WHAT SHOULD CRITIC LEARN? RETHINKING CRITIC DESIGN FOR PREFERENCE-CONDITIONED MULTI-OBJECTIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective reinforcement learning (MORL) increasingly relies on preference-conditioned policies to adapt decisions across competing objectives, yet a basic architectural question remains unresolved: *how should critics represent and estimate objective-wise values?* Existing methods variously use scalarized critics, shared multi-head critics, or fully separated critics, often without a principled account of what design is appropriate. We study this question by separating two sources of critic error that are typically conflated: the loss of objective-specific information induced by scalarized supervision and the cross-objective interference induced by shared trainable representations. We first show that scalarized supervision can mask substantial objective-wise errors, even when the scalarized prediction is accurate. We then characterize the interference that remains in multi-head critics through their shared representation, deriving a local gradient-interaction result and a finite-trajectory decomposition of the prediction residual introduced by parameter sharing. Guided by our proposed theory, we first conduct a controlled toy experiment designed to directly validate the predicted behavior. We then empirically evaluate scalarized, shared multi-head, and fully split critic architectures across five representative MORL algorithms—Momba, CAPQL, GPI-LS, MORL/D, and MO-MPO—on eight continuous-control environments spanning two, three, and five objectives. Across the evaluated settings, our results show that critic architecture materially affects MORL performance, with fully split critics providing a robust default across diverse evaluated settings. More broadly, our findings establish critic architecture as a fundamental design choice in MORL.

1 INTRODUCTION

Multi-objective reinforcement learning (MORL) is becoming increasingly important as reinforcement learning moves from controlled benchmarks toward real-world decision-making, where agents are rarely expected to optimize a single objective (Hayes et al., 2022). Practical systems must often balance competing criteria such as performance, safety, cost, energy consumption, fairness, or user satisfaction, and the appropriate trade-off between these objectives can vary substantially across users, tasks, and deployment conditions (Teoh et al., 2025). This makes *preference* a central component of modern MORL: rather than committing to a single fixed scalar objective, an agent should ideally represent and respond to different preference vectors, allowing its behavior to adapt as the desired trade-off changes (Basaklar et al., 2023; Yang et al., 2025). Consequently, recent MORL methods increasingly aim to learn preference-conditioned policies that generalize across a continuum of objective trade-offs (Basaklar et al., 2023; Liu et al., 2025; Yang et al., 2025), making the quality of the underlying value representation—and in particular the critic architecture used to estimate multiple objectives—fundamental to reliable preference-aware decision making.

Despite rapid progress in MORL, much of the recent literature has focused on improving Pareto-front coverage, preference generalization, and policy performance through increasingly sophisticated preference conditioning, scalarization, and optimization strategies (Abels et al., 2019; Reymond et al., 2022; Lu et al., 2023). However, most existing MORL algorithms treat critic architecture as a default implementation choice, leaving a more fundamental question surprisingly

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.

underexplored: *what should the critic learn, and how should it represent the values of different objectives?* Recent work suggests that architectural choices themselves can substantially affect MORL performance (Terekhov & Gulcehre, 2024; Šťafa et al., 2026), but the representation of objective-wise value information remains comparatively underexplored. This matters because the critic is not merely a predictor: its value estimates and objective-wise advantages directly shape the learning signal used to update the policy (Ambadkar et al., 2026). How multiple objectives are represented in the critic therefore directly determines what information is available for learning preference-dependent decisions. As MORL increasingly aims to support diverse and changing preferences, the choice of critic representation becomes a fundamental part of the learning problem, rather than simply an implementation detail.

In this paper, we take a step back from designing yet another preference-conditioning or scalarization strategy and instead study the critic architecture itself. As shown in Figure 1, we formulate MORL critic design along two distinct axes—*what supervision signal each objective receives* and *how much of the critic representation is shared across objectives*—which yields three canonical designs: scalarized critics, objective-wise multi-head critics with a shared representation, and fully split objective-wise critics. This separation allows us to identify two fundamentally different sources of estimation difficulty. We show that scalarized supervision can hide substantial objective-wise error even when the scalarized value is correct, while separate output heads alone do not eliminate interference introduced through shared trainable representations. We then develop a finite-budget theory that characterizes how this cross-objective coupling propagates through training and establishes conditions under which parameter separation reduces prediction error. The resulting contribution is therefore not simply a new critic variant, but a principled account of *what MORL critic architectures should be used*. Guided by this analysis, we first validate the theoretical predictions in a controlled environment, and then evaluate the three critic architectures across five representative MORL algorithms—Momba (Šťafa et al., 2026), CAPQL (Lu et al., 2023), GPI-LS (Alegre et al., 2023), MORL/D (Felten et al., 2024), and MO-MPO (Abdolmaleki et al., 2020)—on eight continuous-control environments spanning two, three, and five objectives. Our results show that critic architecture materially affects MORL performance, with fully split critics providing a robust default when cross-objective interactions are conflicting or unknown.

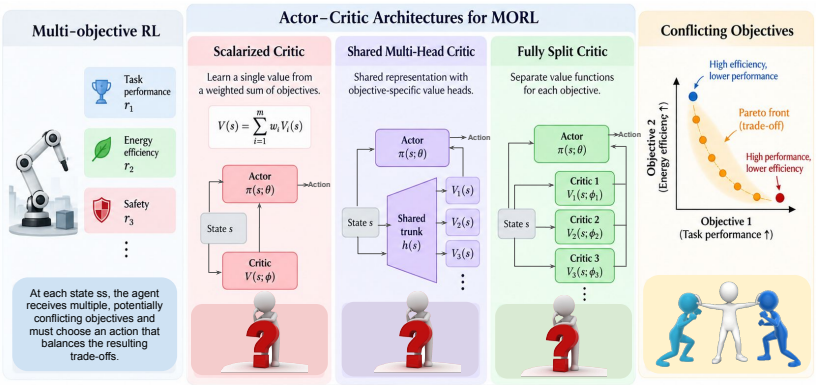


Figure 1: **Critic architecture choices in preference-conditioned multi-objective reinforcement learning.** Three common critic designs: (a) a scalarized critic that directly learns a preference-weighted action value, (b) a shared multi-head critic that predicts objective-wise values from a common representation, or (c) fully split critics with separate representations for each objective.

2 Related Work

MORL has progressively moved from maintaining separate solutions for different preferences toward shared, preference-conditioned architectures. DER (Abels et al., 2019) and Envelope Q-learning (Yang et al., 2019) learn preference-conditioned Q-functions, while PCN (Reymond et al., 2022) conditions policies on desired returns. PD-MORL (Basaklar et al., 2023) extends preference conditioning to continuous control, and PSL-MORL (Liu et al., 2025) generates preference-specific policy parameters through a hypernetwork. Terekhov & Gulcehre (2024) empirically examine multi-body, merged-input, hypernetwork, and shared versus separated actor–critic designs. These studies explore how to represent different preferences; we instead investigate how objective-wise value estimators should share supervision and parameters. CAPQL (Lu et al., 2023) pre-

serves vector-valued supervision within a shared critic, while D^3PO (Ambadkar et al., 2026) uses a shared-body multi-head critic and theoretically analyzes late preference integration in policy optimization. HRA (Van Seijen et al., 2017) learns separate value functions for reward components, and a humanoid loco-manipulation study (Yardımcı, 2026) empirically compares unified and dual critics. PCGrad (Yu et al., 2020) addresses gradient conflicts across shared losses, whereas CoMOGA (Kim et al., 2025) resolves objective-wise policy-gradient conflicts through constrained aggregation. C51 (Bellemare et al., 2017) introduces categorical distributional value learning, which Momba (Šťafa et al., 2026) applies to preference-conditioned MORL. Multivariate distributional RL (Wiltzer et al., 2024) instead models vector-valued return distributions. While recent MORL research has focused heavily on improving empirical performance, rigorous theoretical understanding of the underlying critic architecture and representation remains limited. To our knowledge, no prior work provides a unified theoretical account of *when MORL should use scalarized, shared multi-head, or fully split critics*. We address this architectural question directly by disentangling scalarization-induced error masking from interference caused by parameter sharing, and by establishing conditions under which separating critic representations provably reduces estimation error.

3 Preliminaries and Study Setup

Preference-conditioned multi-objective control: We consider a discounted multi-objective Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \mathbf{r}, \gamma, \rho_0)$, with $\mathbf{r} \in \mathbb{R}^m$ and $\gamma \in [0, 1)$. All reward components are maximized. A preference vector belongs to $\Delta^{m-1} = \{w \in \mathbb{R}^m_+ : \sum_i w_i = 1\}$ and remains fixed within an episode. We learn one preference-conditioned actor $\pi_w(a \mid s) = \pi_\phi(a \mid s, w)$, rather than a separate policy for each objective. Training uses positive reward scales c_i and normalized rewards $\widetilde{r}_i = c_i r_i$. Thus w specifies preferences in normalized reward coordinates. The training objective is

$$J^\alpha(\phi; w) = \mathbb{E}_{\rho_0, P, \pi_w}\left[\sum_{t \geq 0} \gamma^t \left(w^\top \widetilde{\mathbf{r}}_t - \alpha \log \pi_w(a_t \mid s_t)\right)\right], \tag{1}$$

where $\alpha > 0$ is the entropy coefficient. This discounted, normalized objective is distinct from the raw episodic returns used for benchmark evaluation.

Scalarized and objective-wise policy evaluation: For a fixed policy family, reward scales, and entropy coefficient, define

$$Q_i^{\pi,\alpha}(s, a, w) = \mathbb{E}_{\pi_w}\left[\sum_{t \geq 0} \gamma^t \widetilde{r}_{i,t} - \alpha \sum_{t \geq 1} \gamma^t \log \pi_w(a_t \mid s_t) \,\middle|\, s_0 = s, a_0 = a\right]. \tag{2}$$

We assume these discounted returns are absolutely integrable. Every component evaluates the same policy π_w and includes the same future entropy term; they are not independently optimized values Q_i^*. The entropy term starts at $t = 1$ because the initial action is conditioned on. In general, each $Q_i^{\pi,\alpha}$ depends on w through future actions. Write $Q_w^{\pi,\alpha}(s, a)$ for the corresponding soft value with scalar reward $w^\top \widetilde{\mathbf{r}}$. Their exact linear relationship is established in the theoretical analysis.

Critic supervision and parameter sharing: Let $x = (s, a, w)$. We distinguish two design choices: the supervised target and the extent of trainable parameter sharing. A *scalarized critic* (SC) directly predicts $\widehat{Q}_w(x)$ and receives scalarized-return supervision. An *objective-wise multi-head critic* (MH) predicts

$$\widehat{Q}_i^{\mathrm{MH}}(x) = h_{\nu_i}(f_\psi(x)), \qquad \widehat{Q}_i^{\mathrm{SP}}(x) = h_{\nu_i}(f_{\psi_i}(x)), \tag{3}$$

where MH shares the trainable trunk ψ but has objective-specific heads ν_i, whereas *split critics* (SP) have disjoint trainable parameter blocks across objectives. Both MH and SP receive objective-wise losses without multiplication by the current preference weights, and use $\widehat{Q}_w = \sum_i w_i \widehat{Q}_i$ for the common actor. All three designs remain preference-conditioned. These definitions describe supervision and parameterization, not interchangeable names for implementation variants. Scalar or categorical output heads are specified separately for each backbone.

Finite-budget evaluation protocol: Our controlled comparisons fix the evaluated policy and target construction. For an update budget T and a fixed evaluation distribution $\mathcal{D}_{\mathrm{eval}}$ over (s, a, w), the

common value-error metric is $\mathcal{E}_T = \left(\mathbb{E}_{(s,a,w)\sim\mathcal{D}_{\text{eval}}}\left[\left(\widehat{Q}_{w,T}(s,a,w) - Q_w^{\pi,\alpha}(s,a)\right)^2\right]\right)^{1/2}$. For MH and SP the prediction is obtained by late scalarization; for SC it is the directly predicted scalar. One update advances all relevant critic parameter blocks once.

4 Shared or Split Critics? A Theoretical Analysis

In this section, we compare scalarized critics (SC), objective-wise multi-head critics (MH), and fully split critics (SP), separating supervision design from parameter sharing. We analyze how scalarization can mask component errors and how shared trunks introduce cross-objective gradient interference. For smooth nonlinear critics under matched gradient updates, we then derive an exact decomposition of the sharing-induced prediction residual between MH and SP over a finite training trajectory. The analysis characterizes when this residual increases prediction error, rather than asserting a universal advantage of splitting.

4.1 Evaluation Is Separable Even When Policy Improvement Is Not

Fix the preference-conditioned policy family $\pi_w = \pi_\phi(\cdot \mid \cdot, w)$, reward scales $c_i > 0$, and entropy coefficient $\alpha \geq 0$. Let $x = (s, a, w)$ and $\widetilde{r}_i = c_i r_i$. Every objective value $Q_i^{\pi,\alpha}(x)$ evaluates the *same* policy π_w and includes the same future entropy contribution, starting after the conditioned initial action. Preferences remain fixed along a trajectory. Absolute integrability of the discounted returns is assumed.

Lemma 1. *For every $w \in \Delta^{m-1}$, $Q_w^{\pi,\alpha}(s,a) = \sum_{i=1}^m w_i Q_i^{\pi,\alpha}(s,a,w)$.*

Proof sketch. Apply linearity of expectation to the discounted rewards. The repeated future entropy term receives total weight $\sum_i w_i = 1$, so it is counted exactly once. □

Thus scalarized and objective-wise evaluation have compatible true values. This identity neither favors a parameterization nor makes independently fitted approximations equally accurate. In general, objective critics must still depend on w because future actions do. We compare SC, which directly predicts Q_w; MH, with $\widehat{Q}_i(x) = h_{\nu_i}(f_\psi(x))$ and a shared trainable trunk ψ; and SP, which replaces ψ by objective-specific copies ψ_i. Both MH and SP receive unweighted objective-wise losses. They still evaluate one common actor.

4.2 Scalarized Supervision Masks Objective Errors

We first distinguish scalarized supervision, used by SC, from objective-wise supervision, used by both MH and SP. This comparison concerns the information retained in the training signal, independently of whether critic parameters are shared. Although both approaches can recover the same scalarized value, they need not provide the same objective-specific feedback. Fix a preference $w \in \Delta^{m-1}$ and a frozen objective target $y \in \mathbb{R}^m$. Consider a loss that depends on the prediction and target only through their scalar projections: $\mathcal{L}_{\text{sc}}(\widehat{y}; y, w) = \ell\big(w^\top \widehat{y}, w^\top y\big)$. Here $\widehat{y}$ is an output-space analytical device: a direct scalar critic need not internally produce an objective-value vector. By contrast, objective-wise supervision applies losses to the individual components before aggregating them, as in $\sum_i \ell_i(\widehat{y}_i, y_i)$. MH and SP differ in their parameterization, not in this supervision distinction.

Proposition 2. *For every $v \in \text{Null}(w^\top)$, $\mathcal{L}_{\text{sc}}(\widehat{y} + v; y, w) = \mathcal{L}_{\text{sc}}(\widehat{y}; y, w)$. This invariance holds for any scalar loss ℓ defined on the projected arguments; neither a squared loss nor differentiability is required.*

If ℓ is twice differentiable in its first argument, then

$$\nabla^2_{\widehat{y}} \mathcal{L}_{\text{sc}} = \partial_{11}\ell\big(w^\top \widehat{y}, w^\top y\big)\, w w^\top, \tag{4}$$

whose rank is at most one.

In particular, for the squared-loss surrogates

$$\ell_{\text{sc}} = \frac{1}{2}\left[w^\top(\widehat{y} - y)\right]^2, \qquad \ell_{\text{obj}} = \frac{1}{2}\|\widehat{y} - y\|_2^2, \tag{5}$$

the output-space Hessians are $\nabla^2_{\widehat{y}}\ell_{\mathrm{sc}} = ww^\top, \qquad \nabla^2_{\widehat{y}}\ell_{\mathrm{obj}} = I_m$. *If* w *is uniform on the simplex, its second-moment matrix satisfies*

$$M := \mathbb{E}_w[ww^\top] = \frac{I_m + \mathbf{1}\mathbf{1}^\top}{m(m+1)}, \qquad mM = P_+ + \frac{1}{m+1}P_-, \tag{6}$$

where $P_+ = \mathbf{1}\mathbf{1}^\top/m$ *and* $P_- = I_m - P_+$ *project onto the common-objective and objective-contrast subspaces.*

Proof sketch. If $w^\top v = 0$, then $w^\top(\widehat{y} + v) = w^\top\widehat{y}$, which proves the loss invariance. The chain rule gives Eq. 4. For the squared scalar loss, $\partial_{11}\ell = 1$; differentiating the objective-wise squared loss gives I_m. Finally, uniform-simplex preferences satisfy $\mathbb{E}[w_i^2] = 2/[m(m+1)]$ and $\mathbb{E}[w_i w_j] = 1/[m(m+1)]$ for $i \neq j$. These moments give Eq. 6. □

For example, let $y = (3, 1)$, $\widehat{y} = (2, 2)$, and $w = (1/2, 1/2)$. The component errors are $(-1, 1)$, but their scalar projection is zero: $w^\top(\widehat{y} - y) = 0$. Consequently, scalarized supervision cannot distinguish $\widehat{y}$ from the correct vector at this preference. Under Eq. 5, $\ell_{\mathrm{sc}} = 0$ whereas $\ell_{\mathrm{obj}} = 1$. The scalarized value is nevertheless correct at this w: the example demonstrates hidden component errors, not a biased scalar prediction at that preference.

Projection masking holds for any loss based only on the scalar projection; the output-space curvature comparison specifically uses squared losses. Ambiguity at one preference does not imply that SC cannot learn across preferences. Both MH and SP retain objective-wise feedback; we next examine whether sharing their parameters introduces additional interference.

4.3 Separate Heads Do Not Remove Shared-Trunk Interference

Now keep objective-wise supervision fixed. Freeze the policy, batch, sampled next actions, normalization, and stopped targets, including the target-selection decision. Let $L_i(\psi, \nu_i)$ be objective i's loss, and define $g_i = \nabla_\psi L_i, h_i = \nabla_{\nu_i} L_i, G = \sum_i g_i$. MH applies $\psi^+ = \psi - \eta G$ and $\nu_i^+ = \nu_i - \eta h_i$. For a matched one-step intervention, SP copies the current trunk, initializes $\psi_i = \psi$, and applies $\psi_i^+ = \psi - \eta g_i$ with the same head update. Let D_{MH} and D_{SP} denote the decreases in $\sum_i L_i$.

Theorem 3 (Interference remaining after head separation)**.** *If each* L_i *is twice continuously differentiable near the compared updates, then*

$$D_{\mathrm{MH}} = \eta\Big(\|G\|^2 + \sum_i \|h_i\|^2\Big) + O(\eta^2), \tag{7}$$

$$D_{\mathrm{SP}} = \eta \sum_i \Big(\|g_i\|^2 + \|h_i\|^2\Big) + O(\eta^2), \tag{8}$$

$$D_{\mathrm{SP}} - D_{\mathrm{MH}} = -2\eta \sum_{i<j} \langle g_i, g_j \rangle + O(\eta^2). \tag{9}$$

If the summed cross-objective inner product is negative, SP gives strictly greater one-step progress for sufficiently small positive η.

Proof sketch. Expand every loss along its shared and split update. The first-order head contributions are identical and cancel in the difference. Expand $\|\sum_i g_i\|^2$ to expose the cross-objective terms. Appendix B.3 bounds the remainder and gives an explicit admissible step size. □

4.4 Sharing-Induced Prediction Error in Nonlinear Critics

The preceding local result identifies cross-objective gradient interference at a matched parameter point. We now trace the effect of parameter sharing along actual nonlinear training trajectories. This comparison keeps objective-wise supervision fixed and therefore concerns MH versus SP; scalarized supervision is a separate design choice. Rather than asserting that sharing always increases error, we identify its additional prediction residual and characterize when that residual increases a common evaluation error. Let objective i have critic $q_i(x; \psi, \nu_i)$ and per-step loss $L_{i,t}(\psi, \nu_i)$. MH and SP use the same sequence of loss functions, including the same frozen data and target realizations

at each step, and ordinary gradient descent with the same positive step sizes η_t. Initialize SP by copying the MH trunk and the corresponding private heads. The losses may vary across steps and need not be squared losses. Represent both runs in the common space $\Theta = \mathrm{col}_{i=1}^{m}(\psi_i, \nu_i)$, $F_t(\Theta) = \sum_{i=1}^{m} L_{i,t}(\psi_i, \nu_i)$. The repeated trunk blocks of Θ_t^{MH} all equal the actual shared trunk ψ_t^{MH}; this representation does not add independent parameters to MH. The SP trunk blocks evolve independently. Along the MH trajectory, define $g_{i,t} = \nabla_\psi L_{i,t}(\psi_t^{\mathrm{MH}}, \nu_{i,t}^{\mathrm{MH}})$, $c_t = \mathrm{col}_{i=1}^{m}\left(\sum_{j\neq i} g_{j,t},\ 0_{\nu_i}\right)$. Thus c_t contains the updates from other objectives received by each shared trunk block, satisfy:

$$\Theta_{t+1}^{\mathrm{MH}} = \Theta_t^{\mathrm{MH}} - \eta_t\{\nabla F_t(\Theta_t^{\mathrm{MH}}) + c_t\}, \tag{10}$$

$$\Theta_{t+1}^{\mathrm{SP}} = \Theta_t^{\mathrm{SP}} - \eta_t \nabla F_t(\Theta_t^{\mathrm{SP}}). \tag{11}$$

Let $\mathcal{Q}(\Theta) \in \mathbb{R}^n$ collect predictions on one fixed, weighted evaluation set, and let $y \in \mathbb{R}^n$ contain the corresponding common reference values with the same weights. For scalarized value prediction, its ℓth entry is $\sqrt{\omega_\ell}\sum_i w_{\ell i} q_i(x_\ell; \psi_i, \nu_i)$, where $\omega_\ell \geq 0$ and $\sum_\ell \omega_\ell = 1$. Define $e_T^a = \mathcal{Q}(\Theta_T^a) - y$ and $\mathcal{M}_T^a = \|e_T^a\|_2^2$ for $a \in \{\mathrm{MH}, \mathrm{SP}\}$. The reference may be the true values of a common evaluated policy; it need not be realizable by the critic.

