.3386
Additive-16	0.6	0.3890	0.3944	0.3205	0.3385
Additive-6	0.6	0.3890	0.3944	0.3205	0.3385
All-five (0.12)	0.6	0.3744	0.3781	0.3059	0.3278
All-five (0.20)	1.0	0.3985	0.4070	0.3324	0.3501
Random three	0.6	0.3716	0.3774	0.3028	0.3267
Fixed $\{2, 4, 6\}$	0.6	0.3719	0.3752	0.2971	0.3204
Fixed $\{4, 6, 8\}$	0.6	0.3738	0.3787	0.3083	0.3230
Initial-gradient	0.6	0.3891	0.3944	0.3205	0.3385

W/C denote Walker/Cheetah. CVaR is computed within each 64-episode run. All observed selector decisions use three depths; All-five uses five.

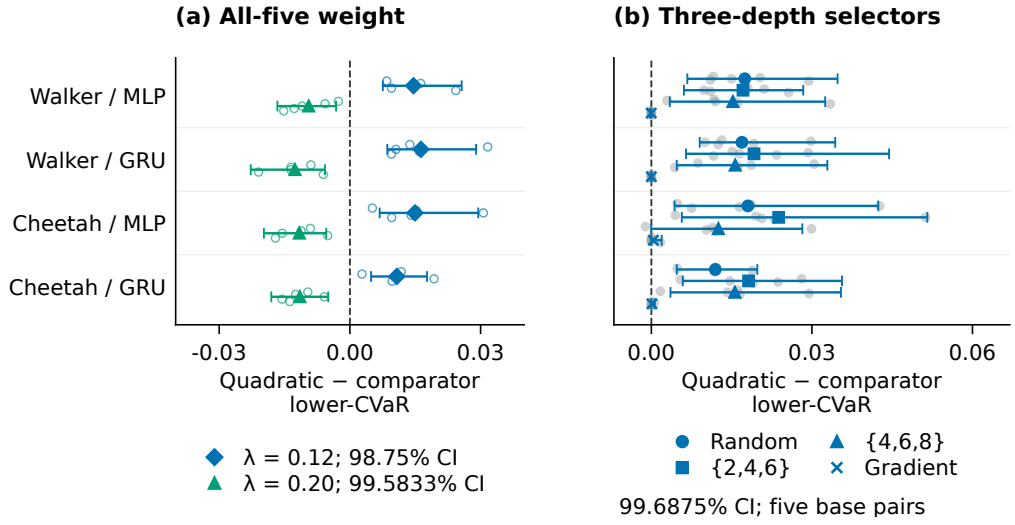

Figure 3: Composition and total weight lead to opposite return orderings. Left: Quadratic minus All-five. Weight 0.12 compares strict-three Quadratic with equal total weight and uses primary 98.75% intervals; weight 0.20 compares at-most-three Quadratic and uses supplementary 99.5833% intervals. The two Quadratic controls have identical observed outcomes. Right: strict-three Quadratic versus trained random, fixed, and gradient controls, with primary 99.6875% intervals. Open points are five paired base means; filled markers aggregate them. All intervals use paired hierarchical bootstrap.

Independent utility checks support a directed correction effect. The tail-fitted Quadratic choice has positive gain at all states under both audits and both tail/full evaluations. It also exceeds matched-norm antithetic random perturbations in mean tail gain in every setting; the random gains average near zero with approximately balanced signs. Exact best-set identities are less stable: the two 26-set audit oracles agree at 103/180 states, despite high all-set rank correlations. Figure 6 and Appendix C.3 report the complete checks.

5.6 Label acquisition competes with continued training

Continued training gives higher lower-tail CVaR means than acquiring pair labels at Quadratic's recorded online budget (Figure 4). Controls continue beyond 24 blocks, with per-run ranges of 40–43 Additive-6 blocks, 41–48 All-five blocks, and 149–175 base-only blocks. Accounting includes

candidate updates, menu construction, fitting utility, selection, copying, and commits; independent audits and final evaluation are excluded.

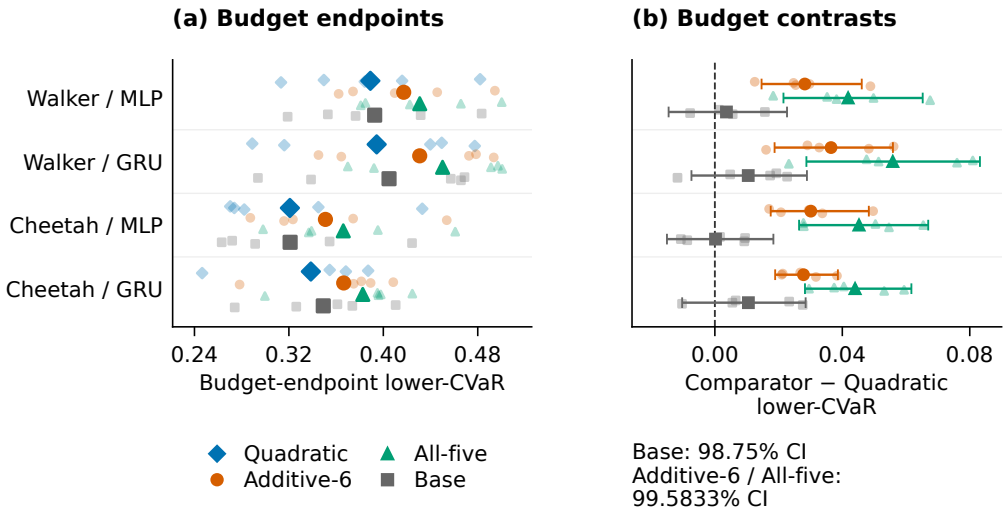

Figure 4: Continued training at the recorded-cost budget. Left: endpoints at each control's last completed checkpoint. Right: comparator minus Quadratic, with primary 98.75% intervals for base-only and supplementary 99.5833% intervals for Additive-6 and All-five. Open points are five base-paired means; filled markers aggregate them. All comparisons use the same paired budgets.

Continued Additive-6 and All-five exceed Quadratic by 0.02782–0.03646 and 0.04182–0.05584, respectively. All eight supplementary 99.5833% intervals exclude zero. Base-only also has higher point estimates in all four settings, although its primary intervals include zero. The descriptive lower-CVaR ordering is All-five(0.20), Additive-6, base-only, then Quadratic in every setting (Table 21). All-five exceeds base-only by 0.03347–0.04536 in these means. Thus multi-step correction retains value under this accounting, while acquiring pair labels is less productive than continued training. Appendix B.6 specifies the accounting and the 304 effective Additive-6 refreshes, excluding late or unused decisions.

5.7 Scope and Implications

The evidence covers two tasks, two backbones, one teacher/readout per task, and 240-step default correction runs. Quadratic fits 15 parameters to 16 labels separately at each state. All-five uses total weights 0.6 and 1.0; triples at weight 1.0 and weights above 1.0 are untested. The weight-1.0 comparison therefore changes both cardinality and per-depth weight. Equal coefficient sums do not match update norms or costs. The finite-grid slopes describe the tested SGD range; Adam uses cold moments, with no closed-loop Adam experiment. Fixed-tail and random-direction audits support directed utility gains but do not eliminate all tail-selection bias or identify pair features as the source of return gains. Teacher-relative utility, lower-tail CVaR, and mean return are distinct targets; recorded costs do not measure hardware latency.

6 Conclusion

Depth-loss selection can help without interaction modelling improving its decisions. Small-step SGD orders raw effects by subset size; ridge coefficients need not follow that hierarchy. Removing ridge improves quadratic prediction but leaves observed choices unchanged, while cold-start Adam exposes failures of both low-order predictors. For short SGD corrections, start with singleton selection or its leading-order gradient-alignment signal, which avoids singleton and pair update probes. Without a cardinality constraint, tune All-five's total weight and compare continued training. Acquiring pair labels is warranted when independent audits establish decision gains that survive this budget comparison.

AI USE STATEMENT

AI tools assisted with writing and revision, methodological and experiment-design discussion, mathematical derivations, literature discovery, data preparation and analysis code, visualization, and interpretation of results. We take responsibility for the scientific content, reported results, citations, and supporting materials.

REPRODUCIBILITY STATEMENT

Section 3 specifies the update operator, utility, label families, fitting objective, and refresh schedule. Appendix B gives the model interfaces, auxiliary losses, and evaluation procedure. Appendix A provides derivations and proofs. The complete endpoint comparisons, independent larger-set audits, operator scans, and selection diagnostics are in Appendix C. Appendices D and E describe the additional fixed-state diagnostics. The accompanying table and plotting sources recreate the displayed tables and figures from the included numerical records and summaries, preserving the paired experimental units. We plan to release the model-training and evaluation code.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://papers.nips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Endre Boros and Peter L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1–3):155–225, 2002. doi: 10.1016/S0166-218X(01)00341-9. URL https://www.sciencedirect.com/science/article/pii/S0166218X01003419.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803. PMLR, 2018. URL https://proceedings.mlr.press/v80/chen18a.html.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity, 2018. URL https://arxiv.org/abs/1812.02224v1. Version 1, first posted 5 December 2018.

Amir-Massoud Farahmand. Iterative value-aware model learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/7a2347d96752880e3d58d72e9813cc14-Paper.pdf.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1486–1494. PMLR, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27503–27516. Curran Associates, Inc., 2021. URL https:

//proceedings.neurips.cc/paper_files/paper/2021/hash/e77910ebb93b511588557806310f78f1-Abstract.html.

Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. Simplifying model-based RL: Learning representations, latent-space models, and policies with one objective. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MQcmfgRxf7a.

Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3bb585ea00014b0e3ebe4c6dd165a358-Paper.pdf.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8387–8406. PMLR, 2022. URL https://proceedings.mlr.press/v162/hansen22a.html.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, 2021. doi: 10.1609/aaai.v35i9.16988. URL https://ojs.aaai.org/index.php/AAAI/article/view/16988.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and Pawan K Mudigonda. In defense of the unitary scalarization for deep multi-task learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 12169–12183. Curran Associates, Inc., 2022. doi: 10.52202/068431-0884. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4f301ae934f396086bfefd1139039dbd-Paper-Conference.pdf.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770. PMLR, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Shikun Liu, Stephen James, Andrew J. Davison, and Edward Johns. Auto-Lambda: Disentangling dynamic task relationships. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=KKeCMim5VN.

Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, and Ethan Fetaya. Auxiliary learning by implicit differentiation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=n7wIfYPdVet.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. doi: 10.1137/100802001. URL https://epubs.siam.org/doi/10.1137/100802001.

Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1632–1641. PMLR, 2015. URL https://proceedings.mlr.press/v37/nutini15.html.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6. URL https://doi.org/10.1016/S0378-4266(02)00271-6.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information*

Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/e58cc5ca94270acaceed13bc82dfedf7-Abstract.html.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132. PMLR, 2020. URL https://proceedings.mlr.press/v119/standley20a.html.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. URL https://arxiv.org/abs/1801.00690.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=4-D6CZkRXxI.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. Do current multi-task optimization methods in deep learning even help? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 13597–13609. Curran Associates, Inc., 2022. doi: 10.52202/068431-0988. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/580c4ec4738ff61d5862a122cdf139b6-Paper-Conference.pdf.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

A Derivations and proofs

The statements in this appendix concern one fixed update state and one fixed utility functional. We suppress the subscript Ω when it does not change. Parameters are measured in the Euclidean norm, and matrix norms are the corresponding operator norms.

A.1 Exact pair representation on a budget-two family

For any set function G with $G(\varnothing) = 0$, define $\mu_j = G(\{j\})$ and $d_{jk} = G(\{j,k\}) - \mu_j - \mu_k$. The pair score formed from these exact quantities reproduces G on every set of size at most two. On the empty set both sums vanish. On a singleton, the score is $\mu_j = G(\{j\})$. On a pair it is $\mu_j + \mu_k + d_{jk} = G(\{j,k\})$. No differentiability or linear update assumption is needed for this identity.

Consequently, an exact pair representation at $B = 2$ does not establish an ability to predict unseen gains. Its coefficients can require precisely the joint labels whose acquisition is costly. It also supplies no restriction on the values of $G(S)$ for larger sets.

A.2 Ridge fitting on singleton and pair labels

Exact pair representation and ridge estimation are different operations. For $J = 5$, omit the zero empty-set row and order the nonempty rows as five singletons followed by ten pairs. The pair design matrix has form

$$X = \begin{bmatrix} I_5 & 0 \\ A & I_{10} \end{bmatrix}, \qquad A^\top A = 3I_5 + \mathbf{1}\mathbf{1}^\top, \tag{9}$$

where A is the unsigned edge–vertex incidence matrix of the complete graph on five vertices. The normalization in Eq. (5) gives $\widehat{\beta} = (X^\top X + 0.016I)^{-1} X^\top y$. The eigenvalues of $X^\top X$ are

$$5 \pm 2\sqrt{6} \quad \text{(once each)}, \qquad \frac{5 \pm \sqrt{21}}{2} \quad \text{(four times each)}, \qquad 1 \quad \text{(five times)}. \tag{10}$$

To obtain these values, decompose A by its singular vectors. For each squared singular value a, the corresponding two-dimensional block of $X^\top X$ has trace $a + 2$ and determinant one. Here $a = 8$ once and $a = 3$ four times; the remaining five directions have eigenvalue one.

Relative to the unregularized fit, ridge retains $\lambda/(\lambda + 0.016)$ in a mode with eigenvalue λ. The minimum factor is approximately 0.8633. This is a property of the design, not a measurement of the experimental fitting error: its effect depends on the observed labels and ranking margins. In particular, a small numerical regularizer does not guarantee that fitting preserves the label argmax.

For any fitted quadratic score and distinct j, k, l,

$$\begin{aligned} \widehat{G}(\{j,k,l\}) = \widehat{G}(\{j,k\}) + \widehat{G}(\{j,l\}) + \widehat{G}(\{k,l\}) \\ - \widehat{G}(\{j\}) - \widehat{G}(\{k\}) - \widehat{G}(\{l\}). \end{aligned} \tag{11}$$

These are fitted scores. Replacing them with the raw labels requires exact interpolation, which ridge does not enforce. Triple prediction error can therefore include fitting error as well as κ.

