CONTROLLABLE SKILL DISCOVERY WITH A SKILL-LINEAR CRITIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised skill discovery (USD) learns diverse behaviors without task rewards, but those behaviors are not necessarily controllable. Existing methods can obtain diversity through small state changes or varied locomotion directions, yet varying their skills can move state properties together or switch them among a few extremes. We ask whether the skill vector can provide a controllable interface for bounded state descriptors such as joint angles. We use *controllable* to mean that one descriptor can be varied gradually (graded control), other descriptors can be held approximately fixed (isolated control), and multiple descriptor targets can be achieved together (composable control). We propose the Skill-Linear Critic (SLC), which retains a mutual-information-based intrinsic reward but restricts skill dependence to a linear combination of skill-independent basis critics. This turns the skill vector from an arbitrary conditioning input into a coefficient vector for a skill-independent family of critics. We add a pairwise orthogonality penalty between the basis embeddings to mitigate observed critic-training instability. In static-pose evaluations of four simulated robots, SLC yields graded, isolated, and composable joint-angle behavior on HalfCheetah, Ant, and Humanoid. It is also the only evaluated learned pipeline with a reliable affine fit in a majority of evaluations on every platform, although this all-platform conclusion is threshold-sensitive for Franka.

1 INTRODUCTION

Reinforcement learning (RL) often depends on carefully designed task-specific rewards. Unsupervised skill discovery (USD) reduces this dependence by learning diverse behaviors without task-specific rewards (Eysenbach et al., 2019; Sharma et al., 2020b; Park et al., 2024; Zheng et al., 2025). These methods condition a policy on a latent *skill* vector. Prior work shows that the learned skills are diverse and useful for downstream tasks such as goal reaching. Yet it remains unclear whether varying a continuous skill can smoothly control one aspect of behavior without changing others.

Existing USD objectives are designed primarily to distinguish skills or induce large state changes, not to make the skill vector a direct control interface. Methods based on mutual information (MI) emphasize distinguishability or predictability (Sharma et al., 2020b), while representation-distance methods such as METRA and CSF encourage large skill-aligned changes (Park et al., 2024; Zheng et al., 2025); neither objective requires smooth intermediate values or fixed non-target descriptors. FSD and SUSD introduce factor-specific structure, but their evaluations do not establish smooth single-variable adjustment and simultaneous target realization under one continuous skill held fixed within a rollout (Colaço Carr et al., 2025; Hosseini & Soleymani Baghshah, 2026).

Beyond distinguishability, we seek controllable skill discovery over selected bounded state descriptors. Intuitively, imagine controlling a multi-jointed robot with a bank of knobs. Control is graded if turning one knob changes one joint angle smoothly, isolated if the other joints remain approximately fixed, and composable if several knob settings are realized together. Throughout this paper, *controllable* refers to these three properties jointly. To encourage these properties, we propose the Skill-Linear Critic (SLC), which retains the DADS-style objective (Sharma et al., 2020b) and training pipeline but represents each critic head as a weighted sum of skill-independent basis critics, with the skill components supplying only their coefficients (Figure 1). Thus, the skill no longer enters the critic as an arbitrary conditioning input; it becomes a coefficient vector for a skill-independent

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.

family of action-value components. To discourage the very large, nearly parallel basis embeddings observed in some unstable runs, we add a pairwise orthogonality penalty between the embeddings.

To determine whether our method yields controllable behavior, we hold one skill fixed within each rollout and evaluate bounded joint angles in static-pose variants of four simulated robots. Previous methods can produce distinguishable behaviors, yet their joint responses may remain coupled, switch abruptly between poses, or fail to realize requested combinations. With a platform-selected penalty weight, SLC provides evidence of approximately controllable behavior on HalfCheetah, Ant, and Humanoid; on Franka, control remains partial with step-like responses (Table 1; Figure 13).

Our main contributions are as follows: (1) We introduce SLC, which constrains the critic's skill dependence to a linear combination of basis critics under an MI-based intrinsic reward. (2) We identify distinct failures of the unregularized critic and show that pairwise regularization can stabilize training, while ablations indicate that orthogonality alone does not explain the behavioral gains. (3) We evaluate controllability through graded, isolated, and composable behavior against a training-free linear reference. Our claims concern bounded joint angles in static poses; Appendices C.1 and C.3 report scope tests on geometric descriptors and unbounded locomotion axes.