Theorem 4 (Sharing-induced prediction error over a finite trajectory)**.** *Assume the paired updates above start from $\Theta_0^{\mathrm{MH}} = \Theta_0^{\mathrm{SP}}$. For each $t < T$, suppose F_t is twice continuously differentiable on a neighborhood of the segment joining the two current parameter vectors. Suppose $\mathcal{Q}$ is continuously differentiable near the final segment. Write $d_t = \Theta_t^{\mathrm{MH}} - \Theta_t^{\mathrm{SP}}$ and define $\overline{\mathcal{H}}_t = \int_0^1 \nabla^2 F_t(\Theta_t^{\mathrm{SP}} + sd_t)\, ds$, $A_t = I - \eta_t \overline{\mathcal{H}}_t$, $\overline{J}_T = \int_0^1 D\mathcal{Q}(\Theta_T^{\mathrm{SP}} + sd_T)\, ds$, $\Phi_{T,t+1} = A_{T-1}\cdots A_{t+1}$ with an empty product equal to I. Then*

$$e_T^{\mathrm{MH}} = e_T^{\mathrm{SP}} + b_T^{\mathrm{share}}, \tag{12}$$

$$b_T^{\mathrm{share}} = -\sum_{t=0}^{T-1} \eta_t \overline{J}_T \Phi_{T,t+1} c_t. \tag{13}$$

Consequently, the excess mean-square prediction error is exactly $\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}} = 2\langle e_T^{\mathrm{SP}}, b_T^{\mathrm{share}}\rangle + \|b_T^{\mathrm{share}}\|_2^2$. In particular, if $b_T^{\mathrm{share}} \neq 0$ and $\langle e_T^{\mathrm{SP}}, b_T^{\mathrm{share}}\rangle \geq 0$, then

$$\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}} \geq \|b_T^{\mathrm{share}}\|_2^2 > 0. \tag{14}$$

Proof sketch. Subtract the paired updates and use the integral mean-value identity to obtain $d_{t+1} = A_t d_t - \eta_t c_t$. Unrolling from $d_0 = 0$ gives $d_T = -\sum_{t<T} \eta_t \Phi_{T,t+1} c_t$. The integral Jacobian identity yields $\mathcal{Q}(\Theta_T^{\mathrm{MH}}) - \mathcal{Q}(\Theta_T^{\mathrm{SP}}) = \overline{J}_T d_T$. Subtract the common reference and expand the squared norm. Appendix C provides the complete proof. □

5 Experiments

5.1 A controlled example of finite-budget critic learning

We first isolate critic optimization from exploration, bootstrapping, and actor learning in a realizable two-objective terminal decision problem. A single state admits three actions with normalized reward vectors $\widetilde{\mathbf{r}}(a_+) = (3, 1), \widetilde{\mathbf{r}}(a_-) = (1, 3), \widetilde{\mathbf{r}}(a_0) = (2.3, 2.3)$. Every action terminates immediately, so the true action values are exactly the displayed rewards and there is no future entropy contribution. Preferences are $w = (t, 1-t), \quad t \sim \mathrm{Unif}[0, 1]$. The corresponding scalarized values are

$$Q_w(a_+) = 1 + 2t, \qquad Q_w(a_-) = 3 - 2t, \qquad Q_w(a_0) = 2.3.$$

Hence a_- is optimal for $t < 0.35$, a_0 for $0.35 \leq t \leq 0.65$, and a_+ for $t > 0.65$. The balanced action therefore makes an underestimated objective contrast decision-relevant: a critic that learns the average value but not the difference between the two objectives can select a_0 over too large a range of preferences. We use the shared-plus-private affine family analyzed in Section 4. Let $\varphi(a) = \sqrt{3}\, e_a$, where e_a is the one-hot encoding of action a. Under uniform evaluation over the three actions, $\Sigma = \mathbb{E}[\varphi(a)\varphi(a)^\top] = I_3$. We compare the scalarized critic (SC), objective-wise multi-head critic (MH), and fully split critics (SP) using the same shared-channel strength $\chi = 0.75$.

Table 1: Executed deterministic affine-critic experiment. $T_{1\%}$ is the first update with relative RMSE at most 0.01. SC/MH/SP all use 12 trainable scalars.

Design	$\mathcal{E}_{20}/\mathcal{E}_0$	$\mathcal{E}_{100}/\mathcal{E}_0$	$T_{1\%}$	$\mathcal{R}_{20}$
SC	0.714521	0.186241	275	0.245000
MH	0.358486	0.005921	90	0.014052
SP	0.011529	2.04×10^{-10}	21	6.12×10^{-6}

All variants begin from the same scalarized value function. For MH and SP, the initial objective predictions are

$$\widehat{\mathbf{Q}}_0(a_+) = \widehat{\mathbf{Q}}_0(a_-) = (2, 2), \qquad \widehat{\mathbf{Q}}_0(a_0) = (2.3, 2.3).$$

SC is initialized to the corresponding scalarized predictions, $\widehat{Q}_{w,0}(a_+) = \widehat{Q}_{w,0}(a_-) = 2$ and $\widehat{Q}_{w,0}(a_0) = 2.3$ for every w. Thus the common-value component is already correct and the remaining error lies entirely in the objective-contrast direction. This initialization is deliberately chosen to isolate contrast learning; it is not intended to model random neural-network initialization. Importantly, all three critic classes can represent the exact target function, so the comparison does not impose an asymptotic approximation error on SC or MH.

Experiment Results. We perform full-population gradient descent with learning rate $\eta = 0.2$ using the loss normalization defined in Section 4. The underlying model parameters are updated directly; the analytical recurrences derived in the theory are not used as the optimizer. Preference expectations can be evaluated exactly in this affine problem.

For all three designs, we report the same scalarized-value error

$$\mathcal{E}_T = \left(\mathbb{E}_{a,w} \left[\left(\widehat{Q}_{w,T}(a) - Q_w(a) \right)^2 \right] \right)^{1/2}, \tag{15}$$

normalized by its initial value. For MH and SP, $\widehat{Q}_{w,T} = \sum_i w_i \widehat{Q}_{i,T}$. We additionally measure whether the critic error changes the induced decision through the expected greedy regret

$$\mathcal{R}_T = \mathbb{E}_w \left[\max_a w^\top \widetilde{\mathbf{r}}(a) - w^\top \widetilde{\mathbf{r}} \left(\arg\max_a \widehat{Q}_{w,T}(a) \right) \right]. \tag{16}$$

This quantity is only a direct readout of critic-induced action ranking; it is not the return of a separately trained actor.

Table 1 shows that the controlled example exposes the two effects separated by our theory. Moving from SC to MH retains objective-wise supervision and sharply reduces the finite-budget error, while moving from MH to SP additionally removes cross-objective parameter coupling. After 20 updates, the relative scalarized-value RMSE decreases from 0.715 for SC to 0.358 for MH and 0.0115 for SP. The same estimation gap is reflected in greedy action ranking: $\mathcal{R}_{20}$ decreases from 0.245 to 0.0141 and then to 6.1×10^{-6}.

5.2 Evaluating Three Critic Architectures Across Five MORL Algorithms

We investigate how sharing critic parameters across objectives affects continuous-action multi-objective reinforcement learning. We implement alternative critic architectures within CAPQL, Momba, MORL/D with MOSAC, GPI-LS, and MO-MPO–TD(0) (Alegre et al., 2023; Felten et al., 2024; Haarnoja et al., 2018a;b; Šťafa et al., 2026). Within each matched comparison, we retain the actor architecture and the algorithm's policy-improvement procedure, allowing us to examine how critic design affects learning across different algorithmic frameworks.

Evaluation Matrix. We evaluate the learned policy set using the Expected Utility Metric (EUM). Let Π denote the set of returned policies and $\hat{\mathbf{J}}(\pi) \in \mathbb{R}^m$ the estimated return vector of policy π, averaged over evaluation episodes. We scale each objective using a fixed positive constant s_i: $\tilde{J}_i(\pi) = \frac{\hat{J}_i(\pi)}{s_i}$. Given a common set of preference vectors $\mathcal{W} = \{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(K)}\}$, where $w_i^{(k)} \geq 0$ and $\sum_{i=1}^m w_i^{(k)} = 1$, the reported EUM is $\mathrm{EUM}(\Pi; \mathcal{W}) = \frac{1}{K} \sum_{k=1}^K \max_{\pi \in \Pi} \left[\sum_{i=1}^m w_i^{(k)} \tilde{J}_i(\pi) \right]$. For each preference vector, the inner maximization selects the policy with the highest weighted

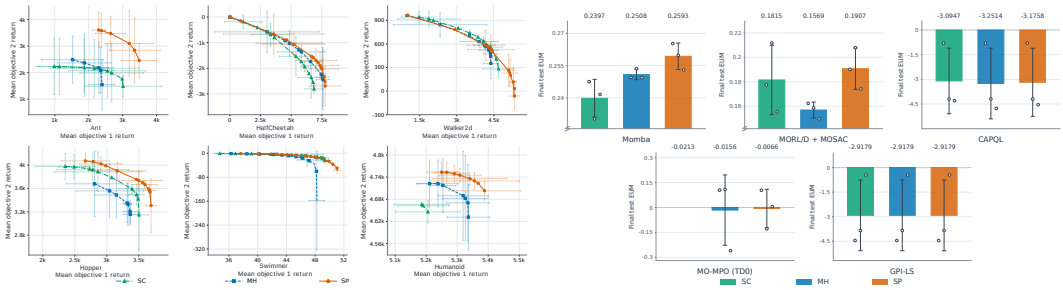
Table 2: MuJoCo configurations: final EUM (↑) at 200k training steps. Each cell shows the mean and sample standard deviation. Bold denotes the highest mean within each method and environment.

Method	Critic	Ant	HalfCheetah	Walker2d	Hopper	Swimmer	Humanoid
Momba	Scalarized (SC)	2450.49 ± 942.30	2540.40 ± 242.42	2588.71 ± 272.62	3575.59 ± 19.46	21.338 ± 0.898	4930.90 ± 45.81
	Multi-head (MH)	2295.21 ± 614.07	2938.16 ± 60.41	2470.98 ± 179.03	3407.29 ± 223.22	20.309 ± 0.931	5012.81 ± 20.88
	Split (SP)	**3291.30** ± 178.54	**3037.73** ± 602.18	**2806.92** ± 147.89	**3715.73** ± 81.15	**21.340** ± 0.307	**5054.65** ± 59.44
CAPQL	Scalarized (SC)	927.74 ± 69.09	1272.18 ± 1021.54	388.66 ± 116.87	2141.89 ± 718.61	14.314 ± 4.459	993.70 ± 289.56
	Multi-head (MH)	876.91 ± 63.60	1855.11 ± 304.39	268.96 ± 37.93	1449.89 ± 141.36	16.470 ± 2.554	983.86 ± 271.44
	Split (SP)	**1165.83** ± 492.02	**2409.61** ± 467.83	**958.36** ± 594.69	**2648.91** ± 435.17	**18.591** ± 1.466	**1408.86** ± 1198.72
MORL/D	Scalarized (SC)	1133.18 ± 248.23	1944.13 ± 136.26	847.12 ± 335.37	2144.18 ± 1019.25	18.448 ± 0.792	459.01 ± 33.81
	Multi-head (MH)	1102.21 ± 144.79	1994.00 ± 55.87	1099.69 ± 167.09	1527.08 ± 466.85	19.167 ± 0.762	614.97 ± 50.11
	Split (SP)	**1246.30** ± 232.19	**2023.73** ± 135.92	**1138.49** ± 263.80	**2296.27** ± 231.76	**19.786** ± 0.228	**691.20** ± 95.84
MO-MPO	Multi-head (MH)	179.62 ± 156.69	-19.18 ± 205.60	395.09 ± 96.91	501.61 ± 164.96	9.747 ± 7.136	474.50 ± 48.06
	Split (SP)	**247.96** ± 118.86	**524.58** ± 63.42	**525.37** ± 227.83	**1181.59** ± 576.56	**13.457** ± 3.781	**484.92** ± 42.59
GPI-LS	Scalarized (SC)	799.52 ± 91.79	2823.92 ± 1616.16	1557.41 ± 501.47	2707.19 ± 321.69	16.458 ± 2.309	2187.79 ± 1886.32
	Multi-head (MH)	1263.46 ± 347.95	2765.89 ± 1748.62	1607.48 ± 552.27	2907.27 ± 177.47	19.652 ± 1.463	**3550.73** ± 2.04
	Split (SP)	**1900.55** ± 1065.25	**3067.43** ± 915.17	**1976.47** ± 862.99	**3066.21** ± 105.40	**20.460** ± 0.244	2673.15 ± 463.83

utility; the outer average measures performance across the considered preferences. Higher EUM therefore indicates greater average utility under the specified preference set and objective scales. Since EUM summarizes utility across a specified set of preferences, it may conceal differences in individual trade-offs. We therefore complement EUM with empirical-front visualizations (see Figure 2a presenting one representative result for Momba, while the remaining results are provided in Appendix D), which reveal the distribution and extent of the observed objective trade-offs. Together, these evaluations characterize average utility and empirical solution-set geometry, without establishing recovery of the complete Pareto front.

Experimental protocol. The MuJoCo study covers Ant, HalfCheetah, Walker2d, Hopper, Swimmer, and Humanoid (Felten et al., 2023). New comparisons use three training seeds and a budget of 200,000 transitions, with evaluations every 10,000 transitions. Population-method budgets are totals across learners. We evaluate returned policy sets using expected utility, and aggregate scores using the mean and sample standard deviation. We also evaluate critic architectures on a more complex three-objective adaptation of GreenLight-Gym (van Laatum et al., 2024), a greenhouse simulator that couples indoor climate and tomato growth. We train each configuration for 500,000 transitions using three random seeds and evaluate its final policy set using the expected utility metric: we use $K = 231$ common preference vectors and three objectives: fruit fresh-mass gain, negative energy consumption, and negative CO_2 input, with fixed scales $\mathbf{s} = (3, 30, 0.5)$. We additionally evaluate the critic designs on a five-objective energy-scheduling task based on the IEEE 33-bus distribution feeder (Thurner et al., 2018; Baran & Wu, 1989), jointly optimizing operating cost, EV service, voltage quality, battery degradation, and renewable-energy utilization . See Appendix F for implementation details.

Architectural variants. Following the theoretical definitions in Section 4, we instantiate three critic designs: a scalarized critic (SC), which learns the preference-weighted return directly; an objective-wise multi-head critic (MH), which predicts individual objective values using a shared representation and objective-specific heads; and fully split critics (SP), which retain the same objective-wise supervision as MH but remove all trainable parameter sharing across objectives. Thus, SC→MH changes the supervision from scalarized to objective-wise, whereas MH→SP isolates the effect of



(a) Pareto fronts of three Momba variants.

(b) Performance on the three-objective GreenLight.

Figure 2: Pareto fronts of Momba variants and performance of all algorithms on the GreenLight

cross-objective parameter sharing. We instantiate these designs within Momba, CAPQL, GPI-LS, MORL/D, and MO-MPO while otherwise preserving each algorithm's actor, policy-improvement rule, replay mechanism, and algorithm-specific components. Where a backbone's native critic already matches one of these designs, we use it directly; otherwise we modify only the critic and its corresponding supervision to realize the theoretical design. Details are provided in Appendix D.

Results. As shown in Table 2, SP achieves the highest mean EUM in **29 of 30** method–environment comparisons, with GPI-LS on Humanoid the only exception. Gains can be substantial: on Walker2d, CAPQL achieves $2.47\times$ the EUM of its strongest shared variant. By contrast, MH does not consistently outperform SC, suggesting that objective-specific heads alone do not reliably recover the benefits of full separation. Improvements also extend to MORL/D+MOSAC, where SC uses scalarized supervision and MH/SP use objective-wise TD targets. Figure 2b shows that Momba with SP achieves the highest observed mean EUM, 0.2593, compared with 0.2397 for its native scalar critic and 0.2508 for MH, corresponding to improvements of 8.2% and 3.4%. SP also improves MORL/D, reaching 0.1907, versus 0.1815 for SC and 0.1569 for MH. For MO-MPO, SP produces a modest increase in mean EUM to -0.007, although variation across seeds is large relative to this gain. For CAPQL and GPI-LS, all critic variants perform poorly under the tested configurations. In this regime, limited learning may obscure the contribution of critic architecture, making the results inconclusive regarding the relative merits of the critic designs. Architectural differences are more apparent in settings where the underlying learners achieve stronger performance. On the five-objective energy-scheduling task, SP achieves the highest mean EUM for Momba, MORL/D+MOSAC, and CAPQL, while MH and SP perform nearly identically for MO-MPO, extending the observed benefit of separation beyond the lower-dimensional control benchmarks (see Appendix F). Together, these results support critic separation as a broadly effective architectural choice across the evaluated methods, rather than an algorithm-specific modification. These comparisons concern mean performance; substantial seed variance in some settings warrants caution about individual gains.

6 Conclusion

This work revisits a basic but largely unresolved question in MORL: how should multiple objective values be represented and learned by the critic? Our results show that this question is not primarily about representational capacity. Even when scalarized, multi-head, and split critics can represent the same value functions, they can induce substantially different finite-budget learning dynamics. Scalarized supervision can suppress objective-specific correction signals, while shared representations can couple otherwise distinct objective losses through their gradients. We characterize these effects separately, derive conditions under which such coupling slows estimation, and also identify regimes in which sharing provides positive transfer. The experiments show that these distinctions matter in practice. Across five MORL algorithms and environments spanning two, three, and five objectives, fully separated critics are consistently competitive and often substantially stronger than shared alternatives, while multi-head separation alone does not reliably recover the same gains. Together, these results shift the critic-design question from *whether one network can represent all objectives* to *what learning dynamics parameter sharing imposes on those objectives*. Critic architecture should therefore be treated as a first-class optimization choice in MORL, rather than as an implementation detail.

AI USE STATEMENT

Generative AI tools were used to edit and polish the manuscript for clarity and readability and helping create scientific figures. Generative AI was not used to develop the fundamental research ideas, including the distinction between critic architectures, the theory, or the interpretation of experimental results. The remaining required disclosure categories, such as dataset reformatting and translation, are not applicable to this work. We additionally used generative AI tools to improve the readability and presentation of the manuscript. All AI-assisted outputs were reviewed by the authors. We independently checked the resulting experimental outputs and validated their consistency with our experimental design. For the theoretical component, the main structure, arguments, and proof strategy were developed by the authors; generative AI was used only to identify potential gaps. The authors take full responsibility for the final content of this work, including all text, claims, code, mathematical arguments, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation details, experimental configurations, and computational resources used in this work in Appendix D. The code and the appendix documents the training and evaluation setup, relevant hyperparameters, and architecture-specific implementation choices required to reproduce the reported experiments. We also provide a link to the accompanying codebase, including the implementations used for the evaluated critic architectures and experimental pipelines. Together, these materials are intended to support independent reproduction of our main empirical results.

REFERENCES

Abbas Abdolmaleki, Sandy Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo Martins, Nicolas Heess, Raia Hadsell, and Martin Riedmiller. A distributional view on multi-objective policy optimization. In *International conference on machine learning*, pp. 11–22. PMLR, 2020.

Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *International conference on machine learning*, pp. 11–20. PMLR, 2019.

Lucas N. Alegre, Ana L. C. Bazzan, Diederik M. Roijers, Ann Nowé, and Bruno C. da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, pp. 2003–2012, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9432-1.

Tanmay Ambadkar, Sourav Panda, Shreyash Kale, Jonathan Dodge, and Abhinav Verma. Preference conditioned multi-objective reinforcement learning: Decomposed, diversity-driven policy optimization. *arXiv preprint arXiv:2602.07764*, 2026.