Exactly additive labels can yield nonzero fitted pair coefficients. Write $y = [u; Au]$, whose unregularized representation is $[u; 0]$. With $q = 0.016$, the pair-coordinate normal equations give $\widehat{d} = A(u - \widehat{\mu})/(1 + q)$. Eliminating $\widehat{\mu}$, put $M = (1+q)^2 I + qA^\top A = aI + q\mathbf{1}\mathbf{1}^\top$, where $a = 1 + 5q + q^2$. The rank-one inverse gives

$$\begin{aligned} M^{-1} &= a^{-1}\left(I - \frac{q}{a + 5q}\mathbf{1}\mathbf{1}^\top\right), \\ \widehat{d}^{\,\mathrm{add}}_{jk} &= \frac{q}{a}\left(u_j + u_k - \frac{2q}{a + 5q}\sum_l u_l\right). \end{aligned} \tag{12}$$

Here $q/a = 0.0148113$ and $2q/(a + 5q) = 0.0275801$. All pair coefficients are positive precisely when the smallest pair sum exceeds this common threshold. Even if every u_j is positive, that condition need not hold.

The measured labels have the form $y = [u; Au + e]$, with e the raw same-fitting-split pair effects. Linearity of the ridge solution yields the exact decomposition

$$\widehat{d} = \underbrace{qAM^{-1}u}_{\widehat{d}^{\mathrm{add}}} + \underbrace{\frac{1}{1+q}\left(I - qAM^{-1}A^{\top}\right)e}_{\widehat{d}^{\mathrm{int}}}. \tag{13}$$

The first term is what ridge would fit after replacing each pair label by its singleton sum; the second is its response to actual interactions. This algebraic counterfactual does not change a model update or add experimental labels.

For the 180 tail-fitting states, $\widehat{d}^{\mathrm{add}}$ and $\widehat{d}$ each contain 1,783 positive entries among 1,800, but their signs disagree at four entries. Their pooled RMS values are 1.55854×10^{-4} and 1.29595×10^{-4} nats; their correlation is 0.96305. The interaction response has RMS 3.55958×10^{-5} nats, or 27.47% of fitted-coefficient RMS. These components need not be orthogonal, so their RMS values do not define additive fractions of explained variance. The additive component predicts much of the fitted sign and scale without explaining every coefficient. Table 2 reports the four settings.

Table 2: Decomposition of fitted pair coefficients on the primary tail13 fitting labels. The additive-label component retains recorded singleton gains and replaces pair labels by their singleton sums; Residual is the exact ridge response to the observed nonadditive label component. RMS values are in 10^{-4} nats, pooled over 450 pair coordinates at 45 states per setting. The final column compares coefficient signs between the additive-label component and the observed fit. RMS ratios are descriptive and are not variance-explained fractions. W/C denote Walker/Cheetah.

Setting	Observed	Additive-label	Residual	Residual/observed (%)	Same sign
W/MLP	1.2465	1.4598	0.3013	24.17	449/450
W/GRU	1.2562	1.5183	0.3592	28.59	449/450
C/MLP	1.4470	1.7616	0.4058	28.04	450/450
C/GRU	1.2215	1.4753	0.3497	28.63	448/450

Additive-16 uses $Z = [I_5; A]$ and the normal equations $(Z^{\top}Z + 0.016I)\widehat{\alpha} = Z^{\top}y$. Additive-6 uses singleton rows and its own objective normalization, giving $(I_5 + 0.006I_5)\widehat{\alpha} = u$. The zero empty-set row changes neither right-hand side, but remains part of the mean-square denominator.

The Additive-16 equations also connect its ranking to gradient alignment. Write $C = Z^{\top}Z = 4I + \mathbf{1}\mathbf{1}^{\top}$ and $y = [u; Au + e]$. For $q = 0.016$, its exact solution is

$$\widehat{\alpha} = cu + b(\mathbf{1}^{\top}u)\mathbf{1} + (C + qI)^{-1}A^{\top}e, \qquad c = \frac{4}{4+q}, \quad b = \frac{q}{(4+q)(9+q)}. \tag{14}$$

Indeed, $(C + qI)^{-1}C = I - q(C + qI)^{-1}$, and a rank-one inverse gives the first two terms. For the fixed utility and common state of Proposition 1, put $a_j = \nabla F(\theta)^{\top} g_j$. Then $u = m\eta a + O(\eta^2)$ and $e = O(\eta^2)$, so $\widehat{\alpha}_j - \widehat{\alpha}_k = cm\eta(a_j - a_k) + O(\eta^2)$ with $c > 0$. Thus a nonzero gap between the third and fourth largest a_j preserves the top-three set for sufficiently small η. For maximization over $\mathcal{S}_3$, a sufficient additional condition is $ca_{(3)} + b\sum_j a_j > 0$ so that three fitted coefficients remain positive. Additive-6 instead shrinks every singleton by $1/(1 + 0.006)$ and has the analogous condition $a_{(3)} > 0$. These are fresh-fit, common-state statements. Finite step sizes, reused fits, or different training trajectories need not preserve ranking. Close environment returns alone do not establish set agreement.

The signal-transfer mechanism also extends beyond this particular design. For a general two-block design $X = [Z, W]$, purely lower-order labels $y = Zu$, and a common ridge penalty $q > 0$, put $Q_Z = Z^{\top}Z + qI$ and $R_W = W^{\top}W + qI - W^{\top}ZQ_Z^{-1}Z^{\top}W$. Eliminating the first coefficient block from the normal equations gives $\widehat{\delta} = qR_W^{-1}W^{\top}ZQ_Z^{-1}u$ for the second block. Correlated feature blocks can therefore transfer lower-order signal under joint ridge penalization. Orthogonal blocks yield zero transfer; correlation alone fixes neither its presence for a given u nor its sign. The numerical factors in Eq. (12) remain specific to the complete singleton–pair design above.

A.3 Taylor remainder for an actual update displacement

For an actual displacement $\Delta_S = \theta_S - b$, the utility expansion is

$$G(S) = -\nabla F(b)^\top \Delta_S - \tfrac{1}{2}\Delta_S^\top H_F(b)\Delta_S - R_S, \qquad |R_S| \le \tfrac{\rho}{6}\|\Delta_S\|^3. \tag{15}$$

Assume that F has a continuous Hessian on the line segment from b to $b+\Delta$, and that $\|H_F(x) - H_F(y)\|_{\mathrm{op}} \le \rho\|x-y\|$ on that segment. The integral expansion is

$$F(b+\Delta) = F(b) + \nabla F(b)^\top \Delta + \tfrac{1}{2}\Delta^\top H_F(b)\Delta + R(\Delta), \tag{16}$$

$$R(\Delta) = \int_0^1 (1-t)\Delta^\top [H_F(b+t\Delta) - H_F(b)]\Delta \, dt. \tag{17}$$

The Lipschitz condition yields

$$|R(\Delta)| \le \rho\|\Delta\|^3 \int_0^1 t(1-t)\, dt = \frac{\rho}{6}\|\Delta\|^3. \tag{18}$$

Subtracting this expansion from $F(b)$ proves Eq. (15) with $R_S = R(\Delta_S)$. The weights and members in Eq. (3) are held fixed in this calculation. Re-selecting them through a fixed CVaR rule defines a different functional, whose smoothness must be checked separately. The expansion above does not automatically extend across its tail-membership changes.

A.4 Single-step SGD as a special case

For a single simultaneous step, all gradients are evaluated at θ:

$$b = \theta - \eta g_0, \quad \theta_S = \theta - \eta(g_0 + \sum_{j\in S} g_j), \quad g_0 = \nabla L_{\mathrm{c}}(\theta), \quad g_j = \lambda_j \nabla L_j(\theta). \tag{19}$$

Proposition 2 (Pair effects of an additive displacement)**.** *Under Eqs.* (15)–(19)*,* $d_{jk} = -\eta^2 g_j^\top H_F(b) g_k + R_j + R_k - R_{jk}$*. The remainder magnitude is at most* $\rho(\|\Delta_j\|^3 + \|\Delta_k\|^3 + \|\Delta_{jk}\|^3)/6$.

Under Eq. (19), $\Delta_j = -\eta g_j$ and $\Delta_{jk} = \Delta_j + \Delta_k$. Write $h = \nabla F(b)$ and $H = H_F(b)$. The joint expansion minus the two singleton expansions gives

$$\begin{aligned} d_{jk} &= -h^\top(\Delta_{jk} - \Delta_j - \Delta_k) \\ &\quad - \tfrac{1}{2}\left(\Delta_{jk}^\top H \Delta_{jk} - \Delta_j^\top H \Delta_j - \Delta_k^\top H \Delta_k\right) - R_{jk} + R_j + R_k \qquad (20) \\ &= -\Delta_j^\top H \Delta_k + R_j + R_k - R_{jk} \qquad (21) \\ &= -\eta^2 g_j^\top H g_k + R_j + R_k - R_{jk}. \qquad (22) \end{aligned}$$

The first term in the first equality vanishes because the displacement is additive, and symmetry of the Hessian combines the two cross terms. Applying the triangle inequality and the remainder bound to the three branches proves the stated approximation-error bound.

The base gradient g_0 changes the expansion point b. It does not appear as another selected loss in the pair decomposition. All g_j in the SGD special case are evaluated at the common pre-update parameter θ, even though the utility expansion is around b.

A.5 Fixed-length SGD blocks

We prove Proposition 1. Write $\ell_0 = L_{\mathrm{c}}$, $\ell_j = \lambda_j L_j$, $\ell_S = \ell_0 + \sum_{j\in S}\ell_j$, $g_0 = \nabla\ell_0(\theta)$, and $A_0 = \nabla^2\ell_0(\theta)$. For each subset put $g_S = g_0 + \sum_{j\in S} g_j$ and $A_S = A_0 + \sum_{j\in S} A_j$. With $c_r = r(r-1)/2$, induction on the step index gives

$$\theta_S^{(r)} = \theta - r\eta g_S + c_r \eta^2 A_S g_S + O(\eta^3). \tag{23}$$

Indeed, expanding the gradient at this iterate yields $\nabla\ell_S(\theta_S^{(r)}) = g_S - r\eta A_S g_S + O(\eta^2)$. The SGD recurrence then replaces c_r by $c_r + r = c_{r+1}$. The formula starts at $r = 0$ and holds uniformly over the finite candidate family through the fixed number m of steps.

Put $v_S = \sum_{j \in S} g_j$ and $B_S = \sum_{j \in S} A_j$. Subtracting the empty endpoint from Eq. (23),

$$\Delta_S = -m\eta v_S + c_m \eta^2 (A_0 v_S + B_S g_0 + B_S v_S) + O(\eta^3). \tag{24}$$

Expanding the fixed utility at b_η gives

$$\begin{aligned} G_\eta(S) = m\eta h_\eta^\top v_S - c_m \eta^2 h_\eta^\top (A_0 v_S + B_S g_0 + B_S v_S) \\ - \frac{m^2 \eta^2}{2} v_S^\top H_\eta v_S + O(\eta^3). \end{aligned} \tag{25}$$

The order-η term is linear in selection indicators, and the order-η^2 terms have degree at most two. Taking the pair difference retains $A_j g_k + A_k g_j$ from $B_S v_S$ and $2 g_j^\top H_\eta g_k$ from the quadratic utility term. This proves Eq. (7). The singleton expression uses $h_\eta = \nabla F(\theta) + O(\eta)$.

For three distinct indices, the third set difference is

$$\begin{aligned} \kappa_\eta(\{j,k,l\}) = G_\eta(\{j,k,l\}) - G_\eta(\{j,k\}) - G_\eta(\{j,l\}) - G_\eta(\{k,l\}) \\ + G_\eta(\{j\}) + G_\eta(\{k\}) + G_\eta(\{l\}). \end{aligned} \tag{26}$$

It annihilates the terms of degree at most two, leaving $O(\eta^3)$. These remainder constants depend on both training-loss and utility derivatives, as well as on m. They are not the utility-only remainder bound for exactly additive displacements.

If $a_j = \nabla F(\theta)^\top g_j$ has nonzero RMS, then $\mathrm{RMS}(\mu) = m\eta\,\mathrm{RMS}(a) + O(\eta^2)$ and $\mathrm{RMS}(d)/\mathrm{RMS}(\mu) = O(\eta)$. The same statement applies to a fixed finite panel with uniform derivative bounds and nonzero pooled leading singleton RMS. A nonzero singleton value at finite η does not by itself establish this first-order condition.

The order need not survive a nonlinear gradient transformation. For example, a first bias-corrected Adam step from zero moments uses $P(g) = g/(|g| + \epsilon)$ componentwise. Its relative displacement is $-\eta[P(g_0 + v_S) - P(g_0)]$, giving

$$\begin{aligned} d_{jk}(\eta) = \eta \nabla F(\theta)^\top [P(g_0 + g_j + g_k) - P(g_0 + g_j) \\ - P(g_0 + g_k) + P(g_0)] + O(\eta^2). \end{aligned} \tag{27}$$

A nonzero bracket allows an $O(\eta)$ pair effect; a common fixed linear preconditioner would cancel it. When $|g| \gg \epsilon$, the cold-start map approaches a componentwise sign update. Common nonzero initial moments do not make Adam a fixed linear map: its next second moment contains $\beta_2 v_0 + (1 - \beta_2) g_S^2$, which still depends on the selected set. Cancellation requires a common linear (or affine) gradient transformation, not merely a warm optimizer state. The ten-step audit evaluates only cold-start Adam; the first-step formula provides a possible mechanism, not a measured decomposition of that block.

A.6 Multi-step blocks and nonadditive update rules

For a general update rule, put $e_{jk} = \Delta_{jk} - \Delta_j - \Delta_k$. Subtracting the two singleton expansions in Eq. (15) from the pair expansion remains valid for actual branch displacements. Expanding the quadratic term further gives

$$\begin{aligned} d_{jk} = -\Delta_j^\top H \Delta_k - h^\top e_{jk} - (\Delta_j + \Delta_k)^\top H e_{jk} - \tfrac{1}{2} e_{jk}^\top H e_{jk} \\ + R_j + R_k - R_{jk}. \end{aligned} \tag{28}$$

Thus even a locally linear utility can produce a pair effect through the optimizer if $e_{jk} \neq 0$. Adaptive preconditioning, clipping the combined gradient, or normalizing corrections by set size can introduce such a residual. Multiple plain-SGD steps can also produce $e_{jk} \neq 0$, because later gradients are evaluated at different branch parameters. This formula uses the actual displacement rather than assuming that these operations preserve additivity.