2 Related Work

2.1 Unsupervised Skill-Discovery Objectives

USD without factor assignments typically optimizes mutual information or representation-space displacement. VIC (Gregor et al., 2017), DIAYN (Eysenbach et al., 2019), and VALOR (Achiam et al., 2018) maximize MI lower bounds over states, transitions, or trajectories. DADS (Sharma et al., 2020b) learns skill-conditioned dynamics for skill-sequence planning, and Off-DADS (Sharma et al., 2020a) extends it off-policy. LSD (Park et al., 2022), CSD (Park et al., 2023), and METRA (Park et al., 2024) instead maximize displacement in a learned representation. CSF (Zheng et al., 2025) relates METRA's objective to a contrastive MI approximation and incorporates successor-feature structure. Unlike METRA and CSF, SLC constrains the critic's dependence on skill: it retains a DADS-style intrinsic reward but allows skill to enter only as coefficients on skill-independent bases.

2.2 Factor-Informed and Disentangled Skills

Factor-informed approaches instead use state-factor or target-variable structure to organize skills (Wang et al., 2024; Hu et al., 2025). DUSDi (Hu et al., 2024) combines factor-isolated components in one policy and evaluates discrete components and hierarchical reuse, although each factor-specific Q_i depends on the whole skill. D3 (Cathomen et al., 2025) studies factor composition, variable-norm speed, interpolation, and switching. FSD (Colaço Carr et al., 2025) targets specified variables, but its tabular experiments use finite skills targeting one variable each, without graded values or simultaneous targets. SUSD (Hosseini & Soleymani Baghshah, 2026) uses factor-specific blocks, embeddings, and distance rewards to evaluate decoding, goal reaching, and compositional downstream tasks. Unlike approaches with prescribed skill-to-factor assignments, SLC uses selected descriptors without such assignments and seeks controllable behavior through the critic structure.

2.3 Linear Value Structure and Successor Features

SLC is architecturally connected to linear value structure and successor features because it represents a skill-conditioned critic as a weighted combination of skill-independent bases. The successor representation (Dayan, 1993) encodes discounted state occupancy, and successor features (SF) make value linear in a reward vector ω for a fixed policy under a linear reward family (Barreto et al., 2017). VISR (Hansen et al., 2020) and APS (Liu & Abbeel, 2021) use SF, while FB (Touati & Ollivier, 2021) uses an analogous successor-measure factorization, for pretraining without external reward. UVFA (Schaul et al., 2015) conditions values directly on goals or tasks, and USFA (Borsa et al., 2019) combines this conditioning with SF. SLC instead makes value linear in the behavior-conditioning variable α through bases that receive no skill input. Its ψ_k is not the discounted expectation of a cumulant, so SF transfer guarantees do not apply. The modular SF structure of MSFA (Carvalho et al., 2023) and SF diversity of Zahavy et al. (2021) are structural precedents; CSF (Zheng et al., 2025) is our closest SF-family baseline.

2.4 Skill Composition and Behavioral Control

Skill composition. "Composition" includes temporal chaining in SNN (Florensa et al., 2017), SPiRL (Pertsch et al., 2021), and OPAL (Ajay et al., 2021); energy or primitive products (Du et al., 2020; Urain et al., 2021; Peng et al., 2019; Jansonnie et al., 2024); ModSkill's modular assembly (Huang et al., 2025); and value or task composition (Todorov, 2009; Haarnoja et al., 2018a; Nangue Tasse et al., 2020; Barreto et al., 2019). We study a different notion: whether one fixed continuous skill can control several behavioral descriptors together. SLC's additive critic gives each head a weighted-product interpretation related algebraically to Haarnoja et al. (2018a). Appendix A.1 derives this single-head interpretation and states its limits; the bases are jointly learned rather than independently trained primitive policies.

Behavioral control and evaluation. Goal-conditioned methods such as UVFA (Schaul et al., 2015) and HER (Andrychowicz et al., 2017) train policies on explicit targets, whereas SLC never conditions its policy on descriptor targets. ICF (Thomas et al., 2017) learns continuous factor and policy embeddings, while Constrained Skill Discovery (Atanassov et al., 2024) updates the skill every step to track position and yaw. Disentanglement metrics—β-VAE score (Higgins et al., 2017), FactorVAE score (Kim & Mnih, 2018), mutual information gap (MIG) (Chen et al., 2018), disentanglement, completeness, and informativeness (DCI) (Eastwood & Williams, 2018), and interventional robustness score (IRS) (Suter et al., 2019)—measure latent–factor correspondence rather than realized behavior. We ask whether fixed continuous skills produce controllable behavior in this sense.