Mesut E. Baran and Felix F. Wu. Network reconfiguration in distribution systems for loss reduction and load balancing. *IEEE Transactions on Power Delivery*, 4(2):1401–1407, 1989. doi: 10.1109/61.25627.

Toygun Basaklar, Suat Gumussoy, and Umit Ogras. PD-MORL: Preference-driven multi-objective reinforcement learning algorithm. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zS9sRyaPFlJ.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. Pmlr, 2017.

Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El-Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4aa8891583f07ae200ba07843954caeb-Abstract.html.

Florian Felten, El-Ghazali Talbi, and Grégoire Danoy. Multi-objective reinforcement learning based on decomposition: A taxonomy and framework. *Journal of Artificial Intelligence Research*, 79: 679–723, 2024. doi: 10.1613/jair.1.15702.

Christopher 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, 2021.

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

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning: Cf hayes et al. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 36498–36527, 2025.

Hojoon Lee, Dongyoon Hwang, Donghu Kim, Hyunseung Kim, Jun Jet Tai, Kaushik Subramanian, Peter R. Wurman, Jaegul Choo, Peter Stone, and Takuma Seno. SimBa: Simplicity bias for scaling up parameters in deep reinforcement learning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a.

Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. Hyperspherical normalization for scalable deep reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33352–33403. PMLR, 2025b.

Erlong Liu, Yu-Chang Wu, Xiaobin Huang, Chengrui Gao, Ren-Jian Wang, Ke Xue, and Chao Qian. Pareto set learning for multi-objective reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18789–18797, 2025.

Haoye Lu, Daniel Herman, and Yaoliang Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations*, 2023.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 1110–1118, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Adam Šťafa, Santeri Heiskanen, Petr Novotný, and Joni Pajarinen. Momba: Network modernization improves multi-objective reinforcement learning. *Reinforcement Learning Journal*, 7, 2026. Presented at the Reinforcement Learning Conference (RLC) 2026.

Jayden Teoh, Pradeep Varakantham, and Peter Vamplew. On generalization across environments in multi-objective reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 80888–80915, 2025.

Mikhail Terekhov and Caglar Gulcehre. In search for architectures and loss functions in multi-objective reinforcement learning. *arXiv preprint arXiv:2407.16807*, 2024.

Leon Thurner, Alexander Scheidler, Florian Schäfer, Jan-Hendrik Menke, Julian Dollichon, Friederike Meier, Steffen Meinecke, and Martin Braun. pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521, 2018. doi: 10.1109/TPWRS.2018.2829021.

Bart van Laatum, Eldert J. van Henten, and Sjoerd Boersma. GreenLight-Gym: Reinforcement learning benchmark environment for control of greenhouse production systems. *arXiv preprint arXiv:2410.05336*, 2024. Accepted at the 8th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture (AgriControl 2025).

Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. Hybrid reward architecture for reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 101297–101336, 2024.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32, 2019.

Yucheng Yang, Tianyi Zhou, Mykola Pechenizkiy, and Meng Fang. Preference controllable reinforcement learning with advanced multi-objective optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Mehmet Turan Yardımcı. Critic architecture matters: Dual vs. unified critics for humanoid loco-manipulation, 2026. URL `https://arxiv.org/abs/2606.11891`.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

A Controlled Experiments on Critic Supervision and Parameter Sharing

This appendix provides controlled numerical experiments that isolate the two mechanisms studied in the main text: (i) loss of objective-specific correction under early scalarization, and (ii) cross-objective interference caused by shared trainable parameters. The experiments use realizable affine critics with exact population gradients, so that value-estimation error can be measured against known ground truth without exploration, replay sampling, target-network lag, or actor optimization. They are designed as numerical checks of the theoretical mechanism, not as substitutes for the CAPQL and Momba experiments with nonlinear critics.

A.1 Three-action terminal problem

We first consider a two-objective terminal decision problem with a single nonterminal state and three actions. Their normalized reward vectors are

$$\widetilde{\mathbf{r}}(a_+) = (3, 1), \qquad \widetilde{\mathbf{r}}(a_-) = (1, 3), \qquad \widetilde{\mathbf{r}}(a_0) = (2.3, 2.3). \tag{17}$$

Every action terminates immediately, so $Q_i(a) = \widetilde{r}_i(a)$ exactly and there is no future entropy contribution. Preferences are

$$w = (t, 1-t), \qquad t \sim \mathcal{U}[0, 1]. \tag{18}$$

The true scalarized action values are

$$Q_w(a_+) = 1 + 2t, \qquad Q_w(a_-) = 3 - 2t, \qquad Q_w(a_0) = 2.3. \tag{19}$$

Hence a_0 is optimal for $0.35 \le t \le 0.65$, while a_- and a_+ are optimal on the two outer preference regions. The balanced action is included so that underestimating the objective contrast can change greedy action selection; with only the two symmetric specialist actions, shrinking the contrast by a positive factor would preserve their pairwise ordering.

Feature representation and initialization. We use

$$\varphi(a) = \sqrt{3}\, e_a, \tag{20}$$

where e_a is the one-hot vector associated with action a. Under uniform evaluation of the three actions,

$$\Sigma = \mathbb{E}_a[\varphi(a)\varphi(a)^\top] = I_3. \tag{21}$$

All critic variants are initialized to the same effective predictions,

$$\widehat{\mathbf{Q}}_0(a_+) = \widehat{\mathbf{Q}}_0(a_-) = (2, 2), \qquad \widehat{\mathbf{Q}}_0(a_0) = (2.3, 2.3). \tag{22}$$

Thus the common component is initially correct and only the objective contrast remains to be learned. This is an intentional warm start used to isolate contrast learning rather than a claim about generic random initialization. All three parameterizations below can represent the exact target functions.

Critic parameterizations. For a shared-channel strength $\chi \in (0, 1)$, we instantiate the three analytical designs from the main text:

$$\widehat{Q}_w^{\mathrm{SC}}(a) = \Big(\sqrt{\frac{\chi}{2}}\, u + \sqrt{\frac{\chi}{2}}\, u + \sqrt{1-\chi} \sum_i w_i v_i\Big)^\top \varphi(a), \tag{23}$$

$$\widehat{Q}_i^{\mathrm{MH}}(a) = \Big(\sqrt{\frac{\chi}{2}}\, u + \sqrt{\frac{\chi}{2}}\, u + \sqrt{1-\chi}\, v_i\Big)^\top \varphi(a), \tag{24}$$

$$\widehat{Q}_i^{\mathrm{SP}}(a) = \Big(\sqrt{\chi}\, u_i + \sqrt{1-\chi}\, v_i\Big)^\top \varphi(a). \tag{25}$$

SC and MH each contain twelve trainable scalars in this three-action construction, while SP contains the same amount.

Losses and optimization. We set $m = 2$, $\chi = 0.75$, and use full-population gradient descent with learning rate $\eta = 0.2$. SC minimizes

$$\mathcal{L}_{\mathrm{SC}} = \frac{m}{2}\,\mathbb{E}_{a,w}\left[\left(\widehat{Q}_w^{\mathrm{SC}}(a) - w^\top \widetilde{\mathbf{r}}(a)\right)^2\right], \tag{26}$$

while MH and SP minimize

$$\mathcal{L}_{\mathrm{obj}} = \frac{1}{2}\,\mathbb{E}_{a}\left[\sum_{i=1}^{m}\left(\widehat{Q}_i(a) - \widetilde{r}_i(a)\right)^2\right]. \tag{27}$$

The factor m in the scalarized loss is the normalization used in the theoretical analysis; it matches the common-mode curvature and prevents the comparison from being driven by a trivial global loss scale. Population gradients are evaluated exactly. For the uniform preference integral, two-point Gauss–Legendre quadrature is exact because the squared loss is quadratic in t; evaluation uses an independent higher-order quadrature rule.

Metrics. All designs are evaluated using the same scalarized-value RMSE,

$$\mathcal{E}_T = \left(\mathbb{E}_{a,w}\left[\left(\widehat{Q}_{w,T}(a) - Q_w(a)\right)^2\right]\right)^{1/2}, \tag{28}$$

where MH and SP are scalarized only at evaluation time. We additionally report the true utility lost by the action chosen greedily under the learned critic,

$$\mathcal{R}_T = \mathbb{E}_w\left[\max_a w^\top \widetilde{\mathbf{r}}(a) - w^\top \widetilde{\mathbf{r}}\left(\arg\max_a \widehat{Q}_{w,T}(a)\right)\right]. \tag{29}$$

This is a diagnostic readout of critic accuracy, not the return of a trained actor.

Figure 3 shows the full finite-budget fitting trajectories. In this pure contrast initialization, SP corrects the missing contrast fastest, MH is slower because the shared channel partially cancels the two objective updates, and SC is slower still because scalarized supervision attenuates the same contrast direction. The measured relative-error trajectories agree with the exact contractions derived in the main text,

$$\rho_{\mathrm{SC}} = 1 - \eta\frac{1-\chi}{3}, \qquad \rho_{\mathrm{MH}} = 1 - \eta(1-\chi), \qquad \rho_{\mathrm{SP}} = 1 - \eta. \tag{30}$$

For $\chi = 0.75$ and $\eta = 0.2$, these are $0.983\overline{3}$, 0.95, and 0.8, respectively. The maximum discrepancy between the directly updated parameters and the closed-form effective updates is on the order of 10^{-15} in double precision.

A.2 Long-budget fit: the gap is not an irreducible approximation bias

To distinguish finite-budget optimization from asymptotic representational failure, we continue the same contrast experiment to 2,000 updates. The resulting scalarized RMSE values are

$$\mathcal{E}_{2000}^{\mathrm{SC}} = 2.15 \times 10^{-15}, \qquad \mathcal{E}_{2000}^{\mathrm{MH}} = 6.18 \times 10^{-16}, \qquad \mathcal{E}_{2000}^{\mathrm{SP}} = 3.94 \times 10^{-16}. \tag{31}$$

All three therefore fit the realizable target to numerical precision. The controlled example supports a finite-update-budget claim, not an unavoidable asymptotic bias of every shared critic.

A.3 Boundary control: sharing can help aligned objectives

The main claim is conditional rather than universal. We therefore repeat the original two-action construction with an initialization whose error lies entirely in the common-objective direction. In this case, the two objectives request aligned changes to the shared representation. At $\chi = 0.75$ and $\eta = 0.2$, after 20 updates the relative RMSE is approximately

$$1.81 \times 10^{-4} \tag{32}$$

for SC/MH, compared with

$$1.15 \times 10^{-2} \tag{33}$$

for SP. Thus sharing can accelerate fitting when the objective gradients are positively aligned, consistent with the sign dependence in the interference analysis.

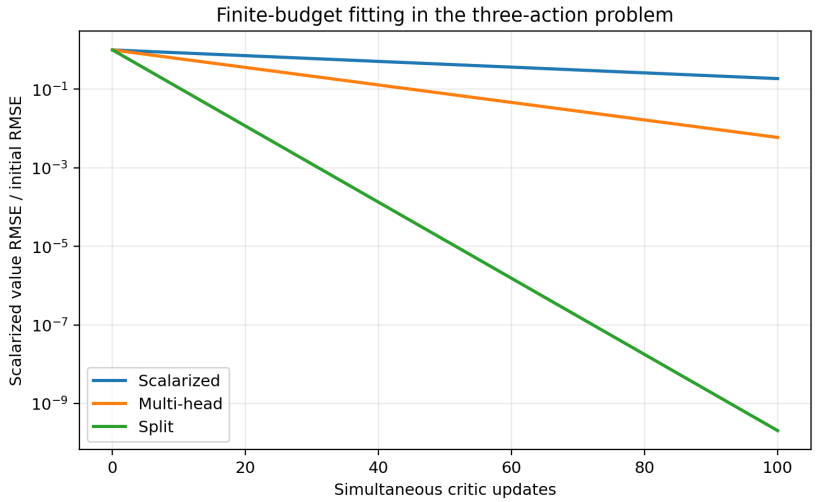


Figure 3: Finite-budget scalarized-value fitting in the three-action terminal problem. Curves are deterministic full-population gradient trajectories, so no confidence intervals are shown. All three methods use the same scalarized-value RMSE metric.



Figure 4: Absolute contrast tracking error in the fixed-reward two-step MDP with an externally changing evaluation policy. One critic update is allowed per policy change.

A.4 Sensitivity to shared-channel strength

We next vary the shared-channel strength

$$\chi \in \{0, 0.25, 0.5, 0.75, 0.9\}, \tag{34}$$

while keeping the contrast initialization fixed. Figure 5 reports the relative RMSE after 20 updates. The $\chi = 0$ condition is an explicit boundary control: the MH shared channel contributes no update, so MH and SP coincide in the contrast dynamics. As χ increases, contrast correction in MH becomes progressively slower, while SP remains unaffected by the cross-objective shared channel. SC additionally retains the scalarized-supervision attenuation.

A.5 Direct one-step test of the gradient-interference prediction

We also test the local identity used in the main theoretical analysis. At MH checkpoints, we freeze the current target and create a virtual split copy with identical current predictions by duplicating the

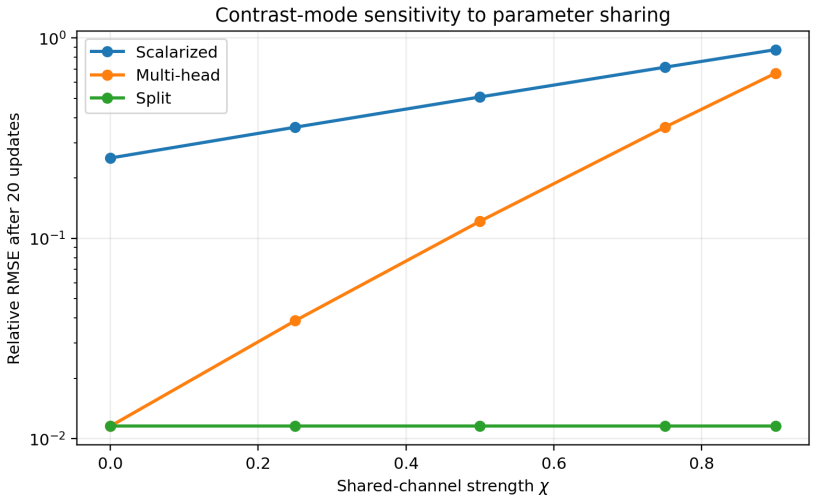


Figure 5: Contrast-mode fitting after 20 updates as a function of the shared-channel strength χ. The complete prespecified sweep is shown.

shared block. Let g_1 and g_2 denote the two objective gradients on the shared block, and let D_{MH} and D_{SP} be the decreases in the same summed objective loss after one shared or split probe update, respectively. The theory predicts

$$\lim_{\eta_{\mathrm{probe}} \to 0} \frac{D_{\mathrm{SP}} - D_{\mathrm{MH}}}{\eta_{\mathrm{probe}}} = -2 g_1^{\top} g_2. \tag{35}$$

For the initial contrast checkpoint, $g_1^{\top} g_2 = -0.75$, so the predicted slope is 1.5. The observed values of $(D_{\mathrm{SP}} - D_{\mathrm{MH}})/\eta_{\mathrm{probe}}$ are 1.40625, 1.490625, 1.4990625, 1.49990625, and 1.499990625 for probe step sizes $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, respectively. Thus the numerical intervention converges to the first-order prediction as the probe step decreases. In the common-mode control the shared gradients instead have positive inner product and the corresponding split-minus-shared gap is negative, matching the predicted positive-transfer case.

A.6 Fixed-reward two-step tracking experiment

Finally, we test whether slower fitting can translate into persistent lag when the value target moves. To give the target drift an RL interpretation without changing rewards, we use a fixed-reward two-step MDP. At s_0, actions a_+ and a_- deterministically transition to s_+ and s_- with zero immediate reward. At either second state, terminal actions b_1 and b_2 yield rewards $(3, 1)$ and $(1, 3)$, respectively. We set $\gamma = 0.9$ and $\alpha = 0$ in this controlled extension.

The evaluated policy is externally prescribed by

$$\pi_k(b_1 \mid s_+) = p_k, \qquad \pi_k(b_1 \mid s_-) = 1 - p_k, \tag{36}$$

with

$$p_k = 0.5 + 0.002k, \qquad k = 0, \ldots, 200. \tag{37}$$

The exact first-step values are therefore

$$Q^{\pi_k}(a_+) = \gamma\,(2 + (2p_k - 1),\ 2 - (2p_k - 1))\,, \tag{38}$$

$$Q^{\pi_k}(a_-) = \gamma\,(2 - (2p_k - 1),\ 2 + (2p_k - 1))\,. \tag{39}$$

The common component remains fixed, whereas the contrast coefficient increases by $d = 0.0036$ per policy change. All critics are initialized exactly at Q^{π_0}. At each step we perform one exact critic update toward the current target, then change the evaluation policy and measure error against the new exact target. This timing gives the scalar recurrence

$$b_{k+1} = \rho b_k - d, \tag{40}$$

for the contrast-mode tracking error.

At $k = 200$, the measured absolute contrast lags are

$$|b_{200}^{\mathrm{SC}}| = 0.208508, \qquad |b_{200}^{\mathrm{MH}}| = 0.071997, \qquad |b_{200}^{\mathrm{SP}}| = 0.018000. \tag{41}$$

These values coincide with the deterministic recurrence predicted by the tracking analysis. SP also has nonzero lag; the experiment shows a relative tracking advantage under the specified finite-update protocol rather than an unbiasedness guarantee.

A.7 Scope of the controlled experiments

Multi-objective reinforcement learning (MORL) is related to, but structurally distinct from, multi-task reinforcement learning (MTRL). MTRL typically learns across separate tasks with different goals, rewards, or dynamics, where parameter sharing is primarily used to promote transfer and generalization. MORL instead optimizes multiple simultaneous objectives within the same task: each transition carries a vector-valued reward, and a preference-conditioned policy trades off the corresponding returns according to a sample-dependent preference vector. This distinction changes the critic-design problem. Objective-wise critics in MORL estimate components of the return under the same policy, share bootstrapped future terms, and are ultimately coupled through preference-dependent scalarization. Consequently, existing analyses of shared representations and gradient interference in multi-task learning (Yu et al., 2020; Fifty et al., 2021) do not directly characterize MORL critics, particularly scalarized supervision (SC), which fits the preference-weighted return rather than a collection of task-specific losses. We therefore study MORL critic architecture through two separate design axes—supervision and parameter sharing—and systematically evaluate their interaction across multiple MORL algorithms.

B Detailed Proofs and Technical Conditions

B.1 Soft evaluation and proof of Lemma 1

Fix $w \in \Delta^{m-1}$, the policy π_w, positive reward scales c_i, and α. Write $\widetilde{r}_i = c_i r_i$ and condition throughout on $(s_0, a_0) = (s, a)$. With future contributions zero after true termination, define

$$Q_i^{\pi,\alpha}(s,a,w) = \mathbb{E}_{\pi_w}\left[\sum_{t=0}^{\infty}\gamma^t \widetilde{r}_{i,t} - \alpha\sum_{t=1}^{\infty}\gamma^t \log \pi_w(a_t \mid s_t)\right], \tag{42}$$

$$Q_w^{\pi,\alpha}(s,a) = \mathbb{E}_{\pi_w}\left[\sum_{t=0}^{\infty}\gamma^t \sum_i w_i\widetilde{r}_{i,t} - \alpha\sum_{t=1}^{\infty}\gamma^t \log \pi_w(a_t \mid s_t)\right]. \tag{43}$$

The initial action is conditioned on, explaining the entropy sum starting at $t = 1$.

Complete proof of Lemma 1. Absolute integrability permits the required exchanges of expectation and sums. Therefore

$$\sum_i w_i Q_i^{\pi,\alpha} = \mathbb{E}_{\pi_w}\left[\sum_{t\geq 0}\gamma^t\sum_i w_i\widetilde{r}_{i,t} - \alpha\sum_{t\geq 1}\gamma^t\Big(\sum_i w_i\Big)\log\pi_w(a_t \mid s_t)\right] \tag{44}$$

$$= \mathbb{E}_{\pi_w}\left[\sum_{t\geq 0}\gamma^t\sum_i w_i\widetilde{r}_{i,t} - \alpha\sum_{t\geq 1}\gamma^t\log\pi_w(a_t \mid s_t)\right] = Q_w^{\pi,\alpha}. \tag{45}$$

All expectations use the same policy; replacing its component values by separately optimized $Q_i^\star$ would not justify this calculation. □

Bellman-level identity. For an objective function Q_i, define

$$(\mathcal{T}_i^{\pi,\alpha}Q_i)(s,a,w) = \mathbb{E}\left[\widetilde{r}_i + \gamma(1-d)\{Q_i(s',a',w) - \alpha\log\pi_w(a' \mid s')\} \mid s,a\right], \tag{46}$$

where $a' \sim \pi_w(\cdot \mid s')$. If $Q_w = \sum_i w_i Q_i$, then

$$\sum_i w_i \mathcal{T}_i^{\pi,\alpha} Q_i = \mathcal{T}_w^{\pi,\alpha} Q_w. \tag{47}$$

Indeed, both the immediate reward and the next-value terms are linear, while the common entropy term again receives total weight one. No independence between objective returns is needed. Their marginal means determine expected linear utility, but their marginal laws need not determine the law of the weighted sum.