A.7 Higher-order set residual

Under additive displacements, let

$$q(S) = -h^\top \Delta_S - \tfrac{1}{2} \Delta_S^\top H \Delta_S. \tag{29}$$

Since $\Delta_S = \sum_{j \in S} \Delta_j$, the function q is quadratic in the selection indicators. For $s = |S| \geq 2$, its exact pair reconstruction is

$$q(S) = \sum_{j<k \in S} q(\{j,k\}) - (s-2) \sum_{j \in S} q(\{j\}). \tag{30}$$

Each singleton contribution is present in $s-1$ pairs; subtracting $s-2$ copies leaves it once. Each cross term is present in precisely one pair. Using $G(S) = q(S) - R_S$ and the definitions of μ and d therefore yields

$$\kappa(S) = -R_S + \sum_{j<k \in S} R_{jk} - (s-2) \sum_{j \in S} R_j. \tag{31}$$

For $s \geq 3$, one resulting bound is

$$|\kappa(S)| \leq \frac{\rho}{6} \left[\|\Delta_S\|^3 + \sum_{j<k \in S} \|\Delta_{jk}\|^3 + (s-2) \sum_{j \in S} \|\Delta_j\|^3 \right]. \tag{32}$$

For $s \leq 2$, the exact identity $\kappa(S) = 0$ is stronger. Without additive displacements, the quadratic part can itself contain higher-order set effects, so this remainder-only expression is not applicable.

A.8 Uniform error and utility regret

Proposition 3 (Value error and selection regret)**.** *If* $\sup_{S \in \mathcal{S}_B} |\widehat{G}(S) - G(S)| \leq E_B$ *and* $\max_S \widehat{G}(S) - \widehat{G}(\widehat{S}) \leq \delta$*, then* $G(S^*) - G(\widehat{S}) \leq 2E_B + \delta$*. Coefficient-error bounds* $\varepsilon_1, \varepsilon_2$ *give the valid choice* $E_B = B\varepsilon_1 + \binom{B}{2}\varepsilon_2 + \max_{S \in \mathcal{S}_B} |\kappa(S)|$.

For a true maximizer S^* and the selected set $\widehat{S}$, add and subtract their predicted scores:

$$G(S^*) - G(\widehat{S}) = G(S^*) - \widehat{G}(S^*) + \widehat{G}(S^*) - \widehat{G}(\widehat{S}) + \widehat{G}(\widehat{S}) - G(\widehat{S}) \tag{33}$$

$$\leq E_B + \delta + E_B. \tag{34}$$

For any feasible set of size s, coefficient-wise bounds and the definition of κ give

$$|\widehat{G}(S) - G(S)| \leq s\varepsilon_1 + \binom{s}{2}\varepsilon_2 + |\kappa(S)|. \tag{35}$$

Maximizing over $s \leq B$ proves the stated expression for E_B. The guarantee requires a uniform bound over the feasible family. A small average error can coexist with a large error at the set selected by maximization, so replacing the supremum by average MAE is invalid.

B Correction and evaluation protocol

B.1 Data collection and model pretraining

Each task uses 2,000 trajectories of 1,000 environment control steps. Of these, 400 use uniformly sampled legal actions, 800 use the mean action of a SAC policy at 200,000 training steps, and 800 use its mean action at 1,000,000 steps (Haarnoja et al., 2018). The two SAC collectors add Gaussian exploration noise. This gives two million recorded control transitions per task.

Each base model is trained for 100,000 parameter updates per seed using 256 trajectory windows per minibatch and only L_0. The task-specific GRU-512 teacher is trained for 300,000 updates with 256 windows, minimizing the average open-loop state-prediction MSE across depths one through eight. The reward readout receives 100,000 updates with batches of 256 state–action–reward samples, minimizing MSE against the observed immediate reward. Teacher and reward parameters are then frozen and shared across all base seeds, backbones, and selectors within the task.

B.2 Residual dynamics and branch state

For state dimension p and action dimension q, the feedforward model is

$$f_\theta(z,a) = W_3 \,\mathrm{SiLU}\big(W_2 \,\mathrm{SiLU}(W_1[z;a] + b_1) + b_2\big) + b_3. \tag{36}$$

Both hidden widths are 256, and the linear output has dimension p. The recurrent model has a single-layer, unidirectional GRU-256:

$$h_{t+1} = \mathrm{GRU}_\theta([z_t; a_t], h_t), \quad f_\theta(z_t, a_t, h_t) = W_o h_{t+1} + b_o. \tag{37}$$

Both models output Δz_t and use the residual state update $\widehat{z}_{t+1} = \widehat{z}_t + \Delta z_t$ during open-loop prediction. Training and candidate updates use float32, zero dropout, and no batch or layer normalization. Feature standardization is fitted on training data and frozen for correction, utility evaluation, and deployment.

Utility evaluation uses float64 throughout. Before each utility rollout, the teacher and candidate transitions, inputs, and GRU hidden states are converted to float64. The frozen reward readout, including its sigmoid, also runs in float64. Discounted reward accumulation, reference standard deviation and scale floor, logits, and stable log-softmax cross-entropy are computed in float64, retaining per-case losses before reduction. Paired differences are formed per case and then averaged with the fixed weights. Raw μ, d, κ and ridge solutions are computed and stored separately in float64. This is not a conversion of already reduced float32 utilities. Double-precision evaluation reduces cancellation in utility differences; it does not remove rounding in float32 updates.

The GRU prefix consists of up to four recorded state–action steps, processed from a zero hidden state. When fewer steps are available, only that available prefix is processed. Candidate branches and CEM samples copy the same initial history and evolve independently. Each candidate recomputes the prefix under its own parameters; loss gradients pass through this computation, while recorded inputs and targets are detached. Thus history is a fixed model input and not a candidate-dependent selection of utility cases.

With $M_{\mathrm{tr}} = 256$ trajectory windows, the loss normalization is

$$L_j = \frac{1}{pM_{\mathrm{tr}}} \sum_i \|\widehat{z}_{i,j} - z_{i,j}\|_2^2, \quad j \in \mathcal{D}, \qquad L_0 = \frac{1}{pM_{\mathrm{tr}}} \sum_i \|\widehat{z}_{i,1} - z_{i,1}\|_2^2. \tag{38}$$

There is no factor $1/2$. The trajectory targets are fixed while the model rolls out along their recorded actions.

B.3 Auxiliary losses and update blocks

Let $\widehat{r}_{\psi_1}$ be a trainable reward head, distinct from the frozen utility readout $\bar{r}$. Reward consistency is

$$L_{\mathrm{rew}} = \frac{1}{M_{\mathrm{tr}} H} \sum_i \sum_{h=0}^{H-1} [\widehat{r}_{\psi_1}(\widehat{z}_{i,h}, a_{i,h}) - r_{i,h}]^2. \tag{39}$$

For $h \geq 1$, its gradient enters the transition through predicted states. An inverse-dynamics head $\widehat{a}_{\psi_2}$ defines

$$L_{\mathrm{inv}} = \frac{1}{qM_{\mathrm{tr}} H} \sum_i \sum_{h=0}^{H-1} \|\widehat{a}_{\psi_2}(\widehat{z}_{i,h}, \widehat{z}_{i,h+1}) - a_{i,h}\|_2^2. \tag{40}$$

Both objectives are enabled with weight 0.1. Their parameters belong to the cloned optimization state; the teacher and utility readout do not. The present comparisons hold these auxiliary objectives fixed and do not estimate their individual contributions.

Starting from a common $\theta^{(0)}$, each candidate computes

$$\theta_S^{(r+1)} = \theta_S^{(r)} - 10^{-3} \nabla \left[L_{\mathrm{c}} + \frac{1}{5} \sum_{j \in S} L_j \right] (\theta_S^{(r)}), \qquad r = 0, \ldots, 9. \tag{41}$$

The 256-window minibatch stays fixed for the ten steps and is shared across all probes at that update state. A committed-block counter increments once per retained block, not per internal SGD step or probe. Empty-set updates retain L_{c}, including both auxiliaries. All-five replaces the selected set by $\mathcal{D}$, with per-depth weight 0.20 in the per-depth-weight control and 0.12 in the equal-total-weight control. The latter matches the coefficient sum of an exact-three update at weight 0.20 per depth, but does not match gradient norm.

Strict-three Quadratic enumerates the ten triples instead of $\mathcal{S}_3$. Random-three samples a triple uniformly at each refresh and reuses it for eight commits. Fixed controls always use $\{2, 4, 6\}$ or $\{4, 6, 8\}$; both are reported. Initial-gradient alignment scores each weighted depth-loss gradient by its inner product with the fitting utility gradient at the common pre-update point, then takes the top three. It neither measures pair labels nor refreshes this gradient score within the ten-step block.

B.4 Reference, fitting, and independent utility audit

The frozen teacher is one single-layer GRU-512 residual dynamics model per task. The task-specific reward readout is

$$\bar{r}(z, a) = \sigma\left(V_3\,\mathrm{SiLU}\big(V_2\,\mathrm{SiLU}(V_1[z; a] + c_1) + c_2\big) + c_3\right), \tag{42}$$

with two 256-unit hidden layers and a scalar output. Its parameters are fixed for both teacher and student rollouts. Uniform menu actions respect the task's action bounds. Teacher and menu construction use horizon eight; discounting is 0.99 without a terminal-value term. Ties in the reference's preferred action sequence are resolved by menu index.

The reference-only energy scale uses the population standard deviation over 16 menu entries:

$$s_i = \max\left\{\left[\frac{1}{16}\sum_{k=1}^{16}(E_{\mathrm{ref}}(i,k) - \overline{E}_{\mathrm{ref}}(i))^2\right]^{1/2}, 10^{-3}\right\}. \tag{43}$$

The reference, target, menu, scale, and tail weights are excluded from differentiation with respect to student parameters. The fixed upper-loss tail is the 13 highest base-only losses out of 128, with case-index tie breaking and equal weights. It differs from fractional CVaR and from a tail reselected after each candidate update. Fitting and audit cases use disjoint trajectories; each batch defines its own tail at the same base-only branch.

Quadratic and Additive-16 fit their respective features on the same 16-label family at a common state. The empty branch both defines the reference endpoint b and contributes the zero label. For n labels, the ridge normal equations are $(X^\top X + n\,10^{-3} I)\widehat{\beta} = X^\top y$. Additive-6 uses the six empty/singleton labels, not pair labels. All selectors enumerate $\mathcal{S}_3$ and use identical tie breaking.

The independent audits measure every set in $\mathcal{S}_5$ at each of the three refresh states on the Quadratic training trajectory. Thus there are $5 \times 3 \times 3 = 45$ common states per task–backbone setting. Fitting, audit-A, and audit-B batches are trajectory-disjoint. Audit labels evaluate prediction and utility regret, not fitting or online selection. The two audit oracles maximize over 26 or 32 measured gains, both including zero for the empty set. They are fixed-state diagnostics with extra information, not deployed policies. Four- and five-depth branches keep weight 0.20 per depth, unlike the equal-total-weight return control.

Each audit fixes its reference, menus, scales, and base-only tail across candidates. Full-case prediction comparisons refit on full-case fitting labels and evaluate on full-case audit labels. Repeated-audit selected-gain comparisons instead retain the tail-fitted choice and evaluate it under both aggregations. No audit-dependent refitting is performed.

B.5 Return aggregation and statistical comparison

For sorted episode scores $x_{(1)} \le \cdots \le x_{(64)}$, the lower-tail score of one correction run is

$$\widehat{\mathrm{CVaR}}_{0.1}(x) = \frac{\sum_{i=1}^{6} x_{(i)} + 0.4\,x_{(7)}}{6.4}. \tag{44}$$

This fractional definition handles a noninteger tail count (Rockafellar & Uryasev, 2002). We average three correction run CVaRs within each base, then the five base means. A paired contrast is a difference of these method-level CVaRs, not CVaR of episode-wise differences. Every method–setting–endpoint cell contains $5 \times 3 \times 64 = 960$ episodes. The combined endpoint table has 49,920 episode records, including the strict-three and at-most-three Quadratic control with identical outcomes. Repeated or identical endpoints are not additional independent evidence.

For each bootstrap replicate, sample five base occurrences with replacement. Within each occurrence, sample three correction run occurrences; within each correction run occurrence, sample 64

Algorithm 1 Correction training with periodic joint-value refresh

Require: Initial parameters, losses, fixed block operator, $B = 3$

for $u = 0, \ldots, 23$ **do**
  Construct the current update state Ω_u
  if $u \in \{0, 8, 16\}$ **then**
    Compute $b = U_{\varnothing}(\Omega_u)$ and fix the fitting utility
    Set $y_{\varnothing} = 0$
    for S in the nonempty fitting family **do**
      Reset to Ω_u; compute $y_S = F(b) - F(U_S(\Omega_u))$
    end for
    Fit the selected feature family by Eq. (5)
    Choose $\widehat{S} = \arg\max_{S \in \mathcal{S}_3} \widehat{G}(S)$
  end if
  Reset to Ω_u; commit $U_{\widehat{S}}(\Omega_u)$
end for

episodes. Paired methods use the same indices. Repeated draws of an outer cluster receive independent inner draws. Each replicate recomputes fractional CVaR before hierarchical averaging. The 20,000 draws use the resampling seed recorded for each contrast.

The four primary contrast families contain four equal-weight, eight weight-diagnostic, sixteen trained-baseline, and four budget-base comparisons. Percentile endpoints use $0.05/(2K)$ and $1 - 0.05/(2K)$ for family size K. The complete 32-row comparison tables give the confidence level of every interval. This is not one simultaneous-coverage statement over all 32 comparisons. An additional post-hoc family contains 12 contrasts: All-five (0.20) minus at-most-three Quadratic at 24 blocks, and continued Additive-6 or All-five (0.20) minus Quadratic at the recorded budget, each in four settings. This family uses 99.583333% intervals with 20,000 draws and the recorded resampling seed. The three resampling levels use independent streams; compared methods share indices within a replicate. The primary intervals are not pooled with these post-hoc intervals. Table 8 reports all 12. Mean scores, selection frequencies, and comparisons without reported intervals are descriptive.

An exact two-sided sign-flip test enumerates all 2^5 sign assignments to the paired base-level mean differences. Under this five-base design, its minimum possible p-value is $2/32 = 0.0625$; zero paired differences can further reduce resolution. We report raw and Bonferroni-adjusted values within each family. No test rejects at 5%. These tests and hierarchical percentile intervals are not dual: an interval excluding zero is not rejection by the sign-flip test. Inference is conditional on the shared teacher/readout and fixed case panels; each setting has five independent base clusters.

Prediction RMSE is the square root of the pooled mean squared error, not the average of state-level RMSEs. A setting contains 450 triples, 225 four-depth sets, and 45 five-depth sets per audit and target. Descriptive means aggregate 45 states, nested within five bases and three runs. Perturbation comparisons first average the ten antithetic direction pairs within each state, then aggregate through correction run and base. Their 95% intervals are individual descriptive intervals, separate from the return-comparison families.