3 Method

We propose the Skill-Linear Critic (SLC), which turns the skill vector α from an arbitrary conditioning input into the coefficient vector of a fixed linear family of critics. In our implementation, SLC retains the DADS-style intrinsic reward and SAC training pipeline but represents each critic head as a skill-weighted sum of skill-independent basis critics. It adds pairwise orthogonality regularization to the basis embeddings, with a platform-selected weight that may be zero.

3.1 Preliminaries and Problem Setting

Task-reward-free MDP and skill-conditioned policies. We study USD in a task-reward-free MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability $P(s' \mid s, a)$, and discount factor $\gamma \in [0, 1)$. Prior USD methods typically denote the skill vector by z (Eysenbach et al., 2019; Sharma et al., 2020b; Park et al., 2024). We instead use α and write $\pi(a \mid s, \alpha)$, because each α_k also serves as the coefficient of a skill-independent basis critic. At the beginning of each episode, we sample $\alpha \sim \mathrm{Unif}(\mathbb{S}^{B-1})$ once from the unit sphere and hold it fixed throughout the episode (Zheng et al., 2025), with B denoting the skill dimension and, in SLC, the number of basis critics.

Descriptors. Each descriptor is a bounded state function $d_i : \mathcal{S} \to \mathbb{R}$ whose value we aim to control, and we collect the K descriptors as $d(s) = (d_1(s), \ldots, d_K(s)) \in \mathbb{R}^K$. Descriptors can be quantities directly included in the robot state, such as joint angles, or quantities computed from it, such as torso-relative distances and heights (Appendix C.1). We use the descriptors both to compute the intrinsic reward and to evaluate behavioral control. Thus, the state quantities of interest are selected in advance, but SLC is not given an assignment between skill components and descriptors. The relation between skill-space directions and descriptors emerges during training, without requiring one-to-one alignment between the coordinates of α and the descriptors. Section 4 explains how we identify a descriptor-specific direction that varies one descriptor while holding the others fixed.

Skill-discovery objective and intrinsic reward. A standard skill-discovery objective is mutual information between the skill and states. Following DADS, the ideal motivation is the MI conditioned on the current descriptor (Sharma et al., 2020b). We write the current and next descriptors as $d = d(s)$ and $d' = d(s')$. A variational lower bound (Barber & Agakov, 2004; Sharma et al., 2020b) with the ideal skill-conditioned descriptor transition model $q_{\mathrm{desc}}(d' \mid d, \alpha)$ gives $I(\alpha; d' \mid d) \geq \mathbb{E}[\log q_{\mathrm{desc}}(d' \mid d, \alpha) - \log p(d' \mid d)]$ and the ideal descriptor-level intrinsic reward

$$r(s, \alpha, s') \;=\; \log q_{\mathrm{desc}}(d' \mid d, \alpha) \;-\; \log p(d' \mid d), \tag{1}$$

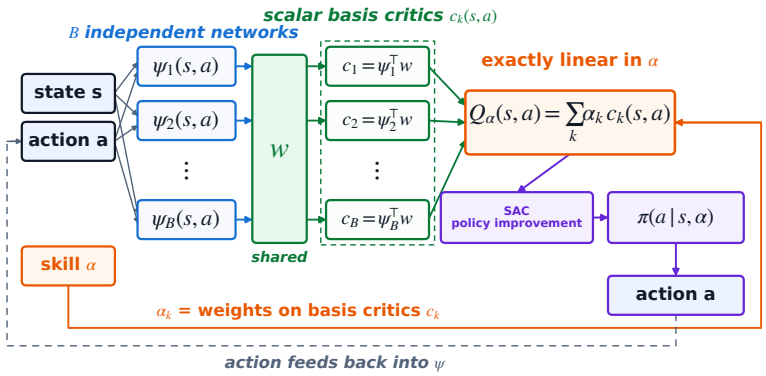

Figure 1: SLC architecture. A projection w shared by the B skill-independent embeddings $\psi_k(s,a)$ produces the scalar critics $c_k(s,a)$. The skill components α_k provide their weights.

where $p(d' \mid d)$ is the conditional marginal induced by the skill prior and rollouts. SLC retains this DADS-style skill-discovery objective but changes the critic structure and adds pairwise orthogonality regularization, as described in the following subsections.

3.2 Skill-Linear Critic

As shown in Figure 1, SLC allows α to enter the critic only as coefficients on skill-independent components. This differs from both standard skill-conditioned critics, which take the skill as a neural-network input, and CSF's $Q(s,a,\alpha) = \psi(s,a,\alpha)^\top \alpha$ (Zheng et al., 2025), whose ψ also receives α and therefore makes Q generally nonlinear in α. SLC shares only the inner-product form of successor features (Barreto et al., 2017), not their semantics or transfer guarantees (Section 2.3).