Why preference conditioning remains. Although an immediate reward may not depend on w, the distribution of a' in Eq. 46 generally does. Thus $Q_i^{\pi,\alpha}(s,a,w)$ need not equal $Q_i^{\pi,\alpha}(s,a,w')$. The finite-budget model below explicitly imposes preference-independent targets. Terminal policy evaluation is one possible instance, not a general property of MORL.

B.2 Proof of Proposition 2

Set $e = \widehat{y} - y$. The scalarized loss expands as

$$\ell_{\mathrm{sc}} = \frac{1}{2} \sum_i w_i^2 e_i^2 + \sum_{i<j} w_i w_j e_i e_j. \tag{48}$$

The cross terms may be negative, and complete cancellation occurs when $w^\top e = 0$. This is cancellation in a *supervision residual*, not yet a statement about parameter-gradient conflict.

Complete proof of Proposition 2. Direct differentiation gives

$$\nabla_{\widehat{y}} \ell_{\mathrm{sc}} = w(w^\top e), \qquad \nabla^2_{\widehat{y}} \ell_{\mathrm{sc}} = w w^\top, \tag{49}$$

whereas $\nabla_{\widehat{y}} \ell_{\mathrm{obj}} = e$ and $\nabla^2_{\widehat{y}} \ell_{\mathrm{obj}} = I_m$. Since $w \neq 0$, the rank of $ww^\top$ is one and its null space has dimension $m-1$.

For completeness, the uniform-simplex density in the first $m-1$ coordinates is $(m-1)!$. Repeated integration over that simplex gives, for nonnegative integer exponents k_i,

$$\mathbb{E}_w \prod_i w_i^{k_i} = \frac{(m-1)! \prod_i k_i!}{(m-1+\sum_i k_i)!}. \tag{50}$$

Setting one exponent to two, or two distinct exponents to one, yields

$$\mathbb{E}[w_i^2] = \frac{2}{m(m+1)}, \qquad \mathbb{E}[w_i w_j] = \frac{1}{m(m+1)} \quad (i \neq j). \tag{51}$$

Hence $M = (I_m + \mathbf{1}\mathbf{1}^\top)/[m(m+1)]$. The orthogonal projectors $P_+ = \mathbf{1}\mathbf{1}^\top/m$ and $P_- = I_m - P_+$ satisfy $P_+ P_- = 0$ and $P_+ + P_- = I_m$. Therefore

$$M = \frac{1}{m} P_+ + \frac{1}{m(m+1)} P_-, \qquad mM = P_+ + \frac{1}{m+1} P_-. \tag{52}$$

Both eigenvalues are positive, and their ratio is $m+1$. □

Interpretive limits. The example $e = (a, -a)$ at $w = (1/2, 1/2)$ has zero scalar-value error at that w, not a biased scalar prediction. Also, $\frac{1}{2}(\sum_i w_i e_i)^2$ is generally not $\sum_i w_i e_i^2/2$. For an arbitrary scalar neural prediction $f_\theta(x)$ with residual $\delta_\theta(x)$, its squared-loss parameter Hessian instead contains

$$\mathbb{E}\left[\nabla_\theta f_\theta \nabla_\theta f_\theta^\top + \delta_\theta \nabla^2_\theta f_\theta\right]. \tag{53}$$

It cannot be replaced by M without specifying the parameterization. The generic linear construction below supplies those additional conditions. Nor does a projection-loss calculation identify a scalar return-distribution cross-entropy with the sum of marginal cross-entropies.

B.3 Proof of Theorem 3 and a finite-step condition

Each L_i depends on the common block ψ and its own disjoint head ν_i. The losses and all stopped target computations are fixed during the comparison. Define

$$D_{\mathrm{MH}} = \sum_i \{L_i(\psi, \nu_i) - L_i(\psi - \eta G, \nu_i - \eta h_i)\}, \tag{54}$$

$$D_{\mathrm{SP}} = \sum_i \{L_i(\psi, \nu_i) - L_i(\psi - \eta g_i, \nu_i - \eta h_i)\}. \tag{55}$$

The second expression evaluates each loss on its own trunk copy.

Complete proof of Theorem 3. Let $z_i = (\psi, \nu_i)$. For a displacement direction v,

$$L_i(z_i - \eta v) = L_i(z_i) - \eta \nabla L_i(z_i)^\top v + R_i(v), \tag{56}$$

$$R_i(v) = \eta^2 \int_0^1 (1-t)\, v^\top \nabla^2 L_i(z_i - t\eta v) v \, dt. \tag{57}$$

For MH, take $v = (G, h_i)$. Its first-order contribution to the decrease of objective i is $\eta(g_i^\top G + \|h_i\|^2)$. Summing over i and using $\sum_i g_i = G$ gives

$$D_{\mathrm{MH}} = \eta\Big(\|G\|^2 + \sum_i \|h_i\|^2\Big) + O(\eta^2). \tag{58}$$

For SP, take $v = (g_i, h_i)$. Summing the corresponding expansions gives

$$D_{\mathrm{SP}} = \eta \sum_i (\|g_i\|^2 + \|h_i\|^2) + O(\eta^2). \tag{59}$$

Finally,

$$\|G\|^2 = \sum_i \|g_i\|^2 + 2\sum_{i<j} g_i^\top g_j. \tag{60}$$

Subtracting cancels all head terms and proves Eq. 9. □

Explicit remainder. Assume that the operator norm of $\nabla^2 L_i$ is bounded by β_i on both update segments in Eq. 55. Then Eq. 57 implies

$$\left| D_{\mathrm{MH}} - \eta\Big(\|G\|^2 + \sum_i \|h_i\|^2\Big) \right| \le \frac{\eta^2}{2} \sum_i \beta_i (\|G\|^2 + \|h_i\|^2), \tag{61}$$

$$\left| D_{\mathrm{SP}} - \eta \sum_i (\|g_i\|^2 + \|h_i\|^2) \right| \le \frac{\eta^2}{2} \sum_i \beta_i (\|g_i\|^2 + \|h_i\|^2). \tag{62}$$

Define

$$S = \sum_{i<j} g_i^\top g_j, \qquad C = \sum_i \beta_i \big(\|G\|^2 + \|g_i\|^2 + 2\|h_i\|^2\big). \tag{63}$$

The triangle inequality yields

$$|D_{\mathrm{SP}} - D_{\mathrm{MH}} + 2\eta S| \le \frac{\eta^2}{2} C. \tag{64}$$

If $S < 0$, $C > 0$, and the stated bounds hold along the update segments, then

$$0 < \eta < \frac{-4S}{C} \quad \Longrightarrow \quad D_{\mathrm{SP}} - D_{\mathrm{MH}} \ge -2\eta S - \frac{\eta^2 C}{2} > 0. \tag{65}$$

If $C = 0$, the remainder vanishes and the same strict advantage follows for any positive step for which those assumptions hold. Dividing Eq. 64 by η also gives

$$\lim_{\eta \downarrow 0} \frac{D_{\mathrm{SP}} - D_{\mathrm{MH}}}{\eta} = -2S. \tag{66}$$

This is a directly testable virtual-update identity.

Objective grouping. For a partition $\mathcal{P}$ of objectives, give every group its own trunk copy and keep private heads. Repeating the proof independently in each group gives

$$D(\mathcal{P}) = \eta \sum_{G\in\mathcal{P}} \Big\| \sum_{i\in G} g_i \Big\|^2 + \eta \sum_i \|h_i\|^2 + O(\eta^2), \tag{67}$$

$$D_{\mathrm{SP}} - D(\mathcal{P}) = -2\eta \sum_{G\in\mathcal{P}} \sum_{\substack{i<j \\ i,j\in G}} g_i^\top g_j + O(\eta^2). \tag{68}$$

The singleton partition removes these direct cross-objective terms. It need not maximize progress when some aggregate within-group interactions are positive. For $m > 2$, a single negative pair does not determine the sign of the full sum.

Optimizer and smoothness qualifications. If the common trunk step uses a fixed symmetric positive-semidefinite preconditioner P, and every copied trunk uses the same P, the first-order difference becomes $-2\eta \sum_{i<j} g_i^\top P g_j$, provided head updates remain matched. This follows by replacing G by PG and g_i by Pg_i in the first-order expansions. Different adaptive states, momentum, or different per-copy preconditioners need not satisfy this formula. Vanilla-gradient virtual updates therefore match the stated diagnostic most directly. For piecewise-smooth networks, the smooth theorem requires a neighborhood in which the compared losses are twice differentiable; differentiability cannot be assumed across arbitrary activation boundaries. The same loss normalization must be used in both compared updates.

B.4 Generic Objective Mixing and Exact Prediction Dynamics

B.4.1 Fixed-target assumptions and realizability

Let $\xi \sim \mu$ and let the fixed feature map $\varphi(\xi) \in \mathbb{R}^d$ have finite second moments and covariance $\Sigma = \mathbb{E}_\mu[\varphi\varphi^\top] \succ 0$. The target $q^\star(\xi) = A^\star\varphi(\xi)$ is fixed during fitting. Both the features and target are independent of the independently sampled $w \sim \rho$. A terminal decision problem whose observed reward is $A^\star\varphi(\xi)$ is one policy-evaluation instance. The derivation itself does not require two objectives, a particular action count, one-hot features, or terminal transitions.

These assumptions do not make arbitrary $Q_i^{\pi,\alpha}(s,a,w)$ preference-independent. When features, targets, or state–action sampling depend on w, the population operator must be rederived; the factorization below cannot simply be assumed.

For SC and MH, parameterize

$$A = B\Theta, \qquad B \in \mathbb{R}^{m\times p}, \quad \mathrm{rank}(B) = m, \qquad \Theta \in \mathbb{R}^{p\times d}. \tag{69}$$

The entries of B are fixed and those of Θ are trainable. MH outputs $B\Theta\varphi$ and SC outputs the single scalar $w^\top B\Theta\varphi$. Using an effective coefficient matrix for this SC family does not assume that an arbitrary scalar neural network exposes objective outputs.

Set $K = BB^\top$. For any nonzero v, $v^\top K v = \|B^\top v\|^2 > 0$, so $K \succ 0$. Every coefficient matrix A is representable, because

$$\Theta = B^\top K^{-1} A \quad \Longrightarrow \quad B\Theta = BB^\top K^{-1} A = A. \tag{70}$$

This proves realizability for every $A^\star$ and permits any shared initial predictor A_0. For instance, SC and MH can use $\Theta_0 = B^\top K^{-1} A_0$, while direct SP coefficient rows start at A_0. Adding N with $BN = 0$ changes parameters but not their predictions. Conversely, every positive-definite $m \times m$ matrix K with unit diagonal admits such a normalized factorization, for example $B = K^{1/2}$. Thus the class is not restricted to the particular K_χ used later.

B.4.2 What parameter separation changes

Write $b_i = B^\top e_i \in \mathbb{R}^p$ for row i of B as a column. The shared effective coefficient is $A_i^\top = \Theta^\top b_i$. To separate objectives, give each objective its own parameter copy Θ_i and define

$$(A_i^{\mathrm{SP}})^\top = \Theta_i^\top b_i. \tag{71}$$

Initializing every copy to Θ_0 preserves the initial effective predictor. Let $R = \nabla_A \mathcal{L}_{\text{obj}} = E\Sigma$ and $r_i = R_i^\top$. The gradient for copy i is $b_i r_i^\top$, so

$$\Delta(A_i^{\text{SP}})^\top = -\eta(b_i r_i^\top)^\top b_i = -\eta\|b_i\|^2 r_i = -\eta K_{ii} r_i. \tag{72}$$

Before normalization, splitting therefore induces

$$E_{t+1}^{\text{SP}} = E_t^{\text{SP}} - \eta D E_t^{\text{SP}}\Sigma, \qquad D = \operatorname{diag}(K_{11}, \ldots, K_{mm}). \tag{73}$$

The main theorem imposes $K_{ii} = 1$, giving $D = I_m$. This explicitly matches within-objective update sensitivity and removes the off-diagonal coupling. Training independent coefficient rows directly induces the same normalized prediction dynamics. Unit diagonal is a comparison assumption, not a claim that every implemented critic already has this normalization.

If unequal diagonal entries are retained and the mode also obeys $Dv = \delta v$, its split factor is $1-\eta\sigma\delta$, not $1 - \eta\sigma$. The analogous ordering then requires $0 < \kappa < \delta$ and $0 < h < 1$, with all factors positive. Thus unnormalized self-sensitivity differences must not be attributed solely to sharing.

B.4.3 Objective-wise population updates

Write $E = A - A^\star$. The objective-wise loss is

$$\mathcal{L}_{\text{obj}}(A) = \tfrac{1}{2}\mathbb{E}_\mu\|E\varphi\|^2 = \tfrac{1}{2}\operatorname{tr}(E\Sigma E^\top), \qquad \nabla_A \mathcal{L}_{\text{obj}} = E\Sigma. \tag{74}$$

For a perturbation ΔA, symmetry of Σ gives directional derivative $\operatorname{tr}((E\Sigma)^\top \Delta A)$. Substituting $\Delta A = B\Delta\Theta$ proves $\nabla_\Theta \mathcal{L}_{\text{obj}} = B^\top E\Sigma$. One full-population gradient step therefore yields

$$\Theta^+ = \Theta - \eta B^\top E\Sigma, \tag{75}$$

$$A^+ = B\Theta^+ = A - \eta BB^\top E\Sigma, \tag{76}$$

$$E_{\text{MH}}^+ = E - \eta K E\Sigma. \tag{77}$$

Normalized copying, or direct SP row updates, gives

$$E_{\text{SP}}^+ = E - \eta E\Sigma. \tag{78}$$

These are induced prediction-space updates from parameter-space gradient descent, not an assumption that gradient descent on A itself produces K.

B.4.4 Scalarized population updates

For fixed $c_\rho > 0$, set $M_\rho = \mathbb{E}_\rho[ww^\top]$ and $H = c_\rho M_\rho$. Independence of ξ and w gives

$$\mathcal{L}_{\text{sc}}(A) = \frac{c_\rho}{2}\mathbb{E}_{\mu,\rho}(w^\top E\varphi)^2 \tag{79}$$

$$= \frac{c_\rho}{2}\mathbb{E}_\rho[w^\top E\Sigma E^\top w] \tag{80}$$

$$= \frac{1}{2}\operatorname{tr}(HE\Sigma E^\top). \tag{81}$$

Since H and Σ are symmetric,

$$\nabla_A \mathcal{L}_{\text{sc}} = HE\Sigma, \qquad \nabla_\Theta \mathcal{L}_{\text{sc}} = B^\top HE\Sigma. \tag{82}$$

Consequently

$$E_{\text{SC}}^+ = E - \eta BB^\top HE\Sigma = E - \eta KHE\Sigma. \tag{83}$$

No commutation between K and H was used. Their order follows from the chain rule: supervision produces $HE\Sigma$ and parameterization multiplies it on the left by K.

B.5 Proof of the Generalized Finite-Budget Theorem

B.5.1 A common observable prediction-error metric

For all designs, evaluate

$$\mathcal{E}(E)^2 = \mathbb{E}_{\mu,\rho}(w^\top E\varphi)^2 = \operatorname{tr}(M_\rho E\Sigma E^\top). \tag{84}$$

This is a nonnegative quadratic form. Its square root is generally a seminorm and becomes a norm when $M_\rho \succ 0$. The main theorem requires positivity only on the compared mode, rather than on every possible error. For its unit vectors v, z and $E = avz^\top$,

$$E\Sigma E^\top = a^2\sigma vv^\top, \qquad \mathcal{E}(E)^2 = a^2\sigma v^\top M_\rho v = a^2\sigma h/c_\rho. \tag{85}$$

The condition $h > 0$ therefore makes the mode observable. This requires positive second moment $\mathbb{E}[(w^\top v)^2]$, not positive variance: even a constant nonzero projection is observable.

For another fixed evaluation moment matrix M_{eval}, the same modal RMSE ratios hold provided $v^\top M_{\text{eval}} v > 0$, with initial error $|a_0|\sqrt{\sigma v^\top M_{\text{eval}} v}$. Without positivity, the metric is identically zero on this mode and cannot have a strict ordering. The factors r^T below concern RMSE or another square-root quadratic metric; squared error instead has factors r^{2T}.

B.5.2 Exact mode preservation and rates

B.6 Finite-Budget Consequences of Objective Coupling

The local comparison above does not by itself imply a multi-step convergence rate. We now analyze a realizable linear class in which parameter coupling and scalarized supervision induce distinct, exactly tractable update operators. Let $\xi = (s, a) \sim \mu$ and let $\varphi(\xi) \in \mathbb{R}^d$ be fixed features with finite second moments. Assume

$$\Sigma = \mathbb{E}_\mu[\varphi\varphi^\top] \succ 0, \qquad q^\star(\xi) = A^\star\varphi(\xi), \qquad A^\star \in \mathbb{R}^{m\times d}. \tag{86}$$

For this analysis, the target and features do not depend on the sampled preference. This is an additional fixed-target assumption, not a property of arbitrary preference-conditioned values $Q_i^{\pi,\alpha}(s, a, w)$.

Parameter coupling and supervision geometry. For MH and SC, parameterize $A = B\Theta$, where $B \in \mathbb{R}^{m\times p}$ is fixed and full row rank and $\Theta \in \mathbb{R}^{p\times d}$ is trainable. MH outputs $A\varphi$; SC directly outputs $w^\top A\varphi$. Define the coupling operator

$$K = BB^\top \succ 0, \qquad K_{ii} = 1 \quad (i = 1, \dots, m). \tag{87}$$

The unit-diagonal assumption matches each objective's own prediction-update sensitivity. Giving each objective independent parameter copies changes this operator to I_m; training independent coefficient rows induces the same SP dynamics. Without unit diagonal, copying gives $\mathrm{diag}(K)$, not necessarily I_m (Appendix B.4).

Let $w \sim \rho$ on Δ^{m-1} independently of ξ, and fix a loss normalization $c_\rho > 0$. With $M_\rho = \mathbb{E}_\rho[ww^\top]$ and $E = A - A^\star$, define

$$H = c_\rho M_\rho \succeq 0, \tag{88}$$

$$\mathcal{L}_{\text{obj}}(A) = \tfrac{1}{2}\mathbb{E}_\mu\|E\varphi\|_2^2, \qquad \mathcal{L}_{\text{sc}}(A) = \tfrac{c_\rho}{2}\mathbb{E}_{\mu,\rho}(w^\top E\varphi)^2. \tag{89}$$

MH and SP use $\mathcal{L}_{\text{obj}}$, whereas SC uses $\mathcal{L}_{\text{sc}}$. All variants can represent any $A^\star$ and start at the same effective predictor A_0. Full-population gradient descent on their trainable parameters induces

$$E_{t+1}^{\text{SP}} = E_t^{\text{SP}} - \eta E_t^{\text{SP}}\Sigma, \tag{90}$$

$$E_{t+1}^{\text{MH}} = E_t^{\text{MH}} - \eta K E_t^{\text{MH}}\Sigma, \tag{91}$$

$$E_{t+1}^{\text{SC}} = E_t^{\text{SC}} - \eta K H E_t^{\text{SC}}\Sigma. \tag{92}$$

Appendix B.4 derives these recurrences directly. Thus K describes parameter-induced coupling and H describes scalarized supervision; neither necessarily attenuates every direction. For every design, we evaluate the same scalarized-value RMSE,

$$\mathcal{E}_T^a = \left(\mathbb{E}_{\mu,\rho}[\{w^\top E_T^a\varphi(\xi)\}^2]\right)^{1/2}, \qquad a \in \{\text{SC}, \text{MH}, \text{SP}\}. \tag{93}$$

Theorem 5 (Finite-budget cost of objective coupling)**.** *Under the preceding setup, let unit vectors $v \in \mathbb{R}^m$ and $z \in \mathbb{R}^d$ satisfy*

$$Kv = \kappa v, \qquad Hv = hv, \qquad \Sigma z = \sigma z, \qquad \kappa, h, \sigma > 0. \tag{94}$$

Suppose every design starts from $E_0 = a_0 v z^\top$, $a_0 \neq 0$. If

$$0 < \eta\sigma \max\{1, \kappa, \kappa h\} < 1, \tag{95}$$

then, for every integer $T \geq 0$,

$$\mathcal{E}_T^{\mathrm{SP}} = (1 - \eta\sigma)^T \mathcal{E}_0, \tag{96}$$

$$\mathcal{E}_T^{\mathrm{MH}} = (1 - \eta\sigma\kappa)^T \mathcal{E}_0, \tag{97}$$

$$\mathcal{E}_T^{\mathrm{SC}} = (1 - \eta\sigma\kappa h)^T \mathcal{E}_0, \tag{98}$$

where $\mathcal{E}_0 = |a_0|\sqrt{\sigma h / c_\rho} > 0$. In particular, if $0 < \kappa < 1$ and $0 < h < 1$, then

$$\mathcal{E}_T^{\mathrm{SP}} < \mathcal{E}_T^{\mathrm{MH}} < \mathcal{E}_T^{\mathrm{SC}} \qquad (T \geq 1), \tag{99}$$

although all three errors converge to zero on this mode.