B.6 Recorded-cost endpoints

The online budget sums recorded costs of candidate updates, commits, teacher-menu construction, fitting-utility evaluation, state copying, fitting, selection, and required I/O. Independent utility audits and final episode evaluation do not affect selection and are excluded. Costs accumulate along each method's own trajectory. This budget analysis uses the online-cost entries, not update counts alone; it makes no accelerator-latency or hardware-throughput claim.

For each paired correction run, the budget is Quadratic's cost after its 24th commit. Base-only trains L_c; Additive-6 and All-five retain their correction rules. Each control is evaluated at its last completed checkpoint no later than that budget, without a 24-block cap. A partially completed update is not an endpoint. The checkpoint time can precede the total time spent within the budget, because selection work after the last completed commit may already have begun or finished.

Selection frequencies count a refresh once only if its decision is used by at least one completed commit included in the endpoint. They exclude decisions finishing after the budget and decisions finishing within it but unused by an included commit. Frequencies are refresh-weighted, not commit-weighted. There are 180 used Quadratic and 304 used Additive-6 refreshes; 43 other Additive-6 decisions finish too late and 13 finish but remain unused. All-five is a fixed prescription and base-only has no selector, so neither contributes to this denominator.

C Complete comparisons and selection diagnostics

The fixed-update endpoint table reports all evaluated methods, including both All-five weights and both fixed triples. The strict-three and at-most-three Quadratic controls have identical episode outcomes in this panel, but remain distinct method definitions. The four contrast tables report all 32 primary comparisons in their stated direction. Positive contrasts favor the named left-hand method; not every row is written as Quadratic minus control. Confidence levels and exact sign-flip p-values are reported separately.

Table 3: Complete fixed-update endpoints. Mean reward and lower-tail return use the same 64-episode evaluation panels.

Method	W/MLP		W/GRU		C/MLP		C/GRU	
	Mean	CVaR	Mean	CVaR	Mean	CVaR	Mean	CVaR
Quadratic	0.6880	0.3890	0.6736	0.3944	0.6170	0.3208	0.6220	0.3386
Quadratic$_{=3}$	0.6880	0.3890	0.6736	0.3944	0.6170	0.3208	0.6220	0.3386
Additive-16	0.6880	0.3890	0.6736	0.3944	0.6168	0.3205	0.6220	0.3385
Additive-6	0.6880	0.3890	0.6736	0.3944	0.6168	0.3205	0.6220	0.3385
All-five (0.12)	0.6767	0.3744	0.6607	0.3781	0.6045	0.3059	0.6114	0.3278
All-five (0.20)	0.6955	0.3985	0.6832	0.4070	0.6272	0.3324	0.6320	0.3501
Random three	0.6750	0.3716	0.6595	0.3774	0.6021	0.3028	0.6107	0.3267
Fixed $\{2, 4, 6\}$	0.6741	0.3719	0.6580	0.3752	0.5980	0.2971	0.6042	0.3204
Fixed $\{4, 6, 8\}$	0.6758	0.3738	0.6607	0.3787	0.6062	0.3083	0.6070	0.3230
Initial-gradient	0.6880	0.3891	0.6736	0.3944	0.6167	0.3205	0.6220	0.3385

Every cell averages 15 correction runs nested within five base seeds. Quadratic$_{=3}$ restricts selection to exactly three depths. All-five weights are per depth.

Table 4: Weight-matched lower-tail-return contrasts.

Comparison	Setting	Δ	Bootstrap interval	CI (%)	p	p_B
All-five (0.12) − Quadratic$_{=3}$	W/MLP	-0.0146	[-0.0257, -0.0076]	98.75	0.0625	0.2500
All-five (0.12) − Quadratic$_{=3}$	W/GRU	-0.0163	[-0.0290, -0.0086]	98.75	0.0625	0.2500
All-five (0.12) − Quadratic$_{=3}$	C/MLP	-0.0149	[-0.0295, -0.0068]	98.75	0.0625	0.2500
All-five (0.12) − Quadratic$_{=3}$	C/GRU	-0.0108	[-0.0177, -0.0049]	98.75	0.0625	0.2500

Intervals are paired hierarchical percentile-bootstrap intervals (20,000 replicates). The exact two-sided sign-flip test uses five independent base-level differences; p_B applies Bonferroni correction within the specified comparison family. The interval and exact test are different procedures, not dual summaries. The design minimum two-sided p is 0.0625; no exact family test rejects at 0.05.

Table 5: Weight-diagnostic lower-tail-return contrasts.

Comparison	Setting	Δ	Bootstrap interval	CI (%)	p	p_B
All-five (0.12) − Quadratic	W/MLP	-0.0146	[-0.0268, -0.0073]	99.375	0.0625	0.5000
All-five (0.20) − All-five (0.12)	W/MLP	0.0241	[0.0131, 0.0379]	99.375	0.0625	0.5000
All-five (0.12) − Quadratic	W/GRU	-0.0163	[-0.0303, -0.0081]	99.375	0.0625	0.5000
All-five (0.20) − All-five (0.12)	W/GRU	0.0289	[0.0175, 0.0445]	99.375	0.0625	0.5000
All-five (0.12) − Quadratic	C/MLP	-0.0149	[-0.0311, -0.0064]	99.375	0.0625	0.5000
All-five (0.20) − All-five (0.12)	C/MLP	0.0265	[0.0156, 0.0428]	99.375	0.0625	0.5000
All-five (0.12) − Quadratic	C/GRU	-0.0108	[-0.0183, -0.0045]	99.375	0.0625	0.5000
All-five (0.20) − All-five (0.12)	C/GRU	0.0222	[0.0140, 0.0330]	99.375	0.0625	0.5000

Intervals are paired hierarchical percentile-bootstrap intervals (20,000 replicates). The exact two-sided sign-flip test uses five independent base-level differences; p_B applies Bonferroni correction within the specified comparison family. The interval and exact test are different procedures, not dual summaries. The design minimum two-sided p is 0.0625; no exact family test rejects at 0.05.

Table 6: Exactly-three selector contrasts after 24 update blocks.

Comparison	Setting	Δ	Bootstrap interval	CI (%)	p	p_B
Quadratic$_{=3}$ − Random three	W/MLP	0.0175	[0.0067, 0.0348]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{2, 4, 6\}$	W/MLP	0.0171	[0.0061, 0.0284]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{4, 6, 8\}$	W/MLP	0.0152	[0.0034, 0.0325]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Initial-gradient	W/MLP	-0.0000	[-0.0002, 0.0000]	99.6875	1.0000	1.0000
Quadratic$_{=3}$ − Random three	W/GRU	0.0169	[0.0090, 0.0344]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{2, 4, 6\}$	W/GRU	0.0192	[0.0065, 0.0444]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{4, 6, 8\}$	W/GRU	0.0157	[0.0047, 0.0329]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Initial-gradient	W/GRU	0.0000	[0.0000, 0.0000]	99.6875	1.0000	1.0000
Quadratic$_{=3}$ − Random three	C/MLP	0.0180	[0.0043, 0.0424]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{2, 4, 6\}$	C/MLP	0.0238	[0.0057, 0.0516]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{4, 6, 8\}$	C/MLP	0.0125	[-0.0000, 0.0282]	99.6875	0.1250	1.0000
Quadratic$_{=3}$ − Initial-gradient	C/MLP	0.0004	[-0.0000, 0.0019]	99.6875	0.5000	1.0000
Quadratic$_{=3}$ − Random three	C/GRU	0.0119	[0.0047, 0.0198]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{2, 4, 6\}$	C/GRU	0.0182	[0.0059, 0.0356]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Fixed $\{4, 6, 8\}$	C/GRU	0.0156	[0.0036, 0.0354]	99.6875	0.0625	1.0000
Quadratic$_{=3}$ − Initial-gradient	C/GRU	0.0001	[0.0000, 0.0006]	99.6875	1.0000	1.0000

Intervals are paired hierarchical percentile-bootstrap intervals (20,000 replicates). The exact two-sided sign-flip test uses five independent base-level differences; p_B applies Bonferroni correction within the specified comparison family. The interval and exact test are different procedures, not dual summaries. The design minimum two-sided p is 0.0625; no exact family test rejects at 0.05.

Table 7: Quadratic minus base-only at the recorded-cost budget.

Comparison	Setting	Δ	Bootstrap interval	CI (%)	p	p_B
Quadratic − Base-only	W/MLP	-0.0036	[-0.0226, 0.0145]	98.75	0.4375	1.0000
Quadratic − Base-only	W/GRU	-0.0105	[-0.0289, 0.0073]	98.75	0.1875	0.7500
Quadratic − Base-only	C/MLP	-0.0002	[-0.0184, 0.0150]	98.75	0.8750	1.0000
Quadratic − Base-only	C/GRU	-0.0105	[-0.0285, 0.0103]	98.75	0.2500	1.0000

Intervals are paired hierarchical percentile-bootstrap intervals (20,000 replicates). The exact two-sided sign-flip test uses five independent base-level differences; p_B applies Bonferroni correction within the specified comparison family. The interval and exact test are different procedures, not dual summaries. The design minimum two-sided p is 0.0625; no exact family test rejects at 0.05.

Table 8: Additional post-hoc return comparisons. Differences are comparator minus Quadratic in lower-tail return (return divided by 1,000). Budget rows share Quadratic's 24-block recorded online cost. All 12 comparisons form one family; intervals are 99.5833% paired hierarchical percentile bootstrap intervals from 20,000 replicates. Every row has five positive paired base differences, exact two-sided sign-flip $p = 0.0625$, and Bonferroni-adjusted $p = 0.75$. The bootstrap intervals and exact tests are distinct procedures. W/C denote Walker/Cheetah.

Setting	Difference	Lower bound	Upper bound
All-five (0.20) − Quadratic, 24 blocks			
W/MLP	0.00946	0.00317	0.01664
W/GRU	0.01262	0.00572	0.02276
C/MLP	0.01160	0.00543	0.01974
C/GRU	0.01148	0.00497	0.01804
Additive-6 − Quadratic, budget			
W/MLP	0.02823	0.01466	0.04610
W/GRU	0.03646	0.01874	0.05589
C/MLP	0.03012	0.01753	0.04832
C/GRU	0.02782	0.01893	0.03861
All-five (0.20) − Quadratic, budget			
W/MLP	0.04182	0.02153	0.06523
W/GRU	0.05584	0.02873	0.08319
C/MLP	0.04524	0.02643	0.06696
C/GRU	0.04400	0.02828	0.06168

C.1 Independent triple audit and raw set effects

Table 9 separates prediction accuracy from selection regret. The latter compares each fitted choice with the best measured audit gain at its common state. Quadratic and Additive-16 have identical mean audit regret in both Walker settings. Quadratic has marginally lower regret in Cheetah/MLP and higher regret in Cheetah/GRU, matching the directions of the two differing choices.

Table 9: Independent triple prediction and 26-set utility regret. Prediction errors aggregate 450 triples per setting; regret averages 45 common states. All quantities are in teacher-relative utility units, with lower better. Regret entries are descriptive.

Task	Model / predictor	Triple RMSE	Triple MAE	Audit regret
Walker	MLP / Quadratic	.004956	.003485	.000341
	MLP / Additive-16	.004866	.003438	.000341
	GRU / Quadratic	.005465	.003952	.000445
	GRU / Additive-16	.005412	.003916	.000445
Cheetah	MLP / Quadratic	.004570	.003466	.000362
	MLP / Additive-16	.004488	.003391	.000364
	GRU / Quadratic	.005100	.003721	.000979
	GRU / Additive-16	.005077	.003700	.000925

The raw pair-effect RMS is 0.574%, 0.703%, 0.722%, and 0.535% of singleton-gain RMS in Walker/MLP, Walker/GRU, Cheetah/MLP, and Cheetah/GRU, respectively. Table 10 reports the underlying scales alongside fitted coefficients and triple residuals. These are finite differences of the fixed audit utility, not estimates of Hessian derivatives.

Table 10: Raw audit effects and fitted coefficients, in displayed utility scale factors. Raw RMS pools five singleton gains, ten pair effects, or ten triple residuals per state over 45 audit states. Fitted coefficients use the fitting split; Table 23 separately compares raw and fitted effects on that same split.

Setting	Singleton gain (10^{-3})	Raw pair effect (10^{-5})	Fitted pair (10^{-4})	Triple κ (10^{-7})
Walker/MLP	5.327	3.059	1.247	1.689
Walker/GRU	5.846	4.112	1.256	2.560
Cheetah/MLP	6.614	4.773	1.447	2.896
Cheetah/GRU	5.940	3.176	1.222	1.746

C.2 Selection margins and composition

Table 11 gives the complete certificate counts. Every certified state has the same additive and pair maximizer. The two differing states have additive margins smaller than twice the maximum score perturbation, as required by the sufficient condition. This post-hoc calculation uses both fitted models and therefore adds no claim of probe savings.

Table 11: Selection stability at common update states. Each setting contains 45 states, nested within five bases and three runs per base. Uncertified agreement means the choices coincide although the sufficient margin condition does not hold.

Setting	Certified agreement	Uncertified agreement	Different	Total
Walker/MLP	28	17	0	45
Walker/GRU	35	10	0	45
Cheetah/MLP	35	9	1	45
Cheetah/GRU	32	12	1	45
Total	130	48	2	180

At the differing Cheetah/MLP state, Quadratic selects $\{2, 3, 6\}$ and Additive-16 selects $\{3, 4, 6\}$. Their audit gains are 0.01025470 and 0.01019614. At the differing Cheetah/GRU state, Quadratic selects $\{4, 6, 8\}$ and Additive-16 selects $\{2, 6, 8\}$, with gains 0.01397114 and 0.01640147. These are common-state utility comparisons, not separate closed-loop return measurements.

Table 12: Triple-set selection counts for Quadratic along its own training trajectories. Each setting contains 45 refresh decisions: five bases, three runs, and three refreshes per correction run. All entries count decisions rather than repeated committed blocks. Sets of size zero, one, or two are never selected by Quadratic, Additive-16, or Additive-6.