SLC learns B skill-independent embeddings $\psi_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ and a projection vector $w \in \mathbb{R}^n$ shared across them. Defining the scalar basis critic $c_k(s,a) \triangleq \langle \psi_k(s,a), w \rangle$ gives

$$Q_\alpha(s,a) \;=\; \Big\langle \sum_{k=1}^{B} \alpha_k \, \psi_k(s,a),\; w \Big\rangle \;=\; \sum_{k=1}^{B} \alpha_k \, c_k(s,a). \tag{2}$$

The first equality combines the embeddings using α_k and then takes the inner product with w. Equivalently, one may project each embedding onto w and then linearly combine the resulting scalar basis critics, as in the second equality. Thus Q_α is exactly linear in the skill components α_k, and the realizable critics lie in the span of the c_k, which need not be linearly independent. Because the intrinsic reward and skill-conditioned policy can depend nonlinearly on α, this linear form is an architectural constraint rather than an exact factorization of the true soft Q-function. The intermediate embedding space also provides the level at which we apply the optional regularizer introduced next.

3.3 Orthogonality Regularization of the Basis

The critic is trained with the standard temporal-difference (TD) loss of SAC (Haarnoja et al., 2018b). Let y_{SAC} denote the corresponding stop-gradient target. The loss is

$$\mathcal{L}_{\mathrm{TD}} \;=\; \tfrac{1}{2}\, \mathbb{E}_{\mathcal{D}}\Big[\big(Q_\alpha(s,a) - y_{\mathrm{SAC}}\big)^2\Big], \tag{3}$$

where $\mathcal{D}$ is the replay buffer. Appendix A.1 gives the full definition of y_{SAC}.

This factorization has two structural limitations. First, $\partial Q_\alpha / \partial \psi_k = \alpha_k w$, so a sample's direct TD gradient on ψ_k is $(Q_\alpha - y_{\mathrm{SAC}})\alpha_k w$ and is parallel to the current w. Orthogonal components change only indirectly through network parameters or rotations of w. Second, TD constrains only the inner product and is invariant under $\psi_k \to \eta\psi_k$, $w \to w/\eta$, leaving their joint scale unidentified.

Motivated by these limitations and by instability in some unregularized runs, we add a pairwise orthogonality penalty between the minibatch-centered basis embeddings. Let $\mathcal{B}$ denote the current

minibatch and $\tilde{\psi}_k$ the embedding after subtracting its minibatch mean:

$$\mathcal{L}_\perp \;=\; \sum_{i<j} \mathbb{E}_{(s,a)\sim\mathcal{B}}\Big[\big\langle \tilde{\psi}_i(s,a),\, \tilde{\psi}_j(s,a)\big\rangle^2\Big], \qquad \tilde{\psi}_k \;=\; \psi_k - \mathbb{E}_{\mathcal{B}}[\psi_k]\,, \tag{4}$$

and the critic loss is the sum of the two terms:

$$\mathcal{L}_Q \;=\; \mathcal{L}_{\mathrm{TD}} \;+\; \lambda_\perp\, \mathcal{L}_\perp. \tag{5}$$

Under the rescaling, centering gives $\tilde{\psi}_k \to \eta\tilde{\psi}_k$. Each inner product scales by η^2 and its square by η^4, so $\mathcal{L}_\perp \to \eta^4\mathcal{L}_\perp$. This breaks the scale invariance of the TD loss but does not identify or bound the joint scale: shrinking the embeddings while enlarging w reduces $\mathcal{L}_\perp$ without changing Q_α. The effective strength of $\lambda_\perp$ depends on the embedding scale and is not directly comparable across platforms. Section 5.4 evaluates its empirical stabilizing effect.

3.4 Training and Implementation

SLC follows the DADS intrinsic-reward estimator (Sharma et al., 2020b) and SAC training pipeline (Haarnoja et al., 2018b). As in standard SAC, DADS and SLC each use two separately parameterized critic heads. Each SLC head instantiates Eq. (2) with its own B width-1024 basis MLPs and projection vector w; the TD loss and optional regularizer are applied to each head. Let $\theta^{(m)}$ collect all parameters of head m, including its basis MLPs and $w^{(m)}$, and let $\bar{\theta}^{(m)}$ denote its target parameters. In Algorithm 1, q_ξ is the skill-conditioned dynamics model from DADS. The implemented reward r_{impl} contrasts the observed transition's log density under the rollout skill against N_{alt} Gaussian alternative skills; δ_t is the episode-end indicator. Each SLC head has about B times the parameters of a baseline head; Appendix C tests a lower-parameter variant that generates all B basis embeddings from one shared MLP. Appendix A.1 gives the exact reward formula and implementation details.