Proof sketch. The recurrences preserve $vz^\top$ and multiply its coefficient by $1 - \eta\sigma$, $1 - \eta\sigma\kappa$, and $1 - \eta\sigma\kappa h$, respectively. Iterate and use $\mathcal{E}(E)^2 = \mathrm{tr}(M_\rho E \Sigma E^\top)$. For $0 < \kappa, h < 1$, the three positive factors are strictly ordered. Appendix B.5 provides the full proof, mixed-mode extensions, and global stability conditions. □

Only the error direction must be a common eigenvector; K and H need not commute globally. The eigenvalue κ measures update strength relative to normalized splitting, and h measures supervision strength under the fixed normalization c_ρ. For $\kappa > 1$, MH instead fits this mode faster than SP under the same nonoscillatory step condition. Consequently, the theorem characterizes a conditional finite-budget cost, not a universal ordering of critic architectures.

Corollary 6 (Shared-plus-private critics under uniform preferences)**.** *Let $B_\chi = [\sqrt{\chi}\,\mathbf{1} \;\; \sqrt{1-\chi}\, I_m]$ with $0 < \chi < 1$, $\rho = \mathrm{Dirichlet}(1, \ldots, 1)$, and $c_\rho = m$. Then*

$$K_\chi = \lambda_+ P_+ + \lambda_- P_-, \qquad H = P_+ + \frac{1}{m+1} P_-, \tag{100}$$

where $\lambda_+ = 1 + (m-1)\chi$ and $\lambda_- = 1 - \chi$. On contrast modes, $\kappa = 1 - \chi$ and $h = 1/(m+1)$, giving the factors $1 - \eta\sigma$, $1 - \eta\sigma(1-\chi)$, and $1 - \eta\sigma(1-\chi)/(m+1)$ for SP, MH, and SC. On the common direction, $\kappa = \lambda_+ > 1$ and $h = 1$, so sharing instead accelerates fitting. All comparisons use the step condition in Theorem 5.

Proof sketch. Compute $B_\chi B_\chi^\top$ and apply Proposition 2. Both operators diagonalize into P_+ and P_-, so Theorem 5 applies on each mode. □

B.7 Why Slower Fitting Matters When Targets Change

Keep B, H, Σ fixed and prescribe a nonrandom target sequence $A_t^\star = A_{\mathrm{base}} + a_t^\star v z^\top$ in the mode of Theorem 5. Other coefficients are already fitted. Let $b_t^a = \mathbb{E}[\widehat{a}_t^a] - a_t^\star$ be its mean tracking error. Each round fits the current target once, then the target changes; additive coefficient-update noise may be integrable and conditionally mean zero. Write $r_{\mathrm{SP}} = 1 - \eta\sigma$, $r_{\mathrm{MH}} = 1 - \eta\sigma\kappa$, and $r_{\mathrm{SC}} = 1 - \eta\sigma\kappa h$.

Corollary 7 (Finite-window tracking lag)**.** *Under the mode and step assumptions of Theorem 5, if $b_0^a = 0$ and $a_{t+1}^\star - a_t^\star = d$ for T rounds, then*

$$b_T^a = -d \sum_{k=0}^{T-1} r_a^k = -d \frac{1 - r_a^T}{1 - r_a}. \tag{101}$$

If additionally $0 < \kappa, h < 1$, $d \neq 0$, and $T \geq 2$, then $|b_T^{\mathrm{SP}}| < |b_T^{\mathrm{MH}}| < |b_T^{\mathrm{SC}}|$.

Proof sketch. Fitting the current target and then changing it gives $b_{t+1}^a = r_a b_t^a - d$. Unroll the recurrence. For $T \geq 2$, the finite sum $\sum_{k=0}^{T-1} r^k$ is strictly increasing on $r \in (0, 1)$. □

This is a conditional tracking-bias result for a specified process, not a claim that practical actor–critic targets follow it. SP also has nonzero lag under drift; mean-square error additionally depends on variance. Appendix C.8 proves the result and a sharp finite-window bound for bounded drift.

Finally, objective errors remain relevant to the actor's scalarized value:

$$\left|\sum_i w_i \widehat{Q}_i(x) - Q_w^{\pi,\alpha}(s,a)\right| \le \sum_i w_i \left|\widehat{Q}_i(x) - Q_i^{\pi,\alpha}(x)\right|. \tag{102}$$

Complete proof of Theorem 5. For $E = avz^\top$, the eigenvector assumptions give

$$E\Sigma = \sigma E, \qquad KE\Sigma = \kappa\sigma E, \qquad KHE\Sigma = \kappa h\sigma E. \tag{103}$$

Each exact update therefore preserves $\{avz^\top : a \in \mathbb{R}\}$. Its one-step factor is, respectively,

$$r_{\mathrm{SP}} = 1 - \eta\sigma, \qquad r_{\mathrm{MH}} = 1 - \eta\sigma\kappa, \qquad r_{\mathrm{SC}} = 1 - \eta\sigma\kappa h. \tag{104}$$

Starting from the same E_0, induction proves

$$E_T^a = r_a^T E_0 \quad (a \in \{\mathrm{SC}, \mathrm{MH}, \mathrm{SP}\}). \tag{105}$$

The step condition places every factor in $(0,1)$. Equation 85 then gives $\mathcal{E}_T^a = r_a^T \mathcal{E}_0$ with $\mathcal{E}_0 = |a_0|\sqrt{\sigma h/c_\rho} > 0$. If $0 < \kappa, h < 1$, then $0 < \kappa h < \kappa < 1$, so

$$0 < r_{\mathrm{SP}} < r_{\mathrm{MH}} < r_{\mathrm{SC}} < 1. \tag{106}$$

Positive integer powers preserve the strict ordering for $T \ge 1$ and all tend to zero. This proves the exact rates, the finite-budget ordering, and convergence on the mode without an asymptotic error floor. □

The proof requires a common eigenvector only for the selected direction; global commutation is unnecessary. A generic direction need not remain one-dimensional when K and H do not share it. A mode with $h = 0$ is excluded: SC receives no correction there and the training-preference RMSE cannot observe that mode.

B.7.1 Stability beyond the selected mode

The main theorem's step bound controls the stated mode; it need not stabilize other modes absent from the initialization. A sufficient global bound follows by defining

$$S_K = K^{1/2} H K^{1/2} \succeq 0, \qquad \Lambda = \max\{1, \lambda_{\max}(K), \lambda_{\max}(S_K)\}. \tag{107}$$

Assume

$$0 < \eta < \frac{1}{\sigma_{\max}(\Sigma)\Lambda}. \tag{108}$$

The SP and MH update operators are symmetric positive definite. For SC, $KH = K^{1/2} S_K K^{-1/2}$, even when $KH \neq HK$. If $\Sigma z_j = \sigma_j z_j$, define $x_{t,j} = E_t^{\mathrm{SC}} z_j$ and $u_{t,j} = K^{-1/2} x_{t,j}$. Then

$$u_{t+1,j} = (I_m - \eta\sigma_j S_K) u_{t,j}. \tag{109}$$

Every positive eigenvalue produces a factor in $(0,1)$ under the bound. If $H \succ 0$, all three designs converge in coefficient error from every initial E_0. If H is singular, some SC coefficient modes may remain unidentified, so convergence of all coefficients to $A^\star$ is not asserted. Similarity to a symmetric operator does not imply monotone contraction of the original Euclidean coefficient norm for nonsymmetric KH; no such claim is needed.

B.7.2 Several modes when the operators commute

When $KH = HK$, symmetry gives a common orthonormal basis $\{v_\ell\}_{\ell=1}^m$ with eigenvalues $\kappa_\ell > 0$ and $h_\ell \ge 0$. Let $\{z_j\}_{j=1}^d$ diagonalize Σ with eigenvalues $\sigma_j > 0$, and write $E_0 = \sum_{\ell,j} a_{\ell j} v_\ell z_j^\top$. Define $d_{\mathrm{SP},\ell} = 1$, $d_{\mathrm{MH},\ell} = \kappa_\ell$, and $d_{\mathrm{SC},\ell} = \kappa_\ell h_\ell$. Exact iteration gives

$$E_T^a = \sum_{\ell,j} a_{\ell j} (1 - \eta\sigma_j d_{a,\ell})^T v_\ell z_j^\top. \tag{110}$$

Orthogonality eliminates cross terms in the trace metric:

$$(\mathcal{E}_T^a)^2 = \sum_{\ell,j} \frac{\sigma_j h_\ell}{c_\rho} a_{\ell j}^2 (1 - \eta\sigma_j d_{a,\ell})^{2T}. \tag{111}$$

If every nonzero initial coefficient has $0 < \kappa_\ell < 1$ and $0 < h_\ell < 1$, and all active factors are positive, each nonzero summand has the SP–MH–SC strict ordering for $T \geq 1$. Their sum has the same ordering. A single product-mode initialization is therefore needed for a single exact factor, not for every conditional multi-mode comparison. When active modes favor different designs, no ranking independent of initialization and budget follows.

B.8 Parameter-Count Matching Without Changing the Coupling

General normalized duplication. Let $b^{(j)}$ be column j of B. Replace it by r_j copies $b^{(j)}/\sqrt{r_j}$, each paired with its own trainable parameter row. The resulting matrix satisfies

$$\widetilde{B}\widetilde{B}^\top = \sum_j \sum_{k=1}^{r_j} \frac{b^{(j)}(b^{(j)})^\top}{r_j} = BB^\top = K. \tag{112}$$

For example, initialize the new rows as $\theta_j^\top/\sqrt{r_j}$ to preserve the old contribution $b^{(j)}\theta_j^\top$ to A_0. The objective and scalarized update derivations then show that the full effective trajectories are unchanged under ordinary gradient descent with matched loss normalization and step size.

Equal counts for the controlled construction. Replacing the shared column of B_χ by r normalized copies gives

$$\widetilde{B}_{\chi,r} = \left[\sqrt{\chi/r}\,\mathbf{1}_m\mathbf{1}_r^\top \quad \sqrt{1-\chi}\,I_m\right], \qquad \widetilde{B}_{\chi,r}\widetilde{B}_{\chi,r}^\top = K_\chi. \tag{113}$$

SC and MH then use $(r+m)d$ trainable scalars; the split construction $B_{\mathrm{SP}} = [\sqrt{\chi}I_m\ \sqrt{1-\chi}I_m]$ uses $2md$. Setting $r = m$ gives exactly $2md$ parameters for every design with the same K_χ, initial predictor, and contraction rates as before. For $m = 2, d = 3$, the matched count is 12 per design; without duplication the original SC/MH use 9 and SP uses 12.

This is an algebraic count-matching construction, not a report that a new parameter-matched benchmark has been run. Duplicated shared columns remain redundant in prediction space; they do not add the independently adjustable objective directions of splitting. The result rules out raw parameter count as a necessary explanation of these particular linear dynamics, but does not replace width-, learning-rate-, or compute-matched comparisons of nonlinear critics. Fixed B and fixed features are not trainable parameters.

C Proof of the Nonlinear Sharing-Error Theorem

This appendix proves Theorem 4 for paired nonlinear gradient-descent trajectories. All vectors use Euclidean inner products; matrix norms are the associated induced operator norms. The comparison changes trunk sharing while holding per-step loss functions, loss normalization, step sizes, and cloned initial parameters fixed. No claim of equal parameter count or wall-clock computation is made.

C.1 Exact representation of the paired updates

Suppose the trunk has p parameters and private head i has r_i parameters. The shared implementation trains one trunk and m heads, while SP trains m independent copies of that trunk with the corresponding heads. For comparison, both are represented in dimension $mp + \sum_i r_i$:

$$\Theta = \mathrm{col}_{i=1}^m(\psi_i, \nu_i), \qquad F_t(\Theta) = \sum_i L_{i,t}(\psi_i, \nu_i).$$

The lifted MH vector repeats a single shared trunk value in all trunk blocks. This is a coordinate representation, not an increase in MH's independently trainable parameters. In particular, the shared

update is not defined as a Euclidean orthogonal projection of the split update; such a projection would introduce a factor $1/m$ into the trunk step. We use the actual gradient step on the summed objective-wise loss.

Along the shared trajectory, define

$$g_{i,t} = \nabla_{\psi} L_{i,t}(\psi_t^{\mathrm{MH}}, \nu_{i,t}^{\mathrm{MH}}), \quad h_{i,t} = \nabla_{\nu_i} L_{i,t}(\psi_t^{\mathrm{MH}}, \nu_{i,t}^{\mathrm{MH}}), \quad G_t = \sum_i g_{i,t}.$$

The actual MH implementation updates

$$\psi_{t+1}^{\mathrm{MH}} = \psi_t^{\mathrm{MH}} - \eta_t G_t, \qquad \nu_{i,t+1}^{\mathrm{MH}} = \nu_{i,t}^{\mathrm{MH}} - \eta_t h_{i,t}.$$

In the lifted space,

$$\nabla F_t(\Theta_t^{\mathrm{MH}}) = \mathrm{col}_i(g_{i,t}, h_{i,t}), \qquad c_t = \mathrm{col}_i(G_t - g_{i,t}, 0_{\nu_i}).$$

Adding these vectors gives $\mathrm{col}_i(G_t, h_{i,t})$. Consequently,

$$\Theta_{t+1}^{\mathrm{MH}} = \Theta_t^{\mathrm{MH}} - \eta_t \nabla F_t(\Theta_t^{\mathrm{MH}}) - \eta_t c_t, \tag{114}$$

$$\Theta_{t+1}^{\mathrm{SP}} = \Theta_t^{\mathrm{SP}} - \eta_t \nabla F_t(\Theta_t^{\mathrm{SP}}). \tag{115}$$

The shared vector stays on the shared diagonal because every trunk block receives the same G_t. SP updates each block at its own actual parameters; it is not reset to MH between updates. Cloning the initial trunk and heads ensures equality of the two initial lifted parameter vectors. Equality of initial predictions alone would not ensure this parameter equality.

C.2 Exact transport of the parameter difference

Define $d_t = \Theta_t^{\mathrm{MH}} - \Theta_t^{\mathrm{SP}}$. For each $t < T$, apply the fundamental theorem of calculus componentwise to the gradient along the segment between the two current parameter vectors:

$$\nabla F_t(\Theta_t^{\mathrm{MH}}) - \nabla F_t(\Theta_t^{\mathrm{SP}}) = \int_0^1 \nabla^2 F_t(\Theta_t^{\mathrm{SP}} + s d_t) d_t \, ds \tag{116}$$

$$= \overline{\mathcal{H}}_t d_t. \tag{117}$$

The stated C^2 condition makes this integral well-defined and the identity exact. The loss may differ from the loss at another time step; only equality between the two runs at the same step is required.

Subtracting Eq. 115 gives

$$d_{t+1} = (I - \eta_t \overline{\mathcal{H}}_t) d_t - \eta_t c_t = A_t d_t - \eta_t c_t. \tag{118}$$

For $0 \le k \le T$, define

$$\Phi_{T,k} = A_{T-1} A_{T-2} \cdots A_k, \qquad \Phi_{T,T} = I.$$

Induction on T yields

$$d_T = \Phi_{T,0} d_0 - \sum_{t=0}^{T-1} \eta_t \Phi_{T,t+1} c_t. \tag{119}$$

Indeed, substituting this expression into the recursion premultiplies all existing terms by A_T and appends the new term $-\eta_T c_T$. This proves the next-step expression without commuting any factors. For the theorem's cloned initialization, $d_0 = 0$.

There is no discarded Taylor remainder: the integrated Hessians encode nonlinearity exactly along the connecting segments. They depend on both actual trajectories. The identity neither assumes that their norms are small nor guarantees contraction or bounded iterates at arbitrary step sizes. It applies to a finite, well-defined trajectory that satisfies the stated segment smoothness assumptions.

C.3 Prediction residual and complete proof of Theorem 4

For a fixed finite evaluation set, let

$$\mathcal{Q}_\ell(\Theta) = \sqrt{\omega_\ell} \sum_i w_{\ell i} q_i(x_\ell; \psi_i, \nu_i), \qquad y_\ell = \sqrt{\omega_\ell}\, y_\ell^\star.$$

Here $y_\ell^\star$ is the same reference value for both runs. The weights ω_ℓ sum to one, so $\|\mathcal{Q}(\Theta)-y\|_2^2$ is their common weighted scalarized-value MSE. The inputs may include preferences; no independence assumption on inputs and preferences is needed. One may instead stack objective-wise predictions with suitable weights; the proof is unchanged.

Complete proof. The continuously differentiable evaluation map satisfies

$$\mathcal{Q}(\Theta_T^{\mathrm{MH}}) - \mathcal{Q}(\Theta_T^{\mathrm{SP}}) = \int_0^1 D\mathcal{Q}(\Theta_T^{\mathrm{SP}} + s d_T) d_T \, ds \tag{120}$$

$$= \overline{J}_T d_T. \tag{121}$$

Subtracting the same reference y from both predictions and inserting Eq. 119 with $d_0 = 0$ gives

$$e_T^{\mathrm{MH}} - e_T^{\mathrm{SP}} = -\sum_{t=0}^{T-1} \eta_t \overline{J}_T \Phi_{T,t+1} c_t = b_T^{\mathrm{share}}. \tag{122}$$

This proves Eq. 13. Expanding the squared Euclidean norm gives

$$\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}} = \|e_T^{\mathrm{SP}} + b_T^{\mathrm{share}}\|_2^2 - \|e_T^{\mathrm{SP}}\|_2^2 \tag{123}$$

$$= 2\langle e_T^{\mathrm{SP}}, b_T^{\mathrm{share}} \rangle + \|b_T^{\mathrm{share}}\|_2^2. \tag{124}$$

If the inner product is nonnegative and $b_T^{\mathrm{share}} \neq 0$, the right-hand side is at least $\|b_T^{\mathrm{share}}\|_2^2 > 0$, proving the final assertion. No part of this proof requires the training loss to equal the evaluation MSE, or the reference to be representable by either critic. □

Initial-parameter mismatch. If only the initial predictions are matched, retain the additional term $\overline{J}_T \Phi_{T,0} d_0$ in the residual difference. Even if the predictions coincide initially, differing parameters can lead to differing Jacobians and subsequent dynamics. The source-only theorem therefore assumes cloned parameters, not merely a common initial output.

A moving common reference. The final reference can be y_T from a common, externally prescribed sequence. Because the same y_T is subtracted in both runs, it cancels in the prediction difference. This does not establish a tracking rate. If the two runs instead use distinct final references, the additional residual includes $-(y_T^{\mathrm{MH}} - y_T^{\mathrm{SP}})$.

C.4 When the additional residual increases error

Write $e = e_T^{\mathrm{SP}}$ and $b = b_T^{\mathrm{share}}$. The exact condition for strictly positive excess MSE is

$$\langle e, b \rangle > -\tfrac{1}{2}\|b\|_2^2. \tag{125}$$

Nonnegative alignment is a sufficient condition, not an assumption of the residual identity. More generally, if for some $\beta \in [0, 1)$,

$$\langle e, b \rangle \geq -\tfrac{\beta}{2}\|b\|_2^2,$$

then

$$\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}} \geq (1-\beta)\|b\|_2^2.$$

If $\langle e, b \rangle = 0$, the additional MSE is exactly $\|b\|_2^2$. A different sufficient condition is $\|b\|_2 > 2\|e\|_2$, since Cauchy–Schwarz yields

$$\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}} \geq \|b\|_2(\|b\|_2 - 2\|e\|_2) > 0.$$

These are conditional statements, not claims that every shared critic satisfies them. The residual can vanish through cancellation, lie in an evaluation-null direction, or improve an existing prediction error.

Random paired trajectories. The identities hold pathwise for the same realized minibatch and target schedule when each realized loss satisfies the smoothness assumptions. If the two residuals have finite second moments, taking expectations gives

$$\mathbb{E}[\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}}] = 2\mathbb{E}\langle e_T^{\mathrm{SP}}, b_T^{\mathrm{share}} \rangle + \mathbb{E}\|b_T^{\mathrm{share}}\|_2^2.$$

The cross expectation being zero is sufficient for a nonnegative expected gap. Marginal mean zero of b_T^{share} alone does not imply that condition. Conditional mean zero given e_T^{SP} would suffice, but is not assumed. Statistical bias is distinct from a pathwise prediction residual: the identity $\mathbb{E}[e_T^{\mathrm{MH}}] = \mathbb{E}[e_T^{\mathrm{SP}}] + \mathbb{E}[b_T^{\mathrm{share}}]$ does not order bias magnitudes.