Selected depths	Walker/MLP	Walker/GRU	Cheetah/MLP	Cheetah/GRU
$\{2, 3, 4\}$	7	2	4	1
$\{2, 3, 6\}$	3	4	4	5
$\{2, 3, 8\}$	3	0	2	3
$\{2, 4, 6\}$	3	3	4	4
$\{2, 4, 8\}$	5	11	0	5
$\{2, 6, 8\}$	5	8	8	3
$\{3, 4, 6\}$	5	4	3	7
$\{3, 4, 8\}$	2	7	1	8
$\{3, 6, 8\}$	9	5	8	5
$\{4, 6, 8\}$	3	1	11	4
Total	45	45	45	45

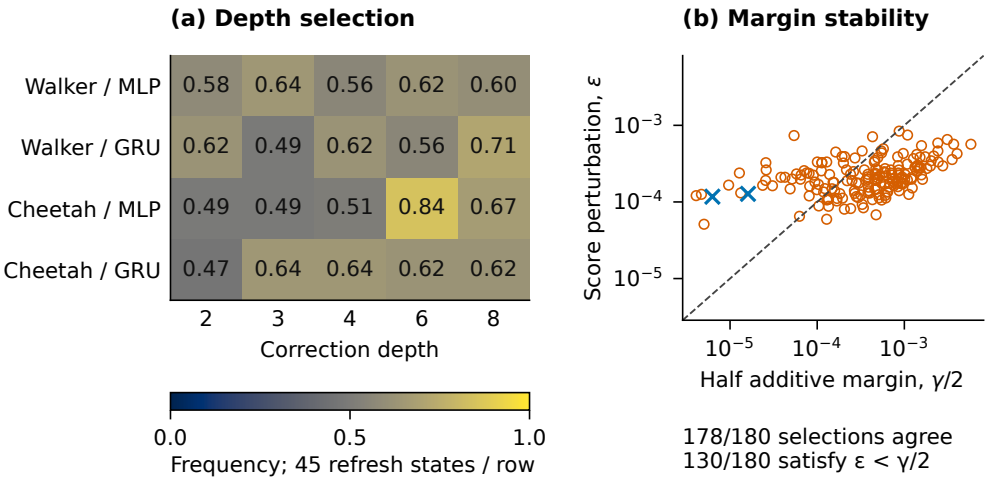

Figure 5: Common-state selection diagnostics. Depth inclusion frequencies use 180 states, not the number of effective refreshes at budget endpoints. The margin diagnostic compares both fitted predictors at the same parameters. Of 178 agreements, 130 satisfy the sufficient certificate; an uncertified state need not have a different decision.

C.3 Two independent audits and actual larger sets

The primary audit extends the 26-set selection panel with five four-depth branches and the five-depth branch. A second independent audit evaluates all 32 branches without changing the fitting split. All-five uses total weight one in these audits. The best-of-32 versus best-of-26 comparison therefore combines a larger candidate family with access to greater total correction weight. Table 13 reports actual oracle gains, while Table 14 reports out-of-fitting-family prediction for sizes four and five.

We distinguish two full-case estimands. For target-fitted prediction, both fitting and evaluation use the 128-case mean. For repeated evaluation of the primary choice, fitting and selection continue to use the 13-case tail, and only the evaluation aggregation changes. Thus the full-case selected gain

Table 13: Candidate-pool expansion evaluated on independent audit batches. Gains are cross-entropy reductions in nats.

Setting	Audit	G^*_{26}	G^*_{32}	Δ	Opt. 4/5 (%)	Quadratic-3	Quadratic-5	Add.-5
tail13: independently target-fitted selectors								
W/MLP	A	0.0158	0.0188	0.0030	8.9/91.1	0.0155	0.0187	0.0187
W/MLP	B	0.0153	0.0183	0.0030	6.7/88.9	0.0148	0.0182	0.0182
W/GRU	A	0.0185	0.0224	0.0039	13.3/86.7	0.0180	0.0223	0.0222
W/GRU	B	0.0187	0.0223	0.0037	11.1/88.9	0.0183	0.0222	0.0221
C/MLP	A	0.0196	0.0237	0.0041	6.7/93.3	0.0193	0.0236	0.0236
C/MLP	B	0.0194	0.0237	0.0042	4.4/95.6	0.0189	0.0236	0.0235
C/GRU	A	0.0172	0.0208	0.0036	13.3/86.7	0.0162	0.0208	0.0207
C/GRU	B	0.0153	0.0184	0.0031	8.9/86.7	0.0145	0.0184	0.0182
full128: independently target-fitted selectors								
W/MLP	A	0.0034	0.0041	0.0007	8.9/88.9	0.0034	0.0041	0.0041
W/MLP	B	0.0034	0.0041	0.0007	6.7/91.1	0.0034	0.0041	0.0041
W/GRU	A	0.0039	0.0047	0.0008	15.6/84.4	0.0038	0.0047	0.0047
W/GRU	B	0.0039	0.0047	0.0008	13.3/86.7	0.0039	0.0047	0.0047
C/MLP	A	0.0042	0.0051	0.0009	4.4/95.6	0.0041	0.0051	0.0051
C/MLP	B	0.0042	0.0051	0.0009	2.2/97.8	0.0042	0.0051	0.0051
C/GRU	A	0.0036	0.0044	0.0008	11.1/86.7	0.0036	0.0044	0.0044
C/GRU	B	0.0034	0.0041	0.0007	11.1/88.9	0.0033	0.0041	0.0041

Each row averages 45 states. Opt. 4/5 gives the frequency of size-four and size-five audit optima. Quadratic/Additive still fit only the 16 empty, singleton and pair labels. B=5 choices are offline candidate selections, not additional return runs. Full128 uses full128-fit labels and a newly selected set.

Table 14: Four- and five-depth prediction error and raw higher-order interaction magnitude.

Setting	Audit	Quadratic RMSE4	Add. RMSE4	I_4 RMS	Quadratic RMSE5	Add. RMSE5	I_5 RMS
tail13							
W/MLP	A	0.0063	0.0061	2.34×10^{-9}	0.0077	0.0073	2.33×10^{-11}
W/MLP	B	0.0055	0.0052	3.95×10^{-9}	0.0067	0.0062	7.93×10^{-11}
W/GRU	A	0.0071	0.0069	4.31×10^{-9}	0.0087	0.0085	9.27×10^{-11}
W/GRU	B	0.0064	0.0063	2.68×10^{-9}	0.0078	0.0075	5.95×10^{-11}
C/MLP	A	0.0058	0.0056	3.92×10^{-9}	0.0072	0.0067	7.47×10^{-11}
C/MLP	B	0.0057	0.0055	3.65×10^{-9}	0.0071	0.0065	9.81×10^{-11}
C/GRU	A	0.0064	0.0064	2.16×10^{-9}	0.0078	0.0077	5.35×10^{-11}
C/GRU	B	0.0072	0.0070	1.36×10^{-9}	0.0089	0.0084	2.24×10^{-11}
full128							
W/MLP	A	0.0010	0.0010	5.63×10^{-10}	0.0012	0.0012	5.36×10^{-12}
W/MLP	B	0.0008	0.0008	7.24×10^{-10}	0.0010	0.0010	8.28×10^{-12}
W/GRU	A	0.0012	0.0011	6.98×10^{-10}	0.0014	0.0014	1.30×10^{-11}
W/GRU	B	0.0010	0.0010	7.49×10^{-10}	0.0013	0.0012	7.84×10^{-12}
C/MLP	A	0.0010	0.0010	6.63×10^{-10}	0.0012	0.0012	7.82×10^{-12}
C/MLP	B	0.0012	0.0011	6.64×10^{-10}	0.0014	0.0013	1.18×10^{-11}
C/GRU	A	0.0010	0.0010	4.87×10^{-10}	0.0012	0.0012	8.13×10^{-12}
C/GRU	B	0.0009	0.0009	6.10×10^{-10}	0.0012	0.0011	4.80×10^{-12}

RMSE pools the corresponding candidate residuals before taking the square root. Raw interactions are same-state, same-audit Möbius differences of candidate gains, not fitted coefficients. All utility quantities use nats.

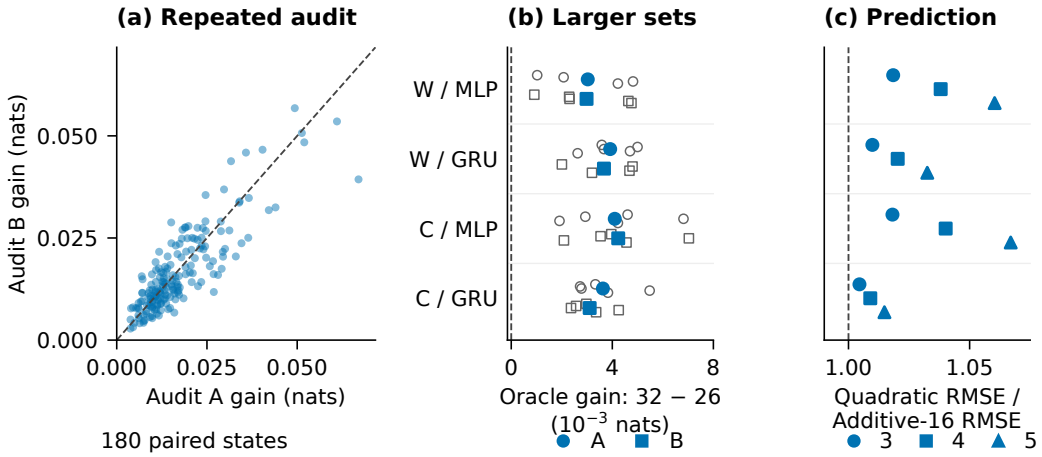

Figure 6: Independent audits of the fixed-tail utility. (a) Audits A and B evaluate the same tail-fitted Quadratic choice at 180 states; the line is identity. (b) Actual oracle-gain increase from 26 to 32 candidates, with five base means (open) and their mean (filled), separately for each audit. (c) Audit-A ridge Quadratic/Additive-16 prediction-RMSE ratios on unseen triples, four-depth sets, and five-depth sets. Ratios above one favor additive prediction. This ridge comparison is complemented by zero-penalty fits in Appendix E. All panels are descriptive; states are nested within five bases and three runs per setting.

in the latter comparison is not required to equal that of a full-case-fitted selector. Tail membership is fixed within an audit, not shared between independent batches.

The all-32 rank correlations in Table 15 can partly reflect differences between cardinalities. Restricting the comparison to ten triples gives mean tail Spearman correlations 0.764, 0.813, 0.800, and 0.783 for Walker/MLP, Walker/GRU, Cheetah/MLP, and Cheetah/GRU. The 26-set audit-oracle choices agree between batches in 21, 26, 29, and 27 of 45 states, respectively. This explains how predominantly positive selected gains can coexist with uncertainty about the exact best set. The full-32 prediction summary is in Table 16.

Table 15: Independent audit-batch stability for the fixed primary tail-selected set.

Setting	Target	$\rho_{A,B}$	G_A	G_B	Sign (%)	RMS$(G_A - G_B)$	d_A RMS	d_B RMS
W/MLP	tail13	0.936	0.0155	0.0148	100	0.0038	3.06×10^{-5}	3.08×10^{-5}
W/MLP	full128	0.959	0.0034	0.0033	100	0.0006	9.02×10^{-6}	8.79×10^{-6}
W/GRU	tail13	0.943	0.0180	0.0183	100	0.0044	4.11×10^{-5}	4.11×10^{-5}
W/GRU	full128	0.959	0.0038	0.0039	100	0.0006	1.18×10^{-5}	1.20×10^{-5}
C/MLP	tail13	0.956	0.0193	0.0189	100	0.0041	4.77×10^{-5}	4.47×10^{-5}
C/MLP	full128	0.970	0.0041	0.0041	100	0.0008	1.30×10^{-5}	1.27×10^{-5}
C/GRU	tail13	0.935	0.0162	0.0145	100	0.0047	3.18×10^{-5}	2.97×10^{-5}
C/GRU	full128	0.958	0.0035	0.0033	100	0.0007	9.49×10^{-6}	9.47×10^{-6}

Spearman correlation is calculated over 32 candidates within each state, then averaged. The same tail-fit Quadratic-3 set defines both selected-gain columns for both objectives. Sign agreement includes zero. Each row contains 45 valid state correlations.

For a selected displacement $\Delta = \theta_{\widehat{S}} - b$, the matched-norm comparison evaluates $b \pm \|\Delta\|_2 v_r$ for ten unit directions v_r. Each antithetic pair is averaged before forming the state-level random gain. Candidate, base, and perturbation branches use the same utility target and case weights. This tests the direction of the selected update relative to random directions of equal length; it does not isolate pair features from the depth losses or equate the underlying training objectives. Table 17 reports both aggregations. Its eight individual 95% intervals are descriptive, not part of the four return-comparison families.

Table 16: Prediction RMSE over all 32 candidates on audit A and audit B.

Setting	Fit/evaluation target	Quadratic A	Additive A	Quadratic B	Additive B
W/MLP	tail13	0.0045	0.0044	0.0039	0.0038
W/MLP	full128	0.0007	0.0007	0.0006	0.0006
W/GRU	tail13	0.0050	0.0049	0.0046	0.0045
W/GRU	full128	0.0008	0.0008	0.0007	0.0007
C/MLP	tail13	0.0042	0.0041	0.0041	0.0040
C/MLP	full128	0.0007	0.0007	0.0008	0.0008
C/GRU	tail13	0.0047	0.0046	0.0052	0.0051
C/GRU	full128	0.0007	0.0007	0.0007	0.0007

Unlike the fixed primary selections in Table 15, each full128 prediction model is fitted using full128 fit labels. No audit labels enter fitting.

Table 17: Selected corrections versus antithetic norm-matched random parameter perturbations.

Setting	Target	Selected gain	Random gain	Difference	95% interval	Random + (%)	Norm resid.
W/MLP	tail13	0.0155	-5.42×10^{-5}	0.0155	[0.0108, 0.0209]	49.9	7.67×10^{-16}
W/MLP	full128	0.0034	-3.61×10^{-5}	0.0034	[0.0024, 0.0046]	49.0	7.67×10^{-16}
W/GRU	tail13	0.0180	-6.80×10^{-5}	0.0181	[0.0139, 0.0228]	49.6	7.95×10^{-16}
W/GRU	full128	0.0038	-4.75×10^{-5}	0.0038	[0.0029, 0.0050]	49.6	7.95×10^{-16}
C/MLP	tail13	0.0193	-7.63×10^{-5}	0.0193	[0.0133, 0.0273]	49.9	6.75×10^{-16}
C/MLP	full128	0.0041	-5.30×10^{-5}	0.0042	[0.0029, 0.0060]	49.3	6.75×10^{-16}
C/GRU	tail13	0.0162	-5.38×10^{-5}	0.0162	[0.0111, 0.0224]	49.7	9.08×10^{-16}
C/GRU	full128	0.0035	-3.80×10^{-5}	0.0035	[0.0023, 0.0049]	49.4	9.08×10^{-16}

Ten directions and their negatives are averaged within each state. Intervals resample base and correction run units while retaining the state/direction panel. Both targets evaluate the primary tail-selected correction. Norm resid. is the relative numerical residual of norm matching, not prediction error.