Algorithm 1 Training procedure for the Skill-Linear Critic (SLC).

Input: $\lambda_\perp$, dimensions B, n, updates N_{opt}, alternatives N_{alt}, target rate $\tau_{\mathrm{tgt}} \in (0, 1]$
Initialize π, the DADS dynamics model q_ξ, and twin SLC critic parameters $\{\theta^{(m)}\}_{m=1}^2$
Set $\bar{\theta}^{(m)} \leftarrow \theta^{(m)}$ for $m \in \{1, 2\}$
for each epoch **do**
  Draw skill vectors $\{\alpha^{(j)}\}$ i.i.d. from $\mathrm{Unif}(\mathbb{S}^{B-1})$
  For each j, roll out $\pi(\cdot \mid s, \alpha^{(j)})$; store $(s, a, \alpha^{(j)}, s', \delta_t)$ in $\mathcal{D}$
  for $i = 1$ to N_{opt} **do**
    Sample a minibatch; update q_ξ; compute r_{impl} with N_{alt} Gaussian alternatives
    For each m, form $Q_\alpha^{(m)} = \sum_{k=1}^B \alpha_k \langle \psi_k^{(m)}(s, a), w^{(m)} \rangle$ by Eq. (2)
    Update each $\theta^{(m)}$ with $\mathcal{L}_Q$ in Eq. (5), using r_{impl} in y_{SAC}; update π and temperature
    $\bar{\theta}^{(m)} \leftarrow \tau_{\mathrm{tgt}}\theta^{(m)} + (1 - \tau_{\mathrm{tgt}})\bar{\theta}^{(m)}, m \in \{1, 2\}$
  end for
end for

4 Evaluation Protocol

We evaluate controllability as graded, isolated, and composable behavior. Since DCI, MIG, and IRS (Eastwood & Williams, 2018; Chen et al., 2018; Suter et al., 2019) do not jointly test smooth single-descriptor changes and simultaneous targets, Figure 2 uses three stages: (1) an affine-fit gate checks whether a global skill–descriptor map is usable; (2) single-descriptor sweeps measure isolated control (iso) and graded continuity (gap; larger is less smooth); and (3) joint-target inversion tests composition. Error and success measure realized control, whereas reach measures only feasibility predicted by the fitted map. Table 1 rounds gap to the nearest integer. Appendices B.1 and B.3 examine decoder and gate robustness; Appendix B.7 reports standard-metric-inspired diagnostics, and Appendix B gives the full protocol.

To make this protocol concrete, consider a supine HalfCheetah with one ideal knob per measured joint angle. Raw skill coordinates are not these knobs; the affine calibration below constructs candidate knob paths. We hold α fixed within each rollout and define its late-window response as $\hat{D}(\alpha) = T_{\mathrm{eval}}^{-1} \sum_{t=T-T_{\mathrm{eval}}+1}^{T} d(s_t)$. Behavioral metrics use new rollout responses, but this summary does not test settling or within-window variation.

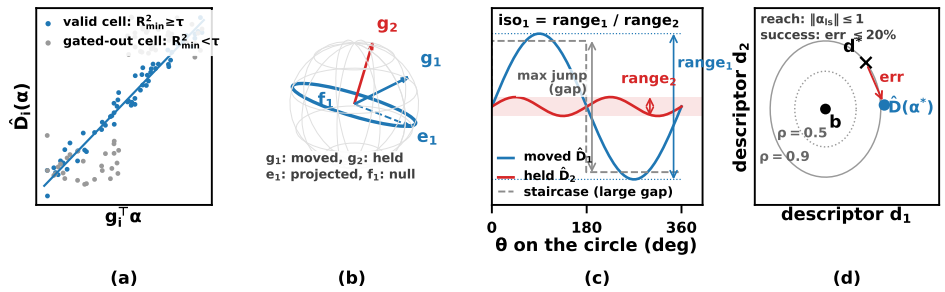

Figure 2: Evaluation overview. (a) Apply the affine-fit gate per (seed, reset) cell. (b) Construct a great-circle intervention for one descriptor. (c) Measure isolation and the largest adjacent jump (gap). (d) Invert the fitted map for joint targets and measure reach, error, and success.