C.5 A MAGNITUDE BOUND AND THE LINK TO ONE-STEP INTERFERENCE

Suppose $\|\nabla^2 F_t\|_{\mathrm{op}} \le L_t$ on the relevant parameter segments and $\|D\mathcal{Q}\|_{\mathrm{op}} \le M_T$ on the final segment. Then

$$\|A_t\|_{\mathrm{op}} \le 1 + \eta_t L_t, \qquad \|\overline{J}_T\|_{\mathrm{op}} \le M_T,$$

and submultiplicativity gives

$$\|b_T^{\mathrm{share}}\|_2 \le M_T \sum_{t=0}^{T-1} \eta_t \|c_t\|_2 \prod_{k=t+1}^{T-1} (1 + \eta_k L_k). \tag{126}$$

Tighter bounds $\|A_t\|_{\mathrm{op}} \le \rho_t$ may replace the factors in the product. They are not required to be below one. An upper bound $\|b_T^{\mathrm{share}}\|_2 \le B_T$ gives

$$|\mathcal{M}_T^{\mathrm{MH}} - \mathcal{M}_T^{\mathrm{SP}}| \le 2\|e_T^{\mathrm{SP}}\|_2 B_T + B_T^2,$$

but neither bound determines the sign of the actual error difference.

At a cloned parameter point, for one fixed loss F, set $\Theta^+_{\mathrm{own}} = \Theta - \eta \nabla F(\Theta)$ and $\Theta^+_{\mathrm{shared}} = \Theta - \eta\{\nabla F(\Theta) + c\}$. Taylor expansion gives

$$F(\Theta^+_{\mathrm{shared}}) - F(\Theta^+_{\mathrm{own}}) = -\eta \langle \nabla F(\Theta), c\rangle + O(\eta^2) \tag{127}$$

$$= -2\eta \sum_{i<j} \langle g_i, g_j \rangle + O(\eta^2). \tag{128}$$

The zero head entries of c remove all head contributions, and each pair of trunk gradients is counted twice. This recovers the local training-loss comparison. For a different smooth evaluation loss V, the first-order term is instead $-\eta\langle \nabla V(\Theta), c\rangle$; training-gradient conflict alone therefore need not increase true-value MSE. For an individual prediction,

$$q_i^{\mathrm{MH},+}(x) - q_i^{\mathrm{SP},+}(x) = -\eta D_\psi q_i(x) \sum_{j \ne i} g_j + o(\eta).$$

If that prediction map is also C^2 near the point, the remainder can be written $O(\eta^2)$. The finite-trajectory theorem transports the corresponding update source using actual evolving curvature rather than repeating a fixed first-order approximation.

C.6 LOSS MISMATCH, OPTIMIZER CHANGES, AND SCOPE

If the two runs use different per-step losses, define the lifted losses F_t^{MH} and F_t^{SP}, and set

$$s_t = \nabla F_t^{\mathrm{MH}}(\Theta_t^{\mathrm{MH}}) - \nabla F_t^{\mathrm{SP}}(\Theta_t^{\mathrm{MH}}).$$

Using the integrated Hessian of F_t^{SP} along the connecting segment gives

$$d_{t+1} = A_t^{\mathrm{SP}} d_t - \eta_t (c_t^{\mathrm{MH}} + s_t).$$

There are now distinct transported sources for direct sharing and for loss mismatch. Different actors, replay samples, target-network values, or discrete twin-selection outcomes can change the loss function and contribute to s_t. These may be indirect consequences of sharing, but cannot be silently assigned to c_t alone. Targets being stopped in backpropagation does not by itself make them identical across the runs.

The displayed dynamics use ordinary gradient descent. Momentum, Adam, unequal preconditioners, or different optimizer states need an appropriate augmented-state analysis and are not covered without modification. The smoothness conditions concern the actual connecting segments; ReLU activation crossings or other nonsmooth operations are not automatically included in the C^2 statement. No target realizability, convexity, fixed-kernel, preference-independence, or contraction assumption is needed for the finite identity, provided the finite trajectories and integrals are well-defined. None of this establishes a universal performance ranking or an exact rate for arbitrary nonlinear critics.

Finally, arbitrary SC networks cannot be obtained from MH merely by untying shared trunks. Scalarized-versus-objective-wise supervision remains a separate comparison. The earlier linear SC/MH/SP rate analysis can be retained as an explicitly structured supporting result; its formulas do not follow from this nonlinear decomposition alone.

C.7 Connection to the General Local Interference Result

For objective-wise loss $F_i(\Theta) = \mathbb{E}_\mu[(E_i\varphi)^2]/2$, write $R = E\Sigma$, $r_i = R_i^\top$, and $b_i = B^\top e_i$. Then

$$\nabla_\Theta F_i = b_i r_i^\top, \qquad \langle \nabla_\Theta F_i, \nabla_\Theta F_j \rangle_F = K_{ij}\, r_i^\top r_j. \tag{129}$$

Thus coupling and residual directions together determine the sign of cross-objective interaction; a nonzero K_{ij} alone does not determine it. Since $K_{ii} = 1$, expansion of the squared gradient sum gives

$$-2\sum_{i<j}\langle \nabla_\Theta F_i, \nabla_\Theta F_j \rangle_F = \sum_i \|\nabla_\Theta F_i\|_F^2 - \Big\| \sum_i \nabla_\Theta F_i \Big\|_F^2 \tag{130}$$

$$= \mathrm{tr}\{R^\top (I_m - K) R\}. \tag{131}$$

On $E = avz^\top$ with $Kv = \kappa v$ and $\Sigma z = \sigma z$,

$$-2\sum_{i<j}\langle \nabla_\Theta F_i, \nabla_\Theta F_j \rangle_F = (1-\kappa)\|E\Sigma\|_F^2. \tag{132}$$

Hence $\kappa < 1$ corresponds exactly to negative aggregate gradient interaction on the selected mode. The leading SP progress advantage from the local theorem is $\eta(1-\kappa)\|E\Sigma\|_F^2$. The generalized multi-step theorem quantifies its persistence under fixed linear dynamics, without summing an unchanged Taylor expansion along arbitrary nonlinear trajectories.

For the original shared-plus-private model, the trunk-only gradients are $g_i = \sqrt{\chi}(E\Sigma)_i^\top$. If $P_- E = E$, then $\mathbf{1}^\top E = 0$, so

$$\sum_i g_i = 0, \qquad \sum_{i<j} g_i^\top g_j = -\frac{\chi}{2}\|E\Sigma\|_F^2 < 0 \quad (E \neq 0). \tag{133}$$

The private heads still learn. Parameter-gradient cancellation therefore does not imply that MH is permanently unable to fit the target. Normalized duplication preserves this aggregate interaction.

C.8 Generalized Tracking and Bounded-Drift Guarantees

Keep B, H, Σ fixed and prescribe $A_t^\star = A_{\mathrm{base}} + a_t^\star v z^\top$ for the eigenmode in Theorem 5. Other coefficients are already fitted. The sequence is nonrandom, and the following scalar process concerns its mode coefficient, not the full feedback dynamics of a learned actor. For design a, the coefficient update is

$$\widehat{a}_{t+1}^a = \widehat{a}_t^a - \eta\omega_a(\widehat{a}_t^a - a_t^\star) + \zeta_{t+1}^a, \qquad \begin{cases} \omega_{\mathrm{SP}} = \sigma, \\ \omega_{\mathrm{MH}} = \sigma\kappa, \\ \omega_{\mathrm{SC}} = \sigma\kappa h. \end{cases} \tag{134}$$

Assume finite first moments and $\mathbb{E}[\zeta_{t+1}^a \mid \mathcal{F}_t] = 0$; deterministic fitting sets this noise to zero. Each round updates against $a_t^\star$ and then compares with $a_{t+1}^\star$, fixing the time index of the lag. Set $e_t^a = \widehat{a}_t^a - a_t^\star$, $d_t = a_{t+1}^\star - a_t^\star$, and $r_a = 1 - \eta\omega_a \in (0,1)$. Subtracting the new target and taking expectations gives

$$e_{t+1}^a = r_a e_t^a - d_t + \zeta_{t+1}^a, \qquad b_{t+1}^a = r_a b_t^a - d_t, \qquad b_t^a = \mathbb{E}[e_t^a]. \tag{135}$$

Induction yields

$$b_T^a = r_a^T b_0^a - \sum_{t=0}^{T-1} r_a^{T-1-t} d_t. \tag{136}$$

Complete proof of Corollary 7. With $b_0^a = 0$ and $d_t = d$, reverse the finite sum to get $b_T^a = -d\sum_{k=0}^{T-1} r_a^k = -d(1-r_a^T)/(1-r_a)$. Under $0 < \kappa, h < 1$, the theorem gives $0 < r_{\mathrm{SP}} < r_{\mathrm{MH}} < r_{\mathrm{SC}} < 1$. For $T \geq 2$, the derivative of this geometric sum is $\sum_{k=1}^{T-1} k r^{k-1} > 0$. Its strict increase proves the ordering of absolute lags when $d \neq 0$. At $T = 1$, all lags equal $|d|$, so strict ordering starts only at $T = 2$. □

Sharp finite-window bound for bounded drift. If $|d_t| \le \overline{d}$, the triangle inequality gives

$$|b_T^a| \le r_a^T |b_0^a| + \overline{d} \sum_{k=0}^{T-1} r_a^k. \tag{137}$$

When $b_0^a = 0$, choosing same-sign constant drift attains the bound:

$$\sup_{|d_0|,\ldots,|d_{T-1}| \le \overline{d}} |b_T^a| = \overline{d} \frac{1 - r_a^T}{1 - r_a}. \tag{138}$$

All coefficients in the sum are positive, so taking every $d_t = \overline{d}$ or every $d_t = -\overline{d}$ gives equality. The worst-case envelopes are strictly ordered for $\overline{d} > 0$ and $T \ge 2$. This is not a pointwise ordering for every arbitrary sign-changing drift sequence.

Bias, variance, and scope. With finite second moments,

$$\mathbb{E}[(e_T^a)^2] = (b_T^a)^2 + \mathrm{Var}(\widehat{a}_T^a). \tag{139}$$

The corollary compares conditional mean lags, not arbitrary update variances or total MSE. Constant drift is required only over a finite window, not indefinitely in a bounded-reward MDP. Actor updates, replay shifts, target networks, and reward normalization can change both the target and the fitting operators, beyond the fixed-B, H, Σ model. SP also has nonzero lag under nonzero drift.

D Implementation Details and Additional Result

Table 3: MuJoCo environments in the proposed evaluation. All tasks have two reward objectives and an episode limit of 1,000 transitions. Dimensions refer to the instantiated environments used in our current implementation.

Environment	Environment identifier	Observation	Action
Ant	`mo-ant-2obj-v5`	105	8
HalfCheetah	`mo-halfcheetah-v5`	17	6
Walker2d	`mo-walker2d-v5`	17	6
Hopper	`mo-hopper-2obj-v5`	11	3
Swimmer	`mo-swimmer-v5`	8	2
Humanoid	`mo-humanoid-v5`	348	17

Interaction budget and evaluation. The training cap is 200,000 transitions per algorithm run, with evaluation every 10,000 training transitions and at the final checkpoint. For methods with five policies, this is the total budget across the population, not 200,000 transitions per policy. MO-MPO assigns 40,000 transitions to each of its five policies. We use paired training seeds $\{4750, 4751, 4752\}$, and identical evaluation episode seeds across the architectures within each environment. Each candidate policy is evaluated for five episodes. GPI-LS retains its native early stopping; actual training steps and algorithmic policy-selection interactions are recorded. Curves are not extended with fabricated points after termination. Learning-curve comparisons use explicitly stated common observed checkpoints, while native-final results are labeled separately.

Code. `https://anonymous.4open.science/r/morl-study-EFC2/README.md`

Computational resources. The main training experiments were conducted on a Slurm-managed GPU cluster equipped with AMD Instinct MI355X accelerators, with eight GPUs per node. Each training run used a single GPU, while independent seeds, environments, and architecture configurations were executed in parallel across available devices. Neural-network training used PyTorch 2.9.1 with ROCm 7.0, with environment simulation performed on CPUs. The eight-GPU batches used for the multi-method architecture comparisons requested 32 CPU cores and 512 GiB of host memory per node.

CAPQL. CAPQL natively uses a vector-valued critic with objective-wise TD supervision. To instantiate the theoretical SC design, we instead scalarize the reward and target value under the sampled preference and train a single shared critic on the resulting scalar TD target. For MH, we retain objective-wise TD supervision and use a shared feature extractor followed by objective-specific heads. SP uses the same objective-wise targets as MH but assigns a dedicated critic path for every objective with a smaller size . All actor updates, entropy regularization, twin-critic reduction, replay, and preference sampling are otherwise unchanged.

Momba. Momba's native implementation uses a scalar distributional critic. We compare this reference against Multi-head and Split vector-valued critics. Multi-head retains a shared Simba feature extractor while assigning a separate categorical prediction head to each objective. Split instead assigns a dedicated Simba critic (Lee et al., 2025a;b), including feature extraction and categorical prediction, to every objective, with reduced width so that the total parameter count matches Multi-head.

For matched Multi-head and Split comparisons, we preserve observation normalization, reward normalization, categorical support, automatic entropy adjustment, and preference sampling. Both variants also retain group-wise twin selection: the preference-weighted value is used to choose one of the twin critics, and the complete objective vector from the twin is then used by the actor. The actor subsequently combines the objective-wise values using the sampled preference weights.

The native scalar critic requires a separate interpretation. Because it directly predicts a scalarized return rather than an objective-wise value vector, its comparison with Multi-head and Split changes not only parameter sharing but also the supervision target and value representation. We therefore treat the Multi-head–Split comparison as the clean architectural comparison, while reporting the native scalar implementation as the original algorithmic reference.

MORL/D with MOSAC. For SC, each MOSAC learner retains scalarized TD supervision under its associated preference. For MH and SP, we decompose this target into objective-wise TD targets while preserving the same policy, replay data, and preference-conditioned actor update. MH shares an initial critic representation across objectives, whereas SP assigns a fully independent critic to every objective, with reduced parameter size. Consequently, SC→MH changes supervision, while MH→SP changes only parameter sharing, matching the theoretical comparison.

GPI-LS. The native GPI-LS implementation uses fully shared vector-valued critics with objective-wise TD supervision. To instantiate SC, we scalarize the reward and target value under the current preference and train a single scalar critic on the resulting scalar TD target. For MH, we retain objective-wise TD supervision with a shared initial representation followed by objective-specific critic paths. SP uses the same objective-wise targets as MH but assigns a dedicated critic path to every objective, with hidden widths chosen so that the total parameter count matches MH.

All other components of GPI-LS are preserved, including prioritized replay, policy updates, generalized policy improvement, linear-support weight selection, normalization, and dropout. Thus, SC→MH changes the supervision from scalarized to objective-wise, whereas MH→SP removes the remaining cross-objective parameter sharing.

MO-MPO. The native TD(0) critic corresponds to our MH design: an input block is shared across objectives, followed by objective-specific hidden networks and outputs. To instantiate SC, we replace the objective-wise critic with a scalar critic trained on the preference-weighted TD target while retaining the original MO-MPO policy-improvement and projection procedure. SP uses the same objective-wise targets as MH but assigns a dedicated critic path to every objective, with hidden widths chosen so that the total parameter count matches MH.

All other components of MO-MPO are retained, including objective-specific policy-improvement steps, Gaussian policy projection, KL constraints, dual optimization, and the policy population. The resulting SC–MH–SP comparison therefore follows the same theoretical decomposition: SC–MH changes the supervision, whereas MH–SP changes parameter sharing.

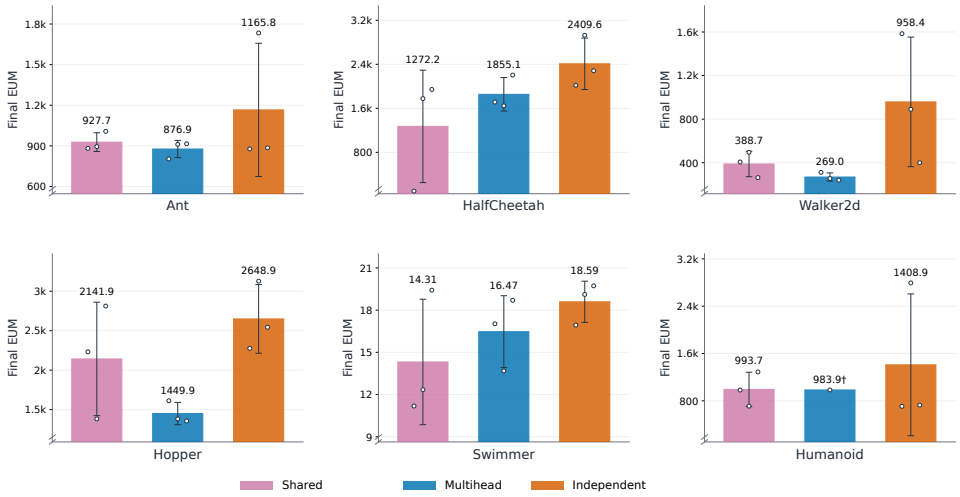


Figure 6: Performance of three Variations of CAPQL.

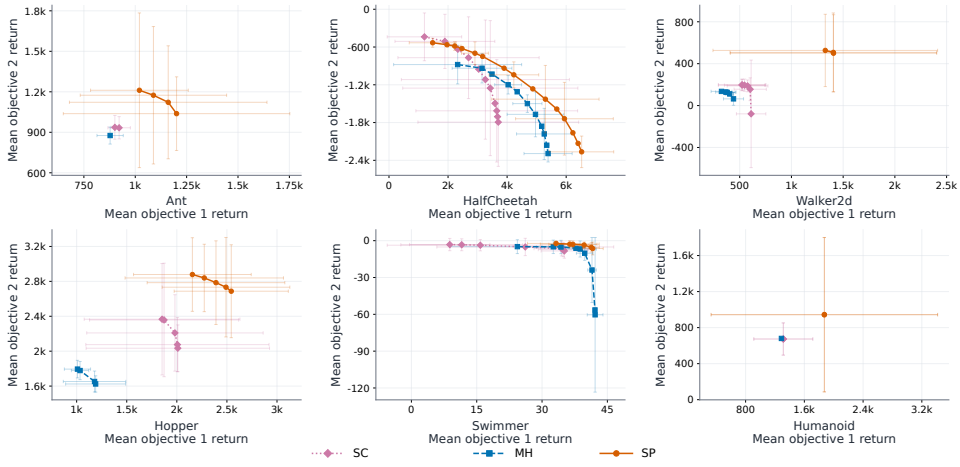


Figure 7: Pareto fronts of three Variations of CAPQL.

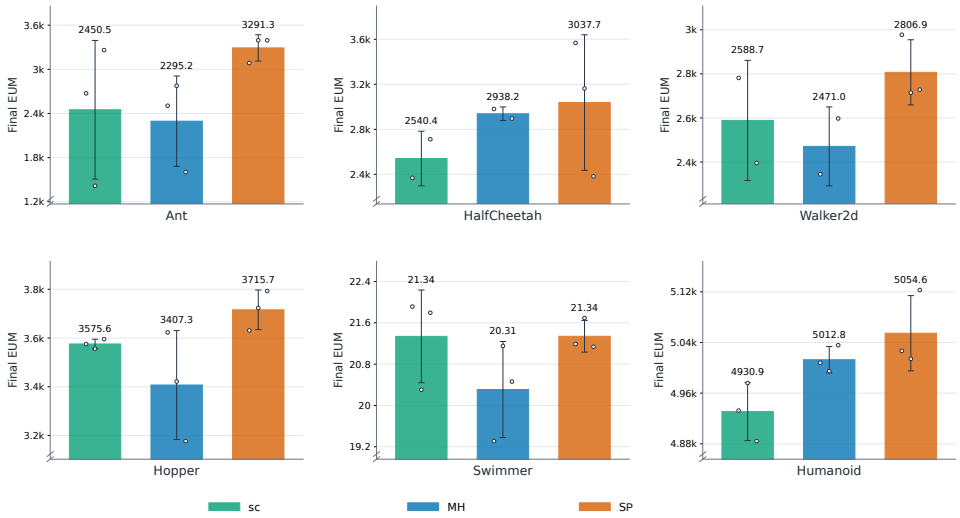


Figure 8: Performance of three Variations of Momba.