C.4 Update-operator sensitivity

We vary one update setting at a time: learning rates $\{0.0003, 0.001, 0.003, 0.01\}$ at ten SGD steps; step counts $\{1, 10, 30, 100\}$ at learning rate 0.001; and ten Adam steps at 0.001. The common default appears only once. Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, with fresh optimizer state cloned across branches, no weight decay, and no clipping. All configurations use the same 256 training windows.

Let q index the update configuration and q_0 the default. The utility anchor, including its tail members, is fixed at q_0, while each configuration has its own base-only endpoint:

$$G_q(S) = F_{q_0}(b_q) - F_{q_0}(\theta_{q,S}), \qquad G_q(\varnothing) = 0.$$

Fixing F_{q_0} avoids mixing operator changes with changes in the tail definition. It does not equate b_q across configurations. Both score models refit the 16 fitting labels for each configuration; audits compare their predictions on $\mathcal{S}_3$. The full-case scan uses full-case fitting and audit values throughout.

Tables 18 and 19 report every configuration, including all negative gains. Figure 7 contrasts absolute selected gain with Quadratic-minus-additive gain. With cold-start Adam, large pair effects are accompanied by large third-order residuals. This observation is consistent with nonadditive displacement and higher-order terms in Eq. (28); it does not identify a measured Hessian contribution. No scan configuration is evaluated as an additional closed-loop policy here.

Table 18: Update-operator scan with the tail13 objective.

Setting	Configuration	G_P	$G_P - G_A$	Agree (%)	RMSE P	RMSE A	d/μ (%)	κ RMS	$\|\delta\|$
W/MLP	SGD, $m = 10$	0.0155	0.00000	100.0	0.0050	0.0049	0.57	1.69×10^{-7}	0.01534
W/MLP	$\eta = 0.0003$	0.0047	-0.00000	97.8	0.0015	0.0015	0.17	4.61×10^{-9}	0.00465
W/MLP	$\eta = 0.003$	0.0448	0.00000	100.0	0.0145	0.0144	1.72	4.44×10^{-6}	0.04475
W/MLP	$\eta = 0.01$	0.1328	-0.00007	97.8	0.0448	0.0455	5.66	1.50×10^{-4}	0.13543
W/MLP	$m = 1$	0.0016	0.00000	100.0	0.0005	0.0005	0.01	1.11×10^{-10}	0.00156
W/MLP	$m = 30$	0.0447	0.00000	100.0	0.0145	0.0143	1.81	5.24×10^{-6}	0.04461
W/MLP	$m = 100$	0.1313	-0.00007	97.8	0.0443	0.0451	5.99	1.81×10^{-4}	0.13379
W/MLP	Adam	-0.0026	0.00068	4.4	0.0120	0.0122	113.56	1.23×10^{-2}	0.01657
W/GRU	SGD, $m = 10$	0.0180	0.00000	100.0	0.0055	0.0054	0.70	2.56×10^{-7}	0.01769
W/GRU	$\eta = 0.0003$	0.0055	0.00000	100.0	0.0017	0.0016	0.21	6.96×10^{-9}	0.00537
W/GRU	$\eta = 0.003$	0.0522	0.00006	97.8	0.0159	0.0158	2.09	6.79×10^{-6}	0.05136
W/GRU	$\eta = 0.01$	0.1534	-0.00013	97.8	0.0475	0.0480	6.75	2.36×10^{-4}	0.15287
W/GRU	$m = 1$	0.0018	0.00000	100.0	0.0006	0.0005	0.02	1.91×10^{-10}	0.00180
W/GRU	$m = 30$	0.0520	0.00001	95.6	0.0159	0.0158	2.20	8.05×10^{-6}	0.05110
W/GRU	$m = 100$	0.1517	-0.00005	93.3	0.0470	0.0475	7.11	2.82×10^{-4}	0.15053
W/GRU	Adam	-0.0035	-0.00160	6.7	0.0114	0.0121	104.71	1.12×10^{-2}	0.02114
C/MLP	SGD, $m = 10$	0.0193	0.00000	97.8	0.0046	0.0045	0.72	2.90×10^{-7}	0.01836
C/MLP	$\eta = 0.0003$	0.0059	0.00000	100.0	0.0014	0.0014	0.22	7.69×10^{-9}	0.00558
C/MLP	$\eta = 0.003$	0.0555	0.00014	91.1	0.0133	0.0132	2.16	8.06×10^{-6}	0.05316
C/MLP	$\eta = 0.01$	0.1598	-0.00192	88.9	0.0404	0.0411	7.06	2.92×10^{-4}	0.15790
C/MLP	$m = 1$	0.0020	0.00000	100.0	0.0005	0.0005	0.02	2.23×10^{-10}	0.00187
C/MLP	$m = 30$	0.0554	0.00014	91.1	0.0133	0.0132	2.27	9.42×10^{-6}	0.05299
C/MLP	$m = 100$	0.1580	-0.00188	88.9	0.0399	0.0408	7.44	3.41×10^{-4}	0.15581
C/MLP	Adam	-0.0042	-0.00223	2.2	0.0124	0.0119	103.22	1.05×10^{-2}	0.01899
C/GRU	SGD, $m = 10$	0.0162	-0.00005	97.8	0.0051	0.0051	0.53	1.75×10^{-7}	0.01560
C/GRU	$\eta = 0.0003$	0.0049	0.00000	100.0	0.0015	0.0015	0.16	4.79×10^{-9}	0.00475
C/GRU	$\eta = 0.003$	0.0470	-0.00017	95.6	0.0149	0.0149	1.61	4.51×10^{-6}	0.04536
C/GRU	$\eta = 0.01$	0.1400	0.00054	93.3	0.0456	0.0458	5.40	1.45×10^{-4}	0.13598
C/GRU	$m = 1$	0.0017	0.00000	100.0	0.0005	0.0005	0.02	1.22×10^{-10}	0.00159
C/GRU	$m = 30$	0.0469	-0.00017	95.6	0.0149	0.0148	1.71	4.83×10^{-6}	0.04522
C/GRU	$m = 100$	0.1385	0.00054	93.3	0.0450	0.0452	5.76	1.61×10^{-4}	0.13431
C/GRU	Adam	-0.0027	-0.00175	2.2	0.0124	0.0110	118.31	1.46×10^{-2}	0.01544

Each configuration is evaluated at 45 states per setting. Default SGD uses $\eta = 0.001$, $m = 10$, batch 256; listed alternatives change one factor. Models fit 16 labels and audit 26 candidates. The q0 reference/menu/tail anchor is fixed across configurations, while gains use each configuration's own base-only update. d/μ is the raw pair-to-singleton RMS ratio; these candidate audits do not define new return endpoints.

Table 19: Update-operator scan with the full128 objective.

Setting	Configuration	G_P	$G_P - G_A$	Agree (%)	RMSE P	RMSE A	d/μ (%)	κ RMS	$\|\delta\|$
W/MLP	SGD, $m = 10$	0.0034	0.00000	97.8	0.0008	0.0008	0.79	7.00×10^{-8}	0.01550
W/MLP	$\eta = 0.0003$	0.0010	0.00000	100.0	0.0002	0.0002	0.24	1.92×10^{-9}	0.00470
W/MLP	$\eta = 0.003$	0.0096	0.00001	97.8	0.0023	0.0022	2.33	1.81×10^{-6}	0.04503
W/MLP	$\eta = 0.01$	0.0272	-0.00008	95.6	0.0069	0.0072	7.48	5.65×10^{-5}	0.13554
W/MLP	$m = 1$	0.0003	0.00000	100.0	0.0001	0.0001	0.03	2.49×10^{-11}	0.00157
W/MLP	$m = 30$	0.0096	0.00001	97.8	0.0023	0.0022	2.42	2.06×10^{-6}	0.04489
W/MLP	$m = 100$	0.0270	-0.00008	95.6	0.0068	0.0071	7.76	6.40×10^{-5}	0.13389
W/MLP	Adam	-0.0006	-0.00015	6.7	0.0026	0.0018	115.58	2.80×10^{-3}	0.01629
W/GRU	SGD, $m = 10$	0.0038	-0.00001	97.8	0.0009	0.0009	0.94	1.07×10^{-7}	0.01757
W/GRU	$\eta = 0.0003$	0.0012	0.00000	100.0	0.0003	0.0003	0.28	2.93×10^{-9}	0.00535
W/GRU	$\eta = 0.003$	0.0108	0.00000	100.0	0.0026	0.0026	2.76	2.77×10^{-6}	0.05106
W/GRU	$\eta = 0.01$	0.0305	0.00004	88.9	0.0078	0.0080	8.69	8.72×10^{-5}	0.15100
W/GRU	$m = 1$	0.0004	0.00000	100.0	0.0001	0.0001	0.04	3.97×10^{-11}	0.00179
W/GRU	$m = 30$	0.0108	0.00000	100.0	0.0026	0.0026	2.86	3.14×10^{-6}	0.05089
W/GRU	$m = 100$	0.0302	0.00004	88.9	0.0077	0.0079	8.98	9.81×10^{-5}	0.14903
W/GRU	Adam	-0.0007	-0.00038	4.4	0.0022	0.0019	106.23	2.26×10^{-3}	0.01998
C/MLP	SGD, $m = 10$	0.0041	0.00000	100.0	0.0008	0.0008	0.91	1.09×10^{-7}	0.01819
C/MLP	$\eta = 0.0003$	0.0013	0.00000	100.0	0.0002	0.0002	0.28	2.97×10^{-9}	0.00552
C/MLP	$\eta = 0.003$	0.0118	0.00000	100.0	0.0023	0.0023	2.71	2.87×10^{-6}	0.05287
C/MLP	$\eta = 0.01$	0.0329	0.00011	91.1	0.0068	0.0071	8.68	9.31×10^{-5}	0.15740
C/MLP	$m = 1$	0.0004	0.00000	100.0	0.0001	0.0001	0.04	4.59×10^{-11}	0.00185
C/MLP	$m = 30$	0.0117	0.00000	100.0	0.0023	0.0023	2.82	3.27×10^{-6}	0.05269
C/MLP	$m = 100$	0.0326	0.00011	91.1	0.0067	0.0070	9.01	1.05×10^{-4}	0.15532
C/MLP	Adam	-0.0005	-0.00038	4.4	0.0023	0.0020	108.13	2.32×10^{-3}	0.01872
C/GRU	SGD, $m = 10$	0.0036	0.00000	100.0	0.0008	0.0008	0.76	6.57×10^{-8}	0.01580
C/GRU	$\eta = 0.0003$	0.0011	0.00000	100.0	0.0002	0.0002	0.23	1.80×10^{-9}	0.00479
C/GRU	$\eta = 0.003$	0.0102	0.00000	97.8	0.0022	0.0023	2.26	1.71×10^{-6}	0.04582
C/GRU	$\eta = 0.01$	0.0290	-0.00025	86.7	0.0068	0.0068	7.29	5.50×10^{-5}	0.13655
C/GRU	$m = 1$	0.0004	0.00000	100.0	0.0001	0.0001	0.03	2.96×10^{-11}	0.00160
C/GRU	$m = 30$	0.0102	0.00000	100.0	0.0022	0.0022	2.35	1.95×10^{-6}	0.04568
C/GRU	$m = 100$	0.0287	-0.00025	86.7	0.0067	0.0067	7.60	6.31×10^{-5}	0.13487
C/GRU	Adam	-0.0004	0.00004	0.0	0.0024	0.0020	123.47	3.00×10^{-3}	0.01476

Each configuration is evaluated at 45 states per setting. Default SGD uses $\eta = 0.001$, $m = 10$, batch 256; listed alternatives change one factor. Models fit 16 labels and audit 26 candidates. The q0 reference/menu/tail anchor is fixed across configurations, while gains use each configuration's own base-only update. d/μ is the raw pair-to-singleton RMS ratio; these candidate audits do not define new return endpoints.

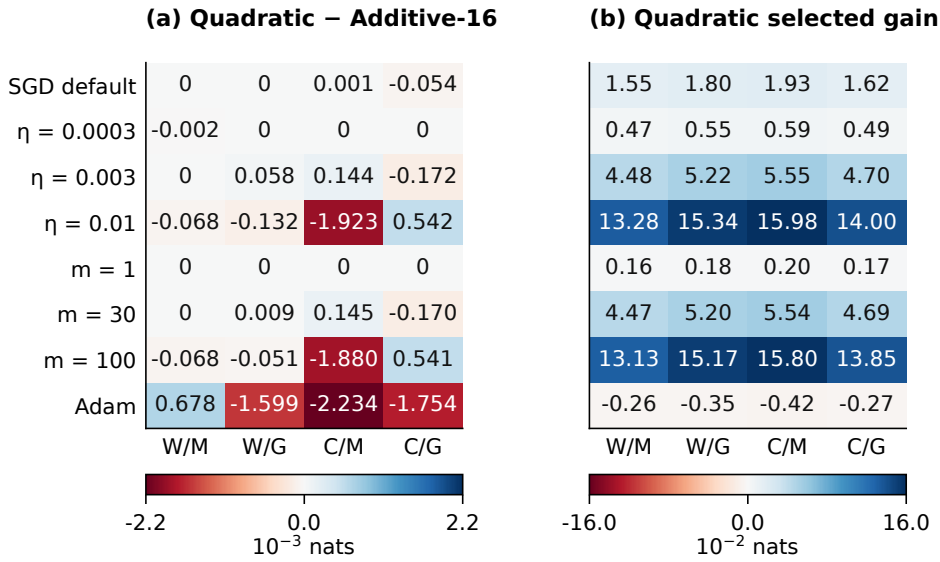

Figure 7: Update-operator sensitivity at the same 180 training states. Absolute selected gain and Quadratic-minus-additive gain answer different questions. Color scales and numeric labels distinguish useful updates from a relative predictor improvement. Each configuration uses its own base-only endpoint with the default utility anchor fixed. Values are descriptive state averages, not additional control-return experiments. Adam uses fresh optimizer state. W/C denote Walker/Cheetah and M/G denote MLP/GRU.

C.5 Budget endpoints and effective selection frequency

Tables 20 and 21 separate fixed-24 operating points from endpoints reached under the paired Quadratic budget. Mean checkpoint cost need not exhaust the budget; unfinished blocks and selection work cannot be credited as completed training. The same distinction determines whether a refresh enters the frequency denominator.

For example, a refresh after block 40 can finish within a budget but fail to contribute if block 41 finishes after it. Counting that refresh as an endpoint decision would describe work done, not a selection used by the evaluated model. The complete scope counts in Table 22 give 304 used Additive-6 decisions out of 360 raw selector records. Fixed All-five prescriptions have no learned-selection denominator.

Table 20: Recorded online cost and return at 24 update blocks.