Affine fit and gate. For sphere-sampled skills, we fit (Figure 2(a))

$$\hat{D}(\alpha) \;\approx\; G\,\alpha + b, \qquad G \in \mathbb{R}^{K\times B},\; b \in \mathbb{R}^{K}. \tag{6}$$

Ordinary least squares gives G, b; row g_i is descriptor i's fitted skill-space direction. Each cell combines one independently trained policy (training seed) and one evaluation initialization (reset). We retain it only if every descriptor satisfies $R^2_{\min} = \min_i R^2_i \geq 0.75$. This gate only licenses tests that depend on G: passing does not establish control, and failure does not rule out another decoder. Appendices B.2, B.3, and B.6 give the remaining validity, robustness, and aggregation details.

Isolation and continuity. On $\mathbb{S}^1$, $\alpha(\theta) = (\cos\theta, \sin\theta)$ turns θ into a dial; this only parameterizes a circular sweep. The fit R^2 is computed on random sphere probes, but restricting an affine map to such a circle predicts a sinusoid in θ. Thus a full-revolution plot is expected to look U-shaped rather than linear (Appendix B.4). In higher dimensions, Figure 2(b) embeds the same dial in a two-dimensional great-circle plane: under the fitted map, e_i changes only descriptor i, whereas $f_i \in \ker G$ changes none. Concretely, e_i is the normalized projection of g_i orthogonal to the held-descriptor gradients g_j, $j \neq i$, and f_i is a unit vector in $\ker G$ orthogonal to e_i. The candidate circle and realized isolation score are

$$\alpha^{(i)}(\theta) = \cos\theta\, e_i + \sin\theta\, f_i, \qquad \mathrm{iso}_i = \frac{\mathrm{range}_\theta\ \hat{D}_i\big(\alpha^{(i)}(\theta)\big)}{\max_{j\neq i} \mathrm{range}_\theta\ \hat{D}_j\big(\alpha^{(i)}(\theta)\big)}, \qquad \mathrm{iso} = \min_i \mathrm{iso}_i. \tag{7}$$

All 210 main-comparison cells satisfy the required rank and nondegeneracy checks. We evaluate each circle in separate fixed-skill rollouts. Iso is the worst-axis ratio of intended to largest unintended range. Gap (Figure 2(c)) measures continuity by normalizing each descriptor's largest cyclic adjacent jump by that of an equally sampled sine with the same range, then averaging across descriptors; 1 is the sinusoidal reference. We set no binary iso or gap threshold. Appendix B.4 gives the checks, construction, calibration, and aggregation details.

Composition. Testing knobs separately does not show that they work together. Figure 2(d) therefore treats the affine map as a calibration rule: collect simultaneous knob settings into one target and invert it to one fixed skill,

$$d^* = (d^*_1, \ldots, d^*_K)^\top, \qquad \alpha_{\mathrm{ls}} = G^+(d^* - b), \tag{8}$$

where d^*_i is knob i's request and G^+ is the Moore–Penrose pseudoinverse. This is map inversion, not skill addition. Using a common across-method half-range h_i for each descriptor, we sample targets at descriptor-space radius ρ around b. If $\|\alpha_{\mathrm{ls}}\| \leq 1$, the first-SVD $\ker G$ direction completes it to the unit sphere without changing the fitted descriptors; otherwise it is radially clipped. Requests are evaluated in new rollouts at the same reset realization used for the fit (reset 0); transfer across resets is not tested. Reach is the fraction of outer targets ($\rho = 0.9$) whose inverse satisfies $\|\alpha_{\mathrm{ls}}\| \leq 1$; it measures feasibility predicted by the affine map, not realized success. In contrast, $\mathrm{err}_{\mathrm{in}}$ is realized axis-mean error over $\rho \leq 0.5$, and success is the fraction of outer requests with normalized axis-mean error at most 20%; the stricter every-axis criterion requires each axis separately to meet this tolerance. Appendix B.5 gives the targets, sphere adjustment, error, null-fill convention, and aggregation.

Table 1: Main results (metrics: Section 4). Valid counts (training seed, reset) cells with $R^2_{\min} \geq 0.75$; iso/gap use valid resets, whereas err_{in}/succ use reset 0 and the first-SVD null fill. Learned methods aggregate three training seeds and the linear torque reference three random matrices; $\pm$ denotes the corresponding standard deviation. The positive-penalty SLC row uses $\lambda_\perp = 1$ on HalfCheetah, Ant, and Humanoid and 100 on Franka. The Obj. column identifies the discovery objective: mutual information (MI), temporal distance (T-Dist.), InfoNCE (NCE), or factor-informed (F). Gray marks majority-invalid learned rows and the unranked reference; composition entries in gray rows are diagnostics, not valid control results. Bold marks the best when at least two learned rows are majority-valid; † marks selected $\lambda_\perp$.