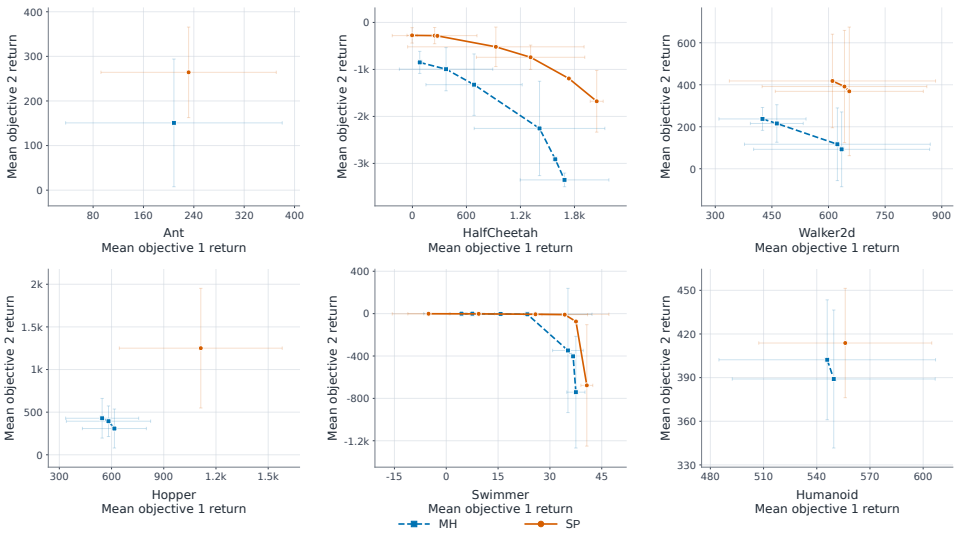


Figure 9: Pareto fronts of three Variations of MO-MPO.

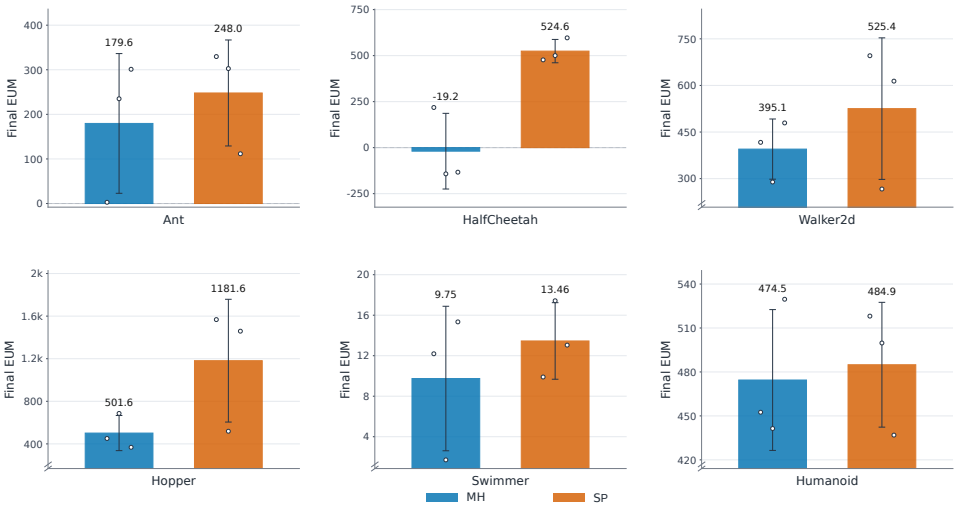


Figure 10: Performance of three Variations of MO-MPO.

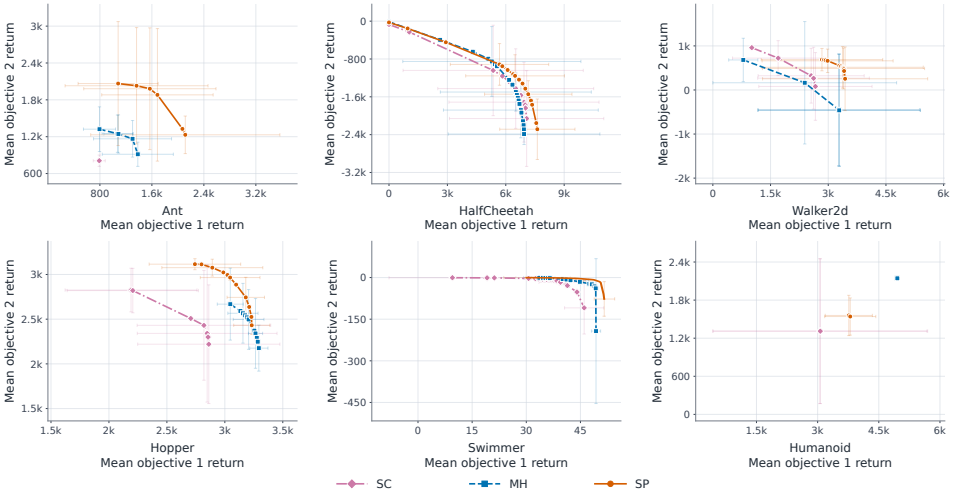


Figure 11: Pareto fronts of three Variations of GPI-LS.

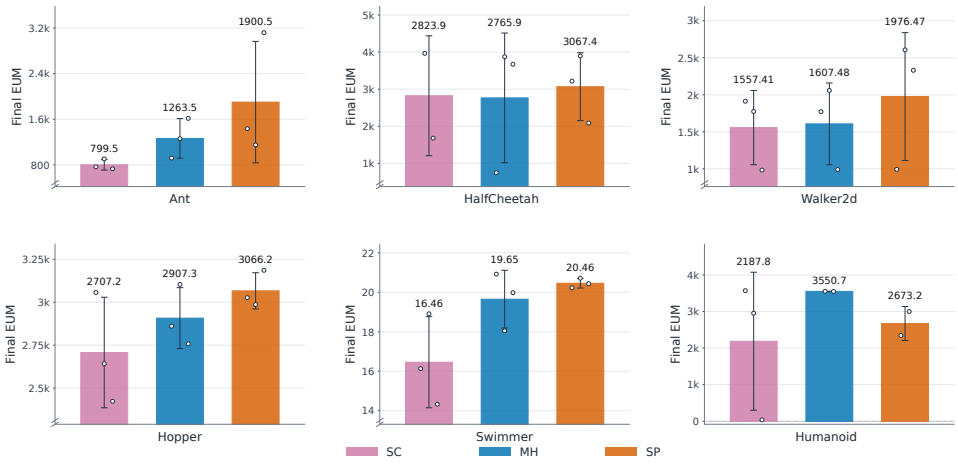


Figure 12: Performance of three Variations of GPI-LS.

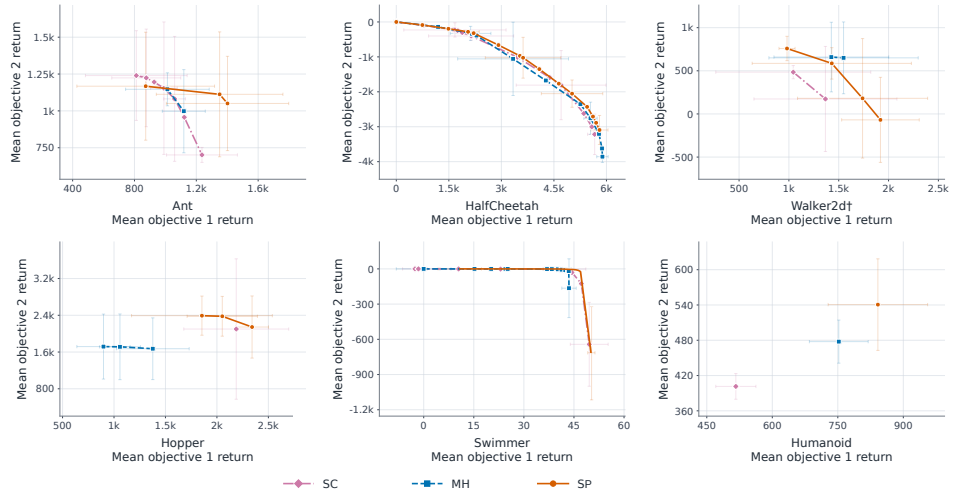


Figure 13: Pareto fronts of three Variations of MORL/D.

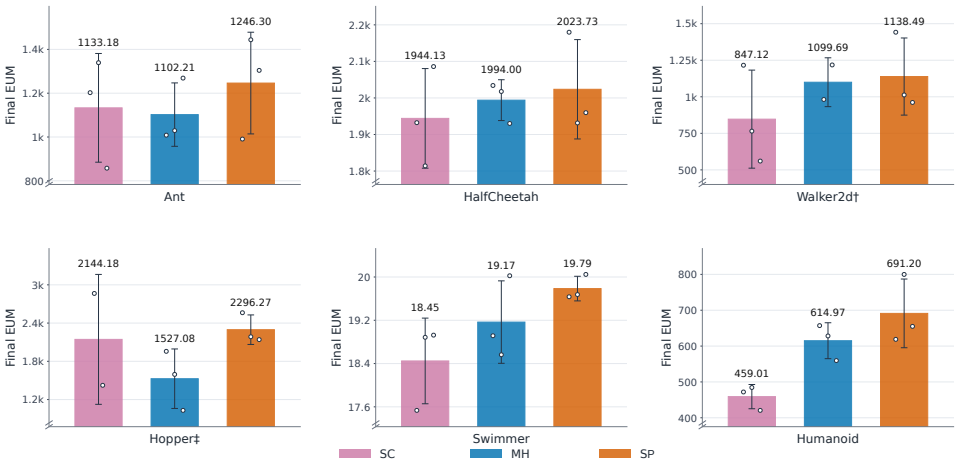


Figure 14: Performance of three Variations of MORL/D.

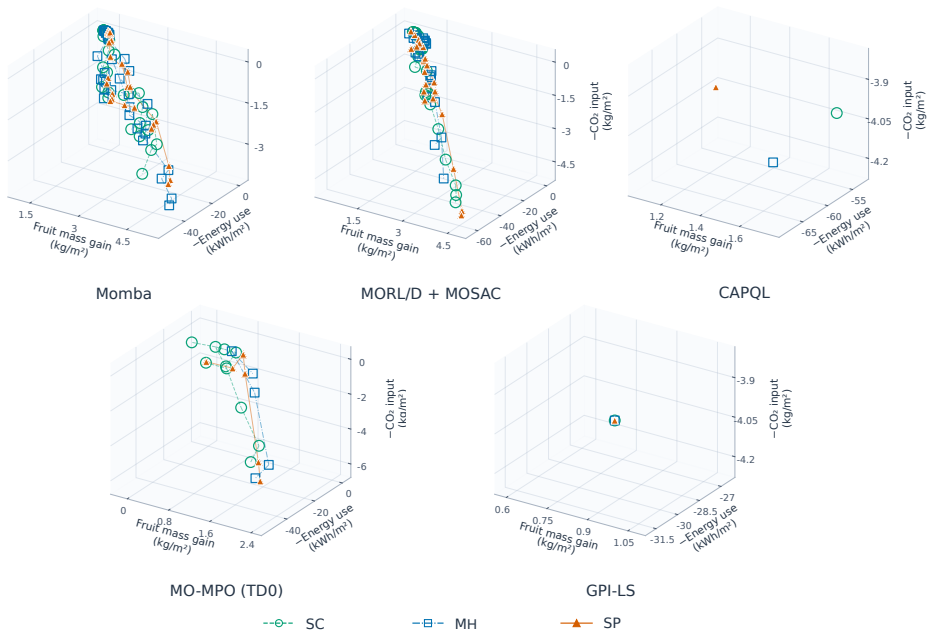


Figure 15: Pareto fronts of GreenLight environment with different algorithm.

D.1 A Ablation of Critic Fitting

We conduct a small, targeted ablation to illustrate the conditional predictions of our analysis in neural critics.

Local gradient interference. Among 80 sampled probes from saved MuJoCo checkpoints at 100k and 200k steps, we examine the 44 probes with negative inter-objective trunk-gradient inner product: We freeze the Bellman targets and create virtual Split (SP) copies of the multi-head (MH) trunk, preserving the initial predictions and matching the head updates. Let D_{SP} and D_{MH} denote the respective decreases in the summed critic loss. At the smallest prespecified SGD step, all 44 probes satisfy $D_{\mathrm{SP}} > D_{\mathrm{MH}}$, consistent with

$$\lim_{\eta \to 0} \frac{D_{\mathrm{SP}} - D_{\mathrm{MH}}}{\eta} = -2 g_1^\top g_2 > 0. \tag{140}$$

Here g_i is the gradient of objective i's critic loss with respect to the shared trunk. The median relative slope error is 0.0263%. These are repeated local probes, not independent training runs.

Estimating objective differences. We additionally report Swimmer as a post-hoc case study from our six-environment Momba ablation. For each frozen policy, MH and SP critics start with identical predictions and receive the same Monte-Carlo soft-return data and 3,000 fitting updates. Using held-out episodes, we decompose the normalized return-prediction errors into common and contrast components, $e_+ = (e_1 + e_2)/\sqrt{2}$ and $e_- = (e_1 - e_2)/\sqrt{2}$. Table 4 shows that SP improves contrast estimation more than common-component estimation under both policy sources. Contrast RMSE decreases in all three checkpoints within each cohort, providing qualitative support for the objective-difference interpretation.

Table 4: SP's relative reduction in mean held-out RMSE compared with MH. Each policy-source cohort contains three checkpoints, averaged equally.

Policy source	Common reduction	Contrast reduction	Contrast wins
SP actor	10.1%	33.2%	$3/3$
MH actor	8.7%	36.7%	$3/3$

E OBJECTIVE-SPECIFIC CRITICS IN MOMBA

Building on the objective-wise soft values defined for Momba, we now specify their approximation and use in policy improvement. Split permits late scalarization, but does not determine whether the corresponding critics should share trainable parameters. Our method uses two categorical critics and their target copies for each objective, with disjoint trainable parameters across objectives. These critics evaluate the same preference-conditioned actor and remain coupled through the replay data, sampled actions, and group-consistent target selection. Each objective receives its own unweighted cross-entropy supervision. Preference weights enter groupwise twin selection and policy improvement, but do not rescale the objective-specific critic losses.

E.1 DECOUPLED DISTRIBUTIONAL CRITICS

For each objective $i \in \{1, \dots, m\}$, we maintain two distributional critics, $Z_{i,1}(s, a, \mathbf{w}; \theta_{i,1}), Z_{i,2}(s, a, \mathbf{w}; \theta_{i,2})$, together with target parameters $\bar{\theta}_{i,1}, \bar{\theta}_{i,2}$. Critics belonging to different objectives share **no trainable parameters**. They nevertheless remain algorithmically coupled through the common actor, replay buffer, sampled preference, transition batch, next action, and twin-selection rule. Each critic predicts a categorical distribution over the fixed support $\mathcal{Z} = \{z_\ell\}_{\ell=0}^{K-1}$, Let $p_{i,k,\ell}(s, a, \mathbf{w}) = \Pr\left(Z_{i,k} = z_\ell \mid s, a, \mathbf{w}\right)$. The corresponding expected value is

$$Q_{i,k}(s, a, \mathbf{w}) = \sum_{\ell=0}^{K-1} z_\ell p_{i,k,\ell}(s, a, \mathbf{w}). \tag{141}$$

The distributional representation is used as the learning target, while policy improvement uses its expectation. We do not assume that the marginal distributions $Z_1, \dots, Z_m$ recover the joint distribution of multi-objective returns. Such joint information is unnecessary for the linear expected utility considered here.

E.1.1 GROUP-CONSISTENT TWIN SELECTION

For a replay transition $(s, a, \mathbf{r}, s', d, \mathbf{w})$, we draw a single next action $a' \sim \pi_\phi(\cdot \mid s', \mathbf{w})$. Both target twin groups evaluate the same a'. We first compute their scalarized expected values, $V'_k = \sum_{i=1}^m w_i \bar{Q}_{i,k}(s', a', \mathbf{w})$, and select $k^- = \arg\min_{k \in \{1,2\}} V'_k$. Crucially, this selection is performed once for the entire objective vector. After selecting k^-, every objective uses the corresponding member of the same twin group. This preserves the semantics of clipped double-Q estimation: $\sum_i w_i \bar{Q}_{i,k^-} = \min_{k \in \{1,2\}} \sum_i w_i \bar{Q}_{i,k}$. In contrast, independently selecting the smaller twin for each objective would produce $\sum_i w_i \min_k \bar{Q}_{i,k} \le \min_k \sum_i w_i \bar{Q}_{i,k}$, and may construct an objective vector that is not predicted by either twin network. We therefore use group-consistent twin selection in both critic targets and actor updates.

E.1.2 OBJECTIVE-WISE DISTRIBUTIONAL BELLMAN UPDATE

Given the selected target group k^-, define the random one-step soft target for objective i as $Y_i = \tilde{r}_i + \gamma(1 - d)\left[Z_{i,k^-}(s', a', \mathbf{w}; \bar{\theta}_{i,k^-}) - \alpha \log \pi_\phi(a' \mid s', \mathbf{w})\right]$. Let $\Pi_{\mathcal{Z}} Y_i$ denote the standard linear projection of this random variable onto the categorical support $\mathcal{Z}$, including clipping at the support boundaries. The resulting target probabilities are denoted by $\hat{p}_{i,\ell}$. Each objective receives its own reward supervision, independently of its current preference weight. The critic loss is

$$\mathcal{L}_Q = -\mathbb{E}_{\mathcal{D}}\left[\sum_{k=1}^{2} \sum_{i=1}^{m} \sum_{\ell=0}^{K-1} \mathrm{sg}(\hat{p}_{i,\ell}) \log p_{i,k,\ell}(s, a, \mathbf{w})\right], \tag{142}$$

where sg denotes stop-gradient. Notice that w_i does not multiply the critic loss. A low-weight objective therefore continues to receive direct supervision from its own observed reward. Preferences influence its value through the future behavior of the common actor, rather than by suppressing its learning signal.

E.2 PREFERENCE-WEIGHTED ACTOR UPDATE

For policy improvement, we sample $a_\phi = f_\phi(s, \mathbf{w}, \epsilon)$ through the reparameterized actor and evaluate it using all online critics. We select the online twin group $k^\pi =$

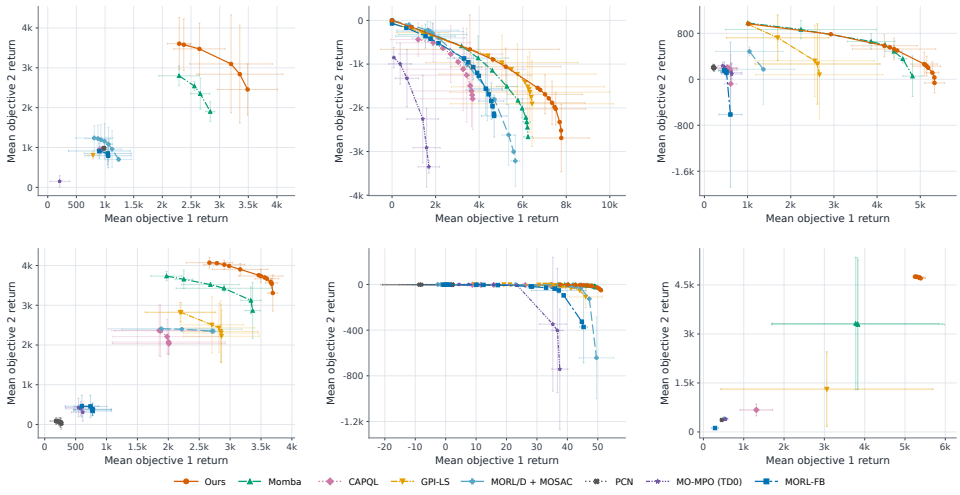


Figure 16: Pareto front comparison of Momba SP (ours) Version over other baselines (original setting).

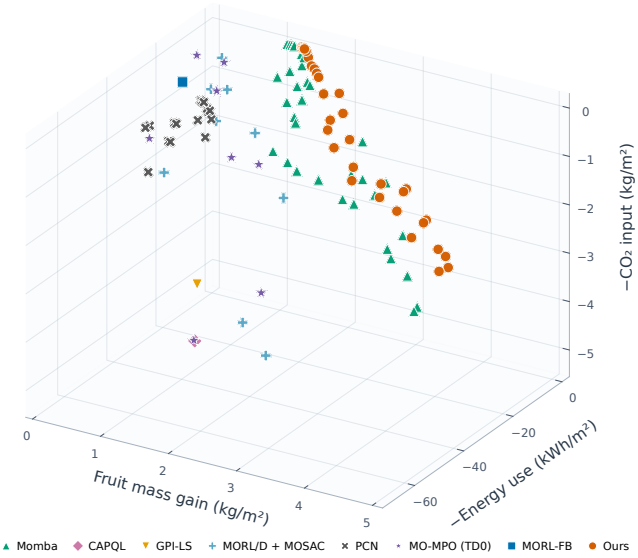


Figure 17: Performance comparison of Momba SP (ours) Version for GreenLight Enviromemt over other baselines (original setting).