Setting	Method	Mean cost (s)	Median (s)	Blocks	Mean return	CVaR	Quadratic budget (s)
W/MLP	Quadratic	122.17	123.10	24.00	0.6880	0.3890	122.17
W/MLP	Additive-6	69.69	70.69	24.00	0.6880	0.3890	122.17
W/MLP	All-five (0.20)	64.97	64.98	24.00	0.6955	0.3985	122.17
W/GRU	Quadratic	149.97	150.23	24.00	0.6736	0.3944	149.97
W/GRU	Additive-6	85.86	86.39	24.00	0.6736	0.3944	149.97
W/GRU	All-five (0.20)	79.15	79.19	24.00	0.6832	0.4070	149.97
C/MLP	Quadratic	124.26	126.59	24.00	0.6170	0.3208	124.26
C/MLP	Additive-6	71.90	74.46	24.00	0.6168	0.3205	124.26
C/MLP	All-five (0.20)	64.96	64.95	24.00	0.6272	0.3324	124.26
C/GRU	Quadratic	150.00	149.96	24.00	0.6220	0.3386	150.00
C/GRU	Additive-6	85.36	84.53	24.00	0.6220	0.3385	150.00
C/GRU	All-five (0.20)	79.28	79.29	24.00	0.6320	0.3501	150.00

Each row averages 15 paired correction runs. Budget endpoints use the last completed model within the paired Quadratic recorded-cost budget.

Table 21: Recorded online cost and return at the paired recorded-cost budget.

Setting	Method	Mean cost (s)	Median (s)	Blocks	Mean return	CVaR	Quadratic budget (s)
W/MLP	Quadratic	122.17	123.10	24.00	0.6880	0.3890	122.17
W/MLP	Additive-6	122.17	123.10	40.13	0.7099	0.4173	122.17
W/MLP	All-five (0.20)	122.17	123.10	44.67	0.7204	0.4309	122.17
W/MLP	Base-only	122.17	123.10	162.47	0.6923	0.3927	122.17
W/GRU	Quadratic	149.97	150.23	24.00	0.6736	0.3944	149.97
W/GRU	Additive-6	149.97	150.23	40.20	0.7010	0.4308	149.97
W/GRU	All-five (0.20)	149.97	150.23	45.07	0.7150	0.4502	149.97
W/GRU	Base-only	149.97	150.23	163.53	0.6822	0.4048	149.97
C/MLP	Quadratic	124.26	126.59	24.00	0.6170	0.3208	124.26
C/MLP	Additive-6	124.26	126.59	40.07	0.6429	0.3510	124.26
C/MLP	All-five (0.20)	124.26	126.59	45.33	0.6554	0.3661	124.26
C/MLP	Base-only	124.26	126.59	165.27	0.6164	0.3210	124.26
C/GRU	Quadratic	150.00	149.96	24.00	0.6220	0.3386	150.00
C/GRU	Additive-6	150.00	149.96	40.00	0.6468	0.3664	150.00
C/GRU	All-five (0.20)	150.00	149.96	45.00	0.6607	0.3826	150.00
C/GRU	Base-only	150.00	149.96	163.27	0.6309	0.3491	150.00

Each row averages 15 paired correction runs. Budget endpoints use the last completed model within the paired Quadratic recorded-cost budget.

Table 22: Selection records contributing to recorded-budget endpoints.

Setting	Method	Raw	Selector	Used	Late	Unused	Fixed
W/MLP	Quadratic	45	45	45	0	0	0
W/MLP	Additive-6	90	90	77	9	4	0
W/MLP	All-five (0.20)	45	0	0	0	0	45
W/GRU	Quadratic	45	45	45	0	0	0
W/GRU	Additive-6	90	90	76	10	4	0
W/GRU	All-five (0.20)	45	0	0	0	0	45
C/MLP	Quadratic	45	45	45	0	0	0
C/MLP	Additive-6	90	90	76	10	4	0
C/MLP	All-five (0.20)	45	0	0	0	0	45
C/GRU	Quadratic	45	45	45	0	0	0
C/GRU	Additive-6	90	90	75	14	1	0
C/GRU	All-five (0.20)	45	0	0	0	0	45

Used counts a selector refresh once only if at least one complete update in the evaluated endpoint applies it. Late decisions and completed-but-unused decisions do not enter the denominator. Fixed prescriptions are not selector decisions; base-only has no selector.

D ADDITIONAL ANALYSES OF RECORDED CANDIDATE BRANCHES

These analyses reuse the 180 common states on Quadratic trajectories: five base checkpoints, three correction runs per checkpoint, and refreshes at 0, 8, and 16 completed blocks in each of four settings. They change the score fit, candidate family, or aggregation of saved utility values, not the recorded parameter updates. Their unit is a candidate block at a fixed state. Closed-loop returns come from the separately executed trained methods. The summaries in this section are descriptive and carry no new interval or significance claims. The independent audit and trained controls in Appendix C complement these fixed-state analyses.

D.1 SAME-LABEL REGULARIZATION SENSITIVITY

At each state, both score models are fitted to the same empty, singleton, and pair labels. We compare the average-squared-error ridge penalty 10^{-3} with zero penalty. For Quadratic at zero penalty, the singleton–pair finite-difference construction is the exact full-rank least-squares solution; the zero row for the empty set adds no parameter. Additive-16 at zero penalty is the five-feature least-squares fit to these same labels, not a singleton-only fit. Both fits use the same 26 candidates and tie-breaking order.

Table 23 and Figure 8(a) compare raw pair effects with fitted pair coefficients *on the same fitting split*. This removes the fit-to-audit change of cases from the coefficient comparison. Ridge affects the numerical coefficients, as expected from Appendix A.2, but removing it changes neither model's choice at any recorded state. The two models agree at 178/180 states under either penalty. The zero-penalty Quadratic fit slightly lowers triple prediction RMSE in all four settings (Table 24). Its exact fit on the measured singleton and pair sets is interpolation, not additional prediction evidence.

Table 23: Raw and fitted effects on the same fitting split. RMS pools five singleton values or ten pair values at each of 45 states. The last column is RMS of the difference between the fitted pair coefficient and raw pair effect. All entries use the displayed utility scale factors.

Setting	μ (10^{-3})	Raw d (10^{-5})	Ridge d (10^{-4})	Difference (10^{-4})
Walker/MLP	5.850	3.295	1.247	1.488
Walker/GRU	6.002	3.945	1.256	1.554
Cheetah/MLP	6.907	4.451	1.447	1.803
Cheetah/GRU	5.928	3.815	1.222	1.508

Table 24: Zero-penalty refitting on the same 16 labels. RMSE pools 450 triples evaluated on independent audit cases per setting; regret averages 45 common states. Choice changes compare each predictor with its own ridge version on the same recorded states. Errors and regret are descriptive utility quantities, not new return measurements.

Setting	Predictor	Triple RMSE	Audit regret	Changed
Walker/MLP	Quadratic, ridge	0.004956	0.000341	0/45
Walker/MLP	Quadratic, zero	0.004861	0.000341	0/45
Walker/MLP	Additive-16, ridge	0.004866	0.000341	0/45
Walker/MLP	Additive-16, zero	0.004880	0.000341	0/45
Walker/GRU	Quadratic, ridge	0.005465	0.000445	0/45
Walker/GRU	Quadratic, zero	0.005410	0.000445	0/45
Walker/GRU	Additive-16, ridge	0.005412	0.000445	0/45
Walker/GRU	Additive-16, zero	0.005420	0.000445	0/45
Cheetah/MLP	Quadratic, ridge	0.004570	0.000362	0/45
Cheetah/MLP	Quadratic, zero	0.004483	0.000362	0/45
Cheetah/MLP	Additive-16, ridge	0.004488	0.000364	0/45
Cheetah/MLP	Additive-16, zero	0.004500	0.000364	0/45
Cheetah/GRU	Quadratic, ridge	0.005100	0.000979	0/45
Cheetah/GRU	Quadratic, zero	0.005075	0.000979	0/45
Cheetah/GRU	Additive-16, ridge	0.005077	0.000925	0/45
Cheetah/GRU	Additive-16, zero	0.005081	0.000925	0/45

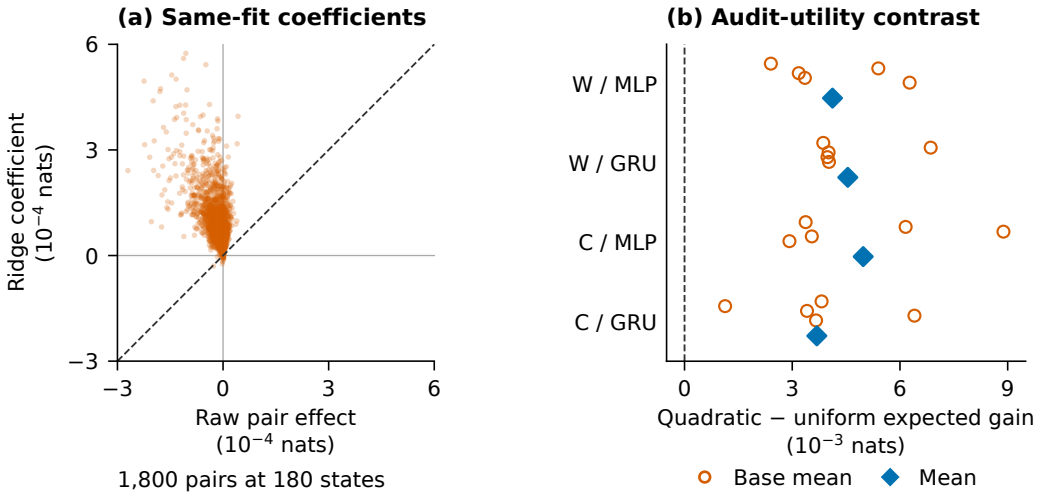

Figure 8: Coefficient sensitivity and local selection value. (a) All 1,800 raw pair effects and fitted ridge pair coefficients on the same fitting splits; the line is the identity. Coefficient differences need not change the selected set. (b) Quadratic minus the exact expected audit gain of a uniform random triple. Each open point averages three runs and three refresh states within one base checkpoint; diamonds average the five bases. The zero line indicates equal utility. These are descriptive fixed-state comparisons, with no new confidence intervals or random-selector training runs in this panel. Separately trained random and fixed controls appear in Figure 3.

D.2 Uniform and fixed-triple utility comparisons

Let $\mathcal{T} = \{S \subseteq \mathcal{D} : |S| = 3\}$ contain all ten triples. For a uniform random choice at a given recorded state,

$$\mathbb{E}_{S\sim\mathrm{Unif}(\mathcal{T})} G_{\mathrm{audit}}(S) = \frac{1}{10} \sum_{S\in\mathcal{T}} G_{\mathrm{audit}}(S).$$

All terms are measured on the same audit split, reference, menu, and tail. Thus this expected utility requires no sampled random choice. Quadratic's selection uses fitting labels only. We compare its audited gain with this expectation and with each of the ten fixed triples, reporting the complete family in Table 25.

The Quadratic-minus-uniform contrast is positive at 177/180 states and at all 20 base-level averages. The three negative states are all in Cheetah/GRU. The setting-level mean is positive against every fixed triple, with differences from 0.00270 to 0.00765 nats. These findings concern usefulness of the selected composition relative to unadapted compositions *at Quadratic's states*. They do not estimate the state distribution or return reached by training with a random or fixed triple, and are not evidence that pair features outperform additive selection.

Table 25: Quadratic minus each utility-only comparison, in 10^{-3} nats. Each cell averages 45 common Quadratic states. Uniform expectation averages all ten triple gains at each state; the remaining rows report every fixed triple. All entries are descriptive independent-audit utility differences, not endpoint return effects or simultaneous significance claims.

Comparison	Walker/MLP	Walker/GRU	Cheetah/MLP	Cheetah/GRU
Uniform expectation	4.121	4.547	4.976	3.685
$\{2,3,4\}$	4.510	5.818	7.650	3.996
$\{2,3,6\}$	4.612	5.327	5.838	3.180
$\{2,3,8\}$	5.038	5.813	5.795	4.143
$\{2,4,6\}$	3.701	4.263	5.157	3.900
$\{2,4,8\}$	4.142	4.750	5.095	4.862
$\{2,6,8\}$	4.224	4.262	3.297	4.027
$\{3,4,6\}$	3.599	3.953	5.353	2.704
$\{3,4,8\}$	4.036	4.444	5.295	3.653
$\{3,6,8\}$	4.123	3.959	3.479	2.834
$\{4,6,8\}$	3.222	2.884	2.802	3.550

D.3 Full-case utility and target sensitivity

For each fit or audit split, we reconstruct candidate and base losses both on the 13 fixed tail members and on all 128 saved cases. Full-case utility is $F_{128}(\vartheta) = 128^{-1}\sum_{i=1}^{128}\ell_i(\vartheta)$. Reference targets, menus, energy scales, and candidate parameters are unchanged. We refit Quadratic and Additive-16 with ridge 10^{-3} on the 16 full-case fit gains, then evaluate their triple predictions and 26-set choices against full-case audit gains. No audit label is used for fitting or selecting.

Table 26 gives the two aggregations side by side. The full-case raw pair-to-singleton RMS ratio is 0.76–0.94%, and Quadratic–Additive-16 agreement remains 178/180. All choices still have size three. This extends the observed approximate additivity to a non-tail summary of the same case panel, rather than establishing that tail selection has no effect. Indeed, changing the target alters 47 Quadratic choices and 47 additive choices. The changed states are not identical for the two predictors.

Table 27 evaluates both Quadratic choices under each common audit target. Fitting the full-case target does not guarantee higher full-case audit gain: its mean difference is slightly negative in three settings and positive in Cheetah/GRU. This is a transfer comparison between disjoint fit and audit cases. Neither the smaller absolute full-case RMSE nor the conditional tail gain is a calibration of repeated-audit uncertainty. The second independent audit and matched-norm comparisons are reported separately in Appendix C.3.

Table 26: Utility aggregation sensitivity at fixed recorded states. Each target uses its own fitting labels and the matching independent-audit aggregation. The ratio uses raw audit pair-effect RMS divided by singleton-gain RMS. Agreement compares Quadratic with Additive-16 under that target. Both use ridge 10^{-3}; absolute errors across the two different targets are not directly comparable.