		HalfCheetah					Ant					Humanoid					Franka				
Method	Obj.	valid	iso ↑	gap ↓	err_{in} / succ ↓ (%) / ↑ (%)		valid	iso ↑	gap ↓	err_{in} / succ ↓ (%) / ↑ (%)		valid	iso ↑	gap ↓	err_{in} / succ ↓ (%) / ↑ (%)		valid	iso ↑	gap ↓	err_{in} / succ ↓ (%) / ↑ (%)	
Linear torque (ref.)	–	3/3	381±58	1	0.1±0.0 / 100		3/3	144±28	1	0.1±0.0 / 100		3/3	7.98±2.23	1	3.3±0.3 / 100		3/3	3.71±1.44	1	7.5±2.7 / 97	
DADS	MI	0/9	–	–	11.4±4.1 / 58		0/9	–	–	13.7±1.5 / 53		0/9	–	–	14.5±2.5 / 53		0/3	–	–	13.8±0.4 / 14	
METRA	T-Dist.	0/9	–	–	21.5±6.3 / 64		0/9	–	–	34.2±6.8 / 42		0/9	–	–	24.3±5.7 / 47		0/3	–	–	25.5±6.0 / 53	
CSF	NCE	9/9	0.95±0.14	8	24.1±2.5 / 64		9/9	1.59±0.30	11	23.2±3.0 / 89		3/9	1.89	12	22.1±4.7 / 75		1/3	0.57	17	29.3±14.3 / 44	
FSD (adapt.)	F	9/9	**5.08±0.39**	5	6.0±0.9 / **100**		9/9	**6.93±4.50**	14	13.0±2.0 / 94		0/9	–	–	10.9±1.6 / 86		0/3	–	–	39.3±2.0 / 28	
SUSD (port)	F	0/9	–	–	8.0±2.4 / 78		0/9	–	–	11.3±1.1 / 3		0/9	–	–	11.9±5.7 / 33		1/3	1.00	19	15.3±4.4 / 64	
SLC, $\lambda_\perp$=0	MI	8/9†	4.09±0.76	**4**	**3.5±0.8 / 100**		0/9	–	–	26.8±5.9 / 44		0/9	–	–	29.4±3.8 / 75		0/3	–	–	15.7±7.3 / 58	
SLC, $\lambda_\perp$>0	MI	9/9	2.81±1.11	9	4.8±0.7 / **100**		9/9†	2.14±0.25	**7**	**6.5±2.4 / 100**		9/9†	2.70±0.79	6	7.4±5.8 / 97		3/3†	1.58±0.32	25	16.3±7.5 / 75	

5 Experimental Results

Our experiments ask four questions. Does SLC provide controllable behavior under the three stated criteria? How does control change as the number of descriptors K grows at fixed skill dimension B, and does increasing B at fixed K help? What roles do the zero-intercept form and pairwise orthogonality penalty play? Where does the method succeed or fail?

5.1 Experimental Setup

We evaluate bounded joint angles in static-pose variants of HalfCheetah, Ant, Humanoid, and Franka that suppress locomotion and contact: HalfCheetah is supine, Ant has stiffened ankles, Humanoid has a fixed torso with free shoulders, and Franka has frozen non-robot objects with object contacts disabled. Main-comparison (K, B) values are $(3, 6), (2, 4), (2, 8), (2, 3)$ in platform order; Franka FSD/SUSD use $B = 4$. Appendices A.2 and A.3 give the environment and stiffness details.

Within each platform, methods share physics, resets, observations, descriptors, actions, and applicable SAC settings; only our FSD adaptation and SUSD port additionally receive a descriptor partition. We train three policies per method and evaluate each under three fixed reset initializations, yielding nine (training seed, reset) cells; a reset reinitializes the same policy rather than retraining it. Composition uses reset 0. Franka's RNG-independent reset yields three cells. Table 1 should be read in three stages: valid licenses the affine interface; iso and gap assess isolation and continuity; err_{in} and succ assess realized composition.