$\arg\min_{k\in\{1,2\}} \sum_{i=1}^{m} w_i Q_{i,k}(s, a_\phi, \mathbf{w})$, and optimize

$$\mathcal{L}_\pi = \mathbb{E}\left[\alpha \log \pi_\phi(a_\phi \mid s, \mathbf{w}) - \sum_{i=1}^{m} w_i Q_{i,k^\pi}(s, a_\phi, \mathbf{w})\right]. \tag{143}$$

Away from the measure-zero boundaries at which the selected twin changes, the value component of the policy gradient is $\nabla_\phi \left[\sum_i w_i Q_i(s, a_\phi, \mathbf{w})\right] = \sum_i w_i \left(\nabla_a Q_i(s, a_\phi, \mathbf{w})\right) \nabla_\phi a_\phi$. Thus, although objective evaluation is parameter-decoupled, policy improvement remains a single coupled optimization problem. We do not separately update one actor per objective, nor do we merge independently trained policies.

Table 5: Principal properties of the `PV_rich` environment. Power and energy ratings are per device unless a total is specified.

Property	Specification
Network	IEEE 33-bus, 32 closed branches, 12.66 kV, 10 MVA base
Decision interval / horizon	15 min / 96 steps (noon to noon)
BESS locations	Buses 18, 22, 25, 33
BESS ratings	500 kWh, $\pm$250 kW; total 2 MWh and 1 MW
EV charging ports	440 at buses 12, 18, 25, 30, 33
EV ratings	60 kWh battery, 7 kW charging per port; no V2G
PV installations	One controllable injection at each bus 2–33
Action / observation / reward	476 / 2712 / 5 dimensions
Voltage operating band	$[0.95, 1.05]$ p.u. (soft objective)
AC solver	pandapower backward/forward sweep (BFSW)

F 5-Objective Experiment: Energy Scheduling

We study a finite-horizon, five-objective control problem on the radial IEEE 33-bus distribution feeder. One episode spans 24 hours, from 12:00 pm to 12:00 pm the following day, with $H = 96$ decisions at intervals of $\Delta t = 0.25$ h. At each decision, one agent jointly schedules four community battery energy storage systems (BESSs), 440 individual electric-vehicle (EV) charging ports, and the curtailment of 32 photovoltaic (PV) injections. A balanced steady-state AC power flow is solved at every step. The five objectives are grid operating cost, EV service shortfall, voltage deviations, battery-throughput degradation, and renewable-energy non-utilization. All are expressed as costs to minimize; the environment returns their negatives as a vector reward. Table 5 lists the principal dimensions and parameters.

F.1 Network and Exogenous Scenario Construction

F.1.1 Distribution feeder

The original feeder has 3.715 MW of active and 2.300 Mvar of reactive demand, which provide bus-specific reference loads. We use balanced, positive-sequence AC power flow, solved with pandapower's BFSW algorithm, voltage angles disabled, a maximum of 30 iterations, and an MVA convergence tolerance of 10^{-8}. Reactive consumption follows each bus's original Q/P ratio; BESS, EV chargers, and PV operate at unity power factor. Only active power is controllable. By default, Bus 1 is the slack bus at 1.0 p.u. Each data source is split chronologically by day into approximately 70% training, 15% validation, and 15% test. The Ausgrid splits have $767/164/165$ days; Electric Nation has $380/82/82$; and AEMO has $254/55/55$.

F.1.2 Household demand and solar generation

The Ausgrid Solar Home Electricity Data provide half-hourly general consumption (`GC`), controlled load (`CL`), and gross generation (`GG`) from 2010–2013. These profiles are linearly interpolated onto the 15-minute grid. The source dataset is available at `https://data.gov.au/data/en/dataset/nsw-solar-home-electricty-data`.

F.1.3 EV demand and schedules

We reconstruct charging-session energy from the Electric Nation Green Flux 15 Minute records. Consecutive rows with a common charger and transaction identifier are grouped into sessions, with sessions split across sufficiently large gaps or overlaps. Energy is estimated from average charging current using 230 V nominal single-phase voltage and the record duration. Quality filtering leaves 57,054 sessions from 290 chargers; the accepted sessions have a median reconstructed energy of 5.342 kWh. The original data provide neither observed vehicle state of charge (SOC) nor a declared target SOC. The source is `https://connecteddata.nationalgrid.co.uk/dataset/electric-nation`. The candidate pool for a sampled source day consists of complete sessions whose original arrivals lie between noon and the following noon and whose departures occur by the episode end. For a sampled source energy e_e^{src}, we set the synthetic energy demand to $D_e =$

$1.15e_e^{src}$, the target energy to $E_e^{req} = 0.80(60) = 48$ kWh, and the arrival energy to

$$E_e^{arr} = \max\{0.10(60), E_e^{req} - D_e\}. \tag{144}$$

The resulting target-energy gap is at most 42 kWh. For short daytime visits, source sessions are further conditioned on whether their amplified energy can be delivered at 7 kW, 95% efficiency, within the synthetic dwell duration. The operative demand distribution is therefore conditional on the constructed schedule, not the unconditional distribution of Electric Nation sessions.

There are 40 work/flexible charging ports at each of buses 12, 25, and 30, and 160 residential ports at each of buses 18 and 33. Work/flexible arrivals fall in 08:00–09:00, 12:00–13:30, or 14:30–17:00, with dwell times sampled uniformly from quarter-hour grid values between 1 and 3 h. Residential arrivals fall in 17:00–23:00 with dwell times between 8 and 12 h.

F.1.4 PRICE

We use the 2019 AEMO VIC1 half-hourly regional reference price (RRP, AUD/MWh), expressed in NEM market time (AEST), and form a bounded dynamic retail-price proxy:

$$\pi_t = 0.20 + 0.80\frac{\text{clip}(\text{RRP}_t, -100, 500)}{1000} \quad \text{AUD/kWh}. \tag{145}$$

Thus $\pi_t \in [0.12, 0.60]$ AUD/kWh. The 30-minute series is linearly interpolated to 15 minutes, and exported energy receives 20% of the contemporaneous import price. The RRP files are published by AEMO at `https://www.aemo.com.au/aemo/data/nem/priceanddemand/`.

F.2 PHYSICAL CONTROLS AND TRANSITIONS

The action $a_t = (a_t^B, a_t^{EV}, c_t)$ belongs to

$$\mathcal{A} = [-1, 1]^4 \times [-1, 0]^{440} \times [0, 1]^{32}. \tag{146}$$

Action values are clipped to this box before device constraints are applied. Every EV action controls one charger, although EV active-power injections at the same bus are summed before power flow.

F.2.1 BATTERY AND CHARGER POWER

Positive device power denotes discharging/injection; negative power denotes charging/consumption. A BESS action maps linearly to requested power up to ± 250 kW. For a battery of energy E_t and limits $[E^{min}, E^{max}]$, its actual power is

$$\bar{P}_t^{ch} = \min\left(P^{ch,max}, \frac{E^{max} - E_t}{\eta_{ch}\Delta t}\right), \tag{147}$$

$$\bar{P}_t^{dis} = \min\left(P^{dis,max}, \frac{(E_t - E^{min})\eta_{dis}}{\Delta t}\right), \tag{148}$$

$$P_t = \text{clip}\left(P_t^{req}, -\bar{P}_t^{ch}, \bar{P}_t^{dis}\right), \tag{149}$$

$$E_{t+1} = E_t - \frac{[P_t]^+}{\eta_{dis}}\Delta t + [-P_t]^+\eta_{ch}\Delta t, \tag{150}$$

where $[x]^+ = \max(x, 0)$ and $\eta_{ch} = \eta_{dis} = 0.95$. The four BESSs each have 500 kWh capacity, 10–90% allowable SOC, and independently sampled initial SOC $U(0.15, 0.30)$. An EV has a 60 kWh battery, 10–90% allowable SOC, and a 7 kW charger with 95% charging efficiency. Its action is mapped from $[-1, 0]$ to a requested power between -7 and 0 kW and is projected using the same state-dependent energy limits. Vehicle-to-grid discharge is disabled, and an unplugged EV applies zero power. At step t, the controller acts on vehicles already connected at boundary t. Battery states are updated, followed by the AC solve for interval $[t, t + 1)$. Departures are processed at boundary $t + 1$ before new arrivals are connected and the next observation is constructed.

F.2.2 PV AND AC NETWORK

For $c_{i,t} \in [0, 1]$, the injected PV power is $P_{i,t}^{PV} = (1 - c_{i,t})\bar{P}_{i,t}^{PV}$. At each non-slack bus, the specified real and reactive injections are

$$P_{i,t}^{inj} = P_{i,t}^{PV} + \sum_{b\in\mathcal{B}_i} P_{b,t}^B + \sum_{e\in\mathcal{E}_i} P_{e,t}^{EV} - P_{i,t}^L, \qquad Q_{i,t}^{inj} = -Q_{i,t}^L. \tag{151}$$

The AC solve determines the slack exchange P_t^{grid} (positive for imports) and 33 bus-voltage magnitudes $V_{i,t}$. The 0.95–1.05 p.u. band is penalized in the objective.

F.3 Five-Dimensional Reward

Let $\mathbf{J}_t = (J_t^{eco}, J_t^{EV}, J_t^V, J_t^{deg}, J_t^{ren})$ and $\mathbf{r}_t = -\mathbf{J}_t$. Raw units and scales are deliberately retained; the environment does not choose a preference weight or rescale the reward dimensions.

- *Grid operating cost.* With $P_{t,\mathrm{kW}}^{grid}$ in kW, the interval energy bill is
$$J_t^{grid} = \left(\pi_t [P_{t,\mathrm{kW}}^{grid}]^+ - 0.2\pi_t [-P_{t,\mathrm{kW}}^{grid}]^+\right) \Delta t. \tag{152}$$
- *EV service.* Let $\mathcal{D}_{t+1}$ be the vehicles departing after interval t, $z_{e,t+1}$ the post-boundary connection indicator, $\mathcal{C}_{t+1}$ the connected set after departures and arrivals, and $n_{e,t}^{rem}$ the remaining feasible charging intervals after the action. With $N_{EV} = 440$, define
$$D_t = \frac{1}{N_{EV}} \sum_{e \in \mathcal{D}_{t+1}} \frac{[E_e^{req} - E_{e,t+1}]^+}{E_e^{req}}, \tag{153}$$
$$U_t = \frac{1}{N_{EV}} \sum_e z_{e,t+1} \left[\frac{[E_e^{req} - E_{e,t+1}]^+}{7 n_{e,t}^{rem} \Delta t (0.95) + 10^{-9}} - 1 \right]^+, \tag{154}$$
$$R_t = \begin{cases} \dfrac{1}{|\mathcal{C}_{t+1}|} \displaystyle\sum_{e \in \mathcal{C}_{t+1}} \frac{[E_e^{req} - E_{e,t+1}]^+}{E_e^{req}}, & |\mathcal{C}_{t+1}| > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{155}$$
$$J_t^{EV} = D_t + 0.05 U_t + 0.02 R_t. \tag{156}$$
The first term measures departure shortfall relative to the *entire* 440-port fleet, the second signals an already unrecoverable charging deficit under maximum future charging, and the third provides a dense online readiness signal. A separate diagnostic counts a departure as successful if its final energy reaches 95% of its 48 kWh target; this threshold does not change J_t^{EV}.
- *Voltage quality.* For all 33 buses, let $\ell_{i,t} = [0.95 - V_{i,t}]^+ + [V_{i,t} - 1.05]^+$. The step cost is
$$J_t^V = 0.075 \,\mathrm{mean}_i \, |V_{i,t} - 1| + 0.175 \,\mathrm{mean}_i \, \ell_{i,t} + 0.75 \max_i \ell_{i,t}. \tag{157}$$
The first term is nonzero even inside the operating band; the last term emphasizes the worst bus. Voltage violations do not themselves end the episode. On AC solver failure, the implementation instead assigns $J_t^V = 10$ and terminates the episode early.
- *Battery-throughput degradation proxy.* Let $W_t^B = \Delta t \sum_b |P_{b,t}^B|$ and $W_t^{EV} = \Delta t \sum_e |P_{e,t}^{EV}|$ be AC-side energy throughput in kWh. With 2,000 kWh of installed BESS capacity and 26,400 kWh of installed EV capacity,
$$J_t^{deg} = \frac{1}{2} \frac{W_t^B}{2(2000)} + \frac{1}{2} \frac{W_t^{EV}}{2(26400)}. \tag{158}$$
Thus a full equivalent cycle of either aggregate battery fleet contributes 0.5 to this objective. This measure is a throughput proxy, not a battery-chemistry or capacity-fade model.
- *Renewable-energy non-utilization.* Define episode-available PV energy $E^{PV,av} = \Delta t \sum_{t,i} \bar{P}_{i,t}^{PV}$, curtailed power $P_t^{curt} = \sum_i (\bar{P}_{i,t}^{PV} - P_{i,t}^{PV})$, and a feeder-level attribution of local absorption and PV export:
$$P_t^{abs} = \sum_i P_{i,t}^L + \sum_b [-P_{b,t}^B]^+ + \sum_e [-P_{e,t}^{EV}]^+, \tag{159}$$
$$P_t^{PV,export} = \left[\sum_i P_{i,t}^{PV} - P_t^{abs} \right]^+, \tag{160}$$
$$J_t^{ren} = \frac{(P_t^{curt} + P_t^{PV,export}) \Delta t}{E^{PV,av}}. \tag{161}$$
Episode PV utilization is $1 - \sum_t J_t^{ren}$. This definition treats both curtailment and PV attributed to feeder export as energy not used locally.

F.4 Agent Observation and Information Structure

The agent receives a 2712-dimensional observation comprising the following information:

- *Time and exogenous conditions:* sine and cosine encodings of the time of day, together with current prices, aggregate load, and available PV and four 15-minute forecasts of each quantity (17 features).
- *Storage and EV states:* the SOCs of the four BESSs and, for each of the 440 EV ports, stored-energy fraction, connection status, remaining time to departure, target-energy fraction, and time to arrival where applicable (2204 features).
- *Network feedback:* exogenous net load and the most recently measured voltage at each of the six buses hosting controllable resources, plus the most recent feeder-wide minimum and maximum voltages and grid import (15 features).
- *Control history:* the 476 setpoints selected at the preceding decision (476 features).

Current exogenous values are known at the decision time. For price, load, and available PV, each future forecast is formed as

$$\hat{x}_{t+k|t} = x_{t+k} \max(0, \epsilon_{t,k}), \qquad \epsilon_{t,k} \sim \mathcal{N}(1, 0.1^2), \quad k = 1, \ldots, 4. \tag{162}$$

Forecast errors are sampled independently. The observation provides only one hour of look-ahead and selected feeder measurements; the complete future profiles and all current bus voltages are not observed. Thus, the agent faces a partially observable decision problem even though the full episode scenario is fixed once sampled.

F.5 Experimental Results

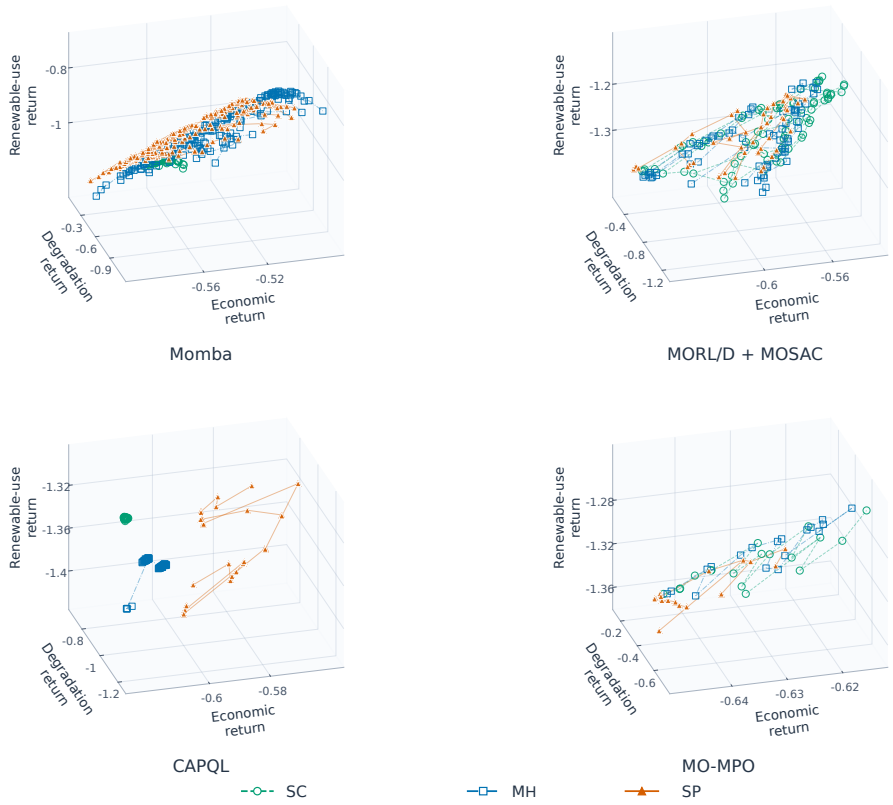


Figure 18: Pareto fronts of 33-bus environment with different algorithm: Economic-Degradation-Renewable.

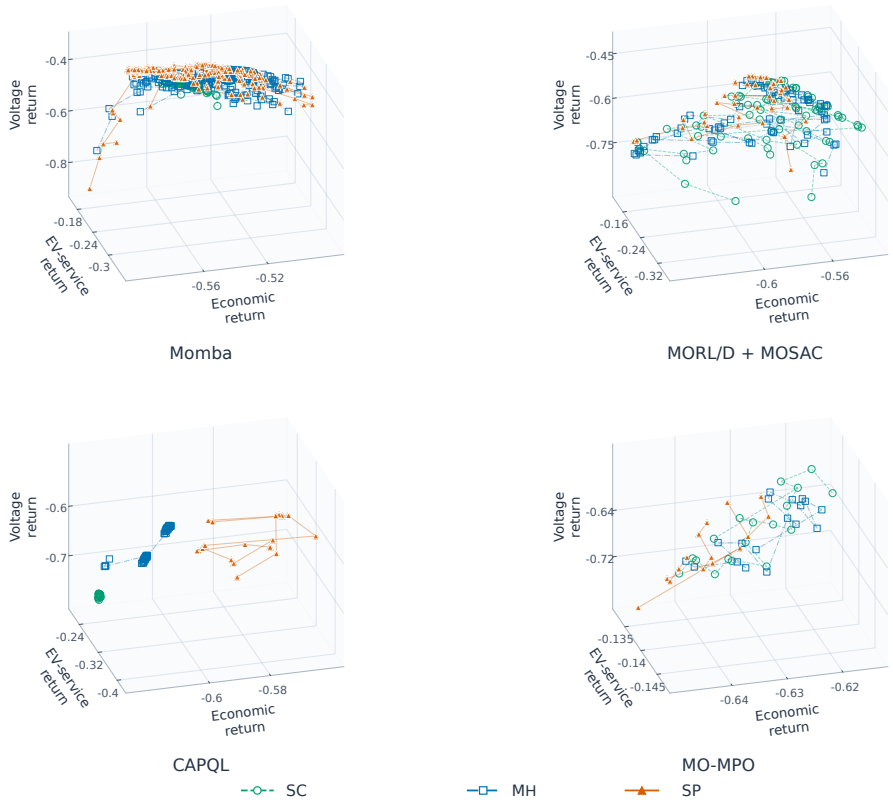


Figure 19: Pareto fronts of 33-bus environment with different algorithm: Economic-EV-Voltage.

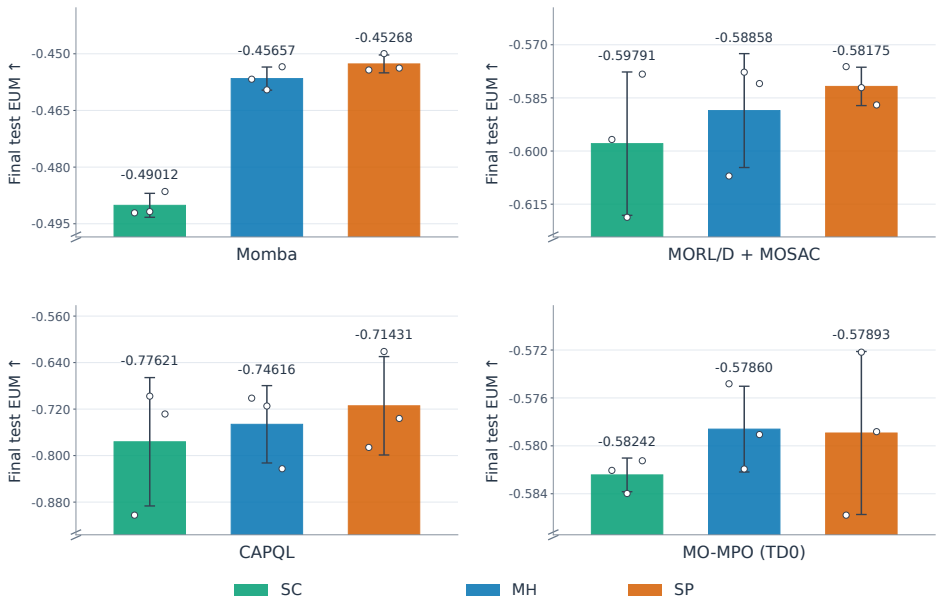


Figure 20: Performance of 33-bus environment with Five Objective.