Setting	Target	d/μ RMS	Quadratic RMSE	Add. RMSE	Agreement
Walker/MLP	Tail 13	0.574%	0.004956	0.004866	45/45
Walker/MLP	Full 128	0.785%	0.000772	0.000757	44/45
Walker/GRU	Tail 13	0.703%	0.005465	0.005412	45/45
Walker/GRU	Full 128	0.935%	0.000890	0.000874	44/45
Cheetah/MLP	Tail 13	0.722%	0.004570	0.004488	44/45
Cheetah/MLP	Full 128	0.914%	0.000785	0.000771	45/45
Cheetah/GRU	Tail 13	0.535%	0.005100	0.005077	44/45
Cheetah/GRU	Full 128	0.758%	0.000770	0.000773	45/45

Table 27: Effect of changing Quadratic's fitting target from tail to full-case mean. The last two columns evaluate the full-case-fitted choice minus the tail-fitted choice under a common audit target, averaged over 45 states, in 10^{-5} nats. A positive value favors the full-case-fitted choice. Changing the fitting target does not guarantee improvement on the independent audit target.

Setting	Choices changed	Tail audit difference	Full-case audit difference
Walker/MLP	14/45	+3.246	-1.698
Walker/GRU	9/45	-8.981	-2.077
Cheetah/MLP	10/45	-12.250	-1.637
Cheetah/GRU	14/45	+61.110	+8.793

D.4 Rescoring beyond the three-loss family

We evaluate the ridge-fitted score at all 32 subsets of the five depths, including the five four-loss sets and All-five. The coefficients still come from the same 16 singleton/pair fitting labels. No new gain labels or parameter branches are created. Table 28 reports all cardinality counts, both with and without ridge.

Quadratic with ridge selects All-five at 164/180 states, versus 149/180 without ridge. Its other ridge choices have size four; without ridge, 28 have size four and three have size three. This score-level sensitivity differs from the unchanged choices within $B = 3$. This subsection isolates the effect of rescoring an expanded family. Actual four- and five-depth gains, measured separately in Appendix C.3, now distinguish predicted preference from audited value. Neither fixed-state analysis measures the return of training with an unrestricted learned selector.

Table 28: Offline score maximizers over all 32 subsets. Each row counts 45 common states; no zero-, one-, or two-loss set is selected. The same 16 fitting labels are retained. This table reports fitted scores rather than measured gains; actual larger-set audits are reported in Appendix C.3. The table does not describe retrained policies.

Setting	Predictor	$\lvert S\rvert = 3$	$\lvert S\rvert = 4$	$\lvert S\rvert = 5$
Walker/MLP	Quadratic, ridge	0	5	40
Walker/MLP	Quadratic, zero	0	7	38
Walker/MLP	Additive-16, ridge	0	7	38
Walker/MLP	Additive-16, zero	0	7	38
Walker/GRU	Quadratic, ridge	0	5	40
Walker/GRU	Quadratic, zero	1	10	34
Walker/GRU	Additive-16, ridge	1	10	34
Walker/GRU	Additive-16, zero	1	10	34
Cheetah/MLP	Quadratic, ridge	0	2	43
Cheetah/MLP	Quadratic, zero	1	3	41
Cheetah/MLP	Additive-16, ridge	1	3	41
Cheetah/MLP	Additive-16, zero	1	3	41
Cheetah/GRU	Quadratic, ridge	0	4	41
Cheetah/GRU	Quadratic, zero	1	8	36
Cheetah/GRU	Additive-16, ridge	1	8	36
Cheetah/GRU	Additive-16, zero	1	8	36

E Regularization, oracle opportunities, and scaling

This section evaluates the same saved candidate values under alternative fits and summarizes the update-operator scan. The candidate parameter branches, fitting/audit split, and closed-loop episode outcomes are unchanged. Removing a score penalty or consulting an audit oracle creates no additional controller-return result.

E.1 Prediction with and without ridge

Tables 29–31 report all combinations of four settings, two independent audits, two utility targets, and three unseen cardinalities. Every fit uses the same 16 labels, with either five additive features or fifteen singleton–pair features. Tail and full-case targets are fitted separately. The zero-penalty quadratic fit interpolates the measured singleton and pair labels; its triple and larger-set values remain predictions. RMSE pools squared errors before taking the square root.

The four-/five-depth comparison favors Quadratic in 2/32 cells with ridge and 30/32 without it. The zero-penalty relative improvement over zero-penalty Additive-16 ranges from -0.3861% to 1.9114%. The two negative differences both occur for Cheetah/GRU in audit A under full-case utility. For triples, 15/16 comparisons favor zero-penalty Quadratic, with relative differences from -0.1078% to 0.6818%. Its one negative case is again Cheetah/GRU, audit A, full-case utility. These small descriptive differences are not a statistical equivalence or superiority test.

Neither predictor changes any of its 180 selections after removing ridge, under either tail or full-case fitting, when the feasible family is held at $\mathcal{S}_3$. This statement is specific to that family: rescoring all 32 sets can change selected cardinality, as shown in Appendix D.4. The 32 larger-set cells and 16 triple cells are views of the same nested state panel, not independent replications.

Table 29: Pooled prediction RMSE for measured triples, in 10^{-3} nats. Each model fits the same 16 empty/singleton/pair labels; Tail and Full are fitted separately. Each cell pools squared error over 45 states and 450 candidate values before taking the square root. Neither audit batch is used for fitting. Ridge is the mean-loss penalty 10^{-3}; Zero removes the penalty. Quadratic and additive selections remain restricted to subsets of size at most three. W/C denote Walker/Cheetah.

			Ridge		Zero	
Setting	Audit	Target	Quadratic	Additive	Quadratic	Additive
W/MLP	A	Tail	4.95601	4.86625	4.86144	4.87952
W/MLP	A	Full	0.77196	0.75686	0.75442	0.75921
W/MLP	B	Tail	4.27497	4.18469	4.17890	4.19788
W/MLP	B	Full	0.64282	0.63236	0.63086	0.63390
W/GRU	A	Tail	5.46543	5.41207	5.41035	5.41968
W/GRU	A	Full	0.88957	0.87421	0.87370	0.87631
W/GRU	B	Tail	5.02347	4.98358	4.98053	4.98952
W/GRU	B	Full	0.79596	0.78182	0.78177	0.78370
C/MLP	A	Tail	4.56976	4.48829	4.48340	4.49992
C/MLP	A	Full	0.78468	0.77086	0.76960	0.77289
C/MLP	B	Tail	4.47892	4.37777	4.36796	4.39262
C/MLP	B	Full	0.89095	0.87490	0.87330	0.87723
C/GRU	A	Tail	5.09992	5.07694	5.07506	5.08103
C/GRU	A	Full	0.77032	0.77337	0.77377	0.77293
C/GRU	B	Tail	5.66454	5.55839	5.54554	5.57546
C/GRU	B	Full	0.73842	0.72522	0.72245	0.72741

Table 30: Pooled prediction RMSE for measured quadruples, in 10^{-3} nats. Each model fits the same 16 empty/singleton/pair labels; Tail and Full are fitted separately. Each cell pools squared error over 45 states and 225 candidate values before taking the square root. Neither audit batch is used for fitting. Ridge is the mean-loss penalty 10^{-3}; Zero removes the penalty. Quadratic and additive selections remain restricted to subsets of size at most three. W/C denote Walker/Cheetah.

			Ridge		Zero	
Setting	Audit	Target	Quadratic	Additive	Quadratic	Additive
W/MLP	A	Tail	6.32331	6.09151	6.06303	6.10796
W/MLP	A	Full	1.00089	0.96044	0.95139	0.96345
W/MLP	B	Tail	5.45950	5.22081	5.19038	5.23759
W/MLP	B	Full	0.82945	0.79962	0.79382	0.80165
W/GRU	A	Tail	7.07793	6.93700	6.92482	6.94630
W/GRU	A	Full	1.15485	1.11280	1.10888	1.11560
W/GRU	B	Tail	6.37273	6.26303	6.24917	6.27027
W/GRU	B	Full	1.02517	0.98594	0.98314	0.98847
C/MLP	A	Tail	5.84094	5.61568	5.59102	5.63073
C/MLP	A	Full	1.00516	0.96538	0.96018	0.96804
C/MLP	B	Tail	5.74165	5.46471	5.42327	5.48416
C/MLP	B	Full	1.15573	1.11143	1.10432	1.11450
C/GRU	A	Tail	6.44261	6.38514	6.37812	6.38892
C/GRU	A	Full	0.96988	0.97589	0.97757	0.97521
C/GRU	B	Tail	7.24789	6.97811	6.93017	6.99867
C/GRU	B	Full	0.94299	0.90664	0.89772	0.90939

Table 31: Pooled prediction RMSE for measured quintuples, in 10^{-3} nats. Each model fits the same 16 empty/singleton/pair labels; Tail and Full are fitted separately. Each cell pools squared error over 45 states and 45 candidate values before taking the square root. Neither audit batch is used for fitting. Ridge is the mean-loss penalty 10^{-3}; Zero removes the penalty. Quadratic and additive selections remain restricted to subsets of size at most three. W/C denote Walker/Cheetah.

			Ridge		Zero	
Setting	Audit	Target	Quadratic	Additive	Quadratic	Additive
W/MLP	A	Tail	7.74798	7.30718	7.24440	7.32682
W/MLP	A	Full	1.24352	1.16450	1.14586	1.16819
W/MLP	B	Tail	6.70751	6.24701	6.18068	6.26741
W/MLP	B	Full	1.02783	0.96657	0.95435	0.96912
W/GRU	A	Tail	8.72668	8.45181	8.42514	8.46285
W/GRU	A	Full	1.43393	1.35068	1.34147	1.35422
W/GRU	B	Tail	7.75020	7.52787	7.49894	7.53647
W/GRU	B	Full	1.26782	1.18897	1.18163	1.19219
C/MLP	A	Tail	7.18427	6.73339	6.68039	6.75199
C/MLP	A	Full	1.24069	1.15832	1.14742	1.16165
C/MLP	B	Tail	7.09326	6.54613	6.45889	6.57033
C/MLP	B	Full	1.43651	1.34782	1.33243	1.35169
C/GRU	A	Tail	7.79076	7.67681	7.66488	7.68020
C/GRU	A	Full	1.16883	1.17551	1.17914	1.17461
C/GRU	B	Tail	8.89079	8.38531	8.29053	8.40928
C/GRU	B	Full	1.15962	1.08697	1.06960	1.09030

E.2 Oracle opportunities under cold-start Adam

Table 32 compares both predicted choices with the best measured audit-A gain among 26 feasible sets. The empty set is included with zero gain, so oracle gain is nonnegative at every state. Positive mean oracle gain does not mean that a strictly beneficial set exists at every state.

For tail utility, oracle gain is positive at 143/180 states, Quadratic gain is negative at 120, and both conditions hold at 83. For full-case utility the corresponding counts are 138, 122, and 81. All eight setting–target means are positive for the oracle and negative for both fitted selectors. The mean oracle-minus-Quadratic gap is therefore a missed fixed-state opportunity; it is not the return difference between an oracle-trained policy and Quadratic. The oracle uses audit information unavailable to the selectors.

Table 32: Cold-start Adam audit at 45 states per setting. Gains and regrets are in 10^{-3} nats. $G^\star$ maximizes audit-A gain over all 26 subsets of size at most three, including the empty set; G_Q, G_A evaluate the Quadratic and additive fit-selected sets. R_Q, R_A are the corresponding oracle regrets. Counts $n_+^\star$ and n_-^Q denote positive oracle gain and negative Quadratic gain; $n_{+,-}$ counts their intersection. All means and counts are descriptive. Tail and Full denote separately fitted tail13 and full128 objectives; W/C denote Walker/Cheetah.

Setting	Target	$G^\star$	G_Q	G_A	R_Q	R_A	$n_+^\star$	n_-^Q	$n_{+,-}$
W/MLP	Tail	2.15090	-2.56191	-3.24028	4.71281	5.39118	38	31	24
W/GRU	Tail	2.64322	-3.46905	-1.87032	6.11227	4.51354	34	28	17
C/MLP	Tail	2.62946	-4.24316	-2.00898	6.87262	4.63843	34	31	20
C/GRU	Tail	2.23166	-2.67424	-0.91983	4.90590	3.15149	37	30	22
W/MLP	Full	0.24522	-0.59608	-0.44910	0.84130	0.69432	34	29	18
W/GRU	Full	0.32851	-0.74520	-0.36547	1.07372	0.69398	33	34	23
C/MLP	Full	0.43281	-0.53320	-0.15504	0.96601	0.58785	33	29	17
C/GRU	Full	0.31037	-0.35737	-0.40103	0.66773	0.71140	38	30	23

E.3 Descriptive scales across SGD configurations

Table 33 fits an ordinary least-squares line to log RMS versus log learning rate at fixed $m = 10$, or versus log block length at fixed $\eta = 10^{-3}$. The learning-rate grid is $\{0.0003, 0.001, 0.003, 0.01\}$. The block-length slope uses $\{10, 30, 100\}$; the one-step measurements remain in the complete scan tables but do not enter this multi-step slope. Each RMS pools raw finite differences over the same 45 states per setting and uses audit A. The default utility anchor is fixed throughout the scan.

The learning-rate slopes are consistent with first-, second-, and third-order effects. These are finite-grid descriptive fits without confidence intervals. In particular, the block-length slopes are not tests of a theorem holding m fixed, and none identifies a causal Hessian contribution. Mean selected displacement norms under cold-start Adam are close in scale to the default SGD norms, but this does not match per-state directions or separate the optimizer's nonadditive displacement from the curvature of the evaluated utility.

Table 33: Descriptive log–log slopes for SGD. The learning rates are 0.0003, 0.001, 0.003, and 0.01 at ten steps; the block lengths are 10, 30, and 100 at $\eta = 0.001$. μ, d, κ denote pooled RMS of raw audit-A singleton, pair, and triple effects over 45 states per setting. $\|\Delta\|$ denotes mean selected displacement norm. Each entry is the slope of an ordinary least-squares fit in log coordinates, not a confidence bound. Tail/Full denote tail13/full128; W/C denote Walker/Cheetah.

		Learning rate η				Block length m			
Setting	Target	μ	d	κ	$\|\Delta\|$	μ	d	κ	$\|\Delta\|$
W/MLP	Tail	0.969	1.966	2.965	0.963	0.951	1.969	3.028	0.940
W/GRU	Tail	0.969	1.957	2.975	0.956	0.952	1.956	3.040	0.929
C/MLP	Tail	0.970	1.964	3.009	0.955	0.953	1.966	3.069	0.928
C/GRU	Tail	0.971	1.974	2.944	0.958	0.954	1.986	2.963	0.934
W/MLP	Full	0.958	1.943	2.937	0.960	0.935	1.929	2.959	0.936
W/GRU	Full	0.958	1.936	2.940	0.954	0.935	1.917	2.961	0.928
C/MLP	Full	0.959	1.944	2.955	0.957	0.937	1.930	2.984	0.931
C/GRU	Full	0.959	1.947	2.947	0.957	0.936	1.937	2.981	0.931