SLC introduces one additional regularization hyperparameter, $\lambda_\perp$, selected per platform from seed-0 mean iso without consulting the gate or composition (selected values $0, 1, 1, 100$), whereas the baseline implementations use their fixed settings. We therefore interpret the all-platform majority-validity comparison with this tuning difference in mind. The Humanoid choice was fixed by the historical sweep before the task-RNG evaluation correction; the main Humanoid results were reevaluated under the corrected protocol, but the sweep itself was not repeated. A single value, $\lambda_\perp = 1$, yields 9/9 valid cells on HalfCheetah, Ant, and Humanoid; Franka requires its selected value. Table 13 gives the sweep; Appendices A.1 and B.2 give implementation and evaluation provenance.

5.2 Controllable Behavior

Table 1 summarizes the results, while Figure 3 shows the underlying behavior. On HalfCheetah, Ant, and Humanoid, platform-selected SLC passes at least 8/9 cells, has inner composition errors of 3.5/6.5/7.4%, and realizes 100/100/94% of outer requests under the stricter every-axis criterion (Table 8). These reset-0 results use the fixed first-SVD null fill. Together with iso and gap, they show candidate knobs that vary, remain approximately isolated, and combine, although neither isolation nor smoothness is exact. This joint pattern—not dominance on every metric—is our main evidence of controllable behavior.

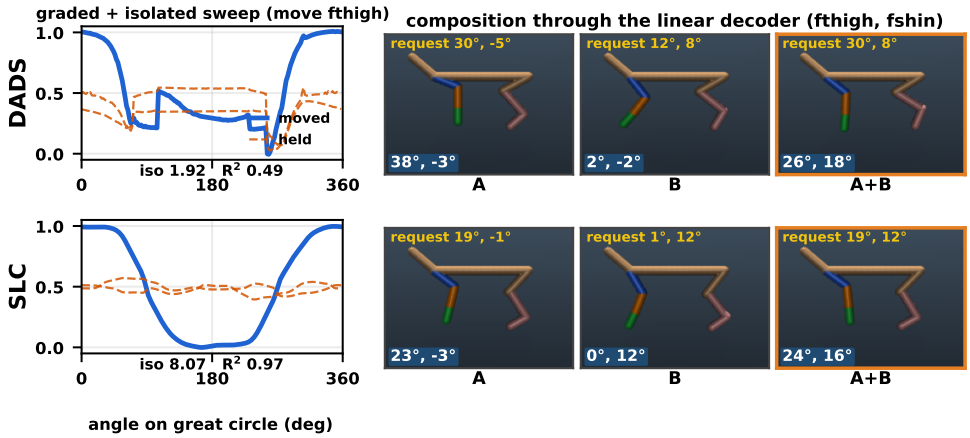

Figure 3: HalfCheetah (drawn upright, trained supine), with DADS above and SLC below. Left: a great-circle sweep tests graded `fthigh` variation (solid) and isolation of held descriptors (dashed). Right: A/B request $0.5h_i$ on one descriptor and A+B on both (yellow: requested; blue: realized). All methods and platforms are in Appendix C.2.

Franka supports only a weaker conclusion. At the gate threshold $\tau = 0.75$, it passes 3/3 cells, but only 1/3 at $\tau = 0.80$; moreover, at its $B = 3$, a synthetic step response already attains mean $R^2 \approx 0.76$. The fitted map predicts all outer targets feasible (100% reach), yet the realized response has 16.3% inner error, a $25\times$ gap, and only 53% every-axis success (75% axis-mean; Tables 12 and 8). Thus Franka contributes to the declared gate-validity result, but not to the full controllability.

For the HalfCheetah panel in Figure 3, selected SLC uses $\lambda_\perp = 0$, so DADS and SLC differ in critic architecture (and capacity) while sharing the remaining pipeline. SLC varies `fthigh` smoothly while keeping `fshin` and `ffoot` nearly fixed, and better realizes A, B, and their joint A+B request. This single-seed panel illustrates the contrast without isolating its cause.

As a positive control for the mechanics, a training-free reference applies the full action vector $a = M\alpha$ at every step. It produces near-sinusoidal responses and inner composition errors of at most 7.5% (Tables 1 and 10). It is neither skill discovery nor an upper bound: it hand-codes the skill-to-action map and covers a narrower pose range. Its role is to show that the testbeds physically admit the requested response; the learned-method question is whether reward-free USD can discover such an interface without target poses or a hand-designed controller (Appendix A.3).

The previous methods fail in different ways. DADS and METRA have no gate-valid cell, whereas CSF is affine-readable on every HalfCheetah and Ant cell but remains coupled and inaccurate. Readability is therefore necessary for our affine-map-based tests but insufficient for behavioral control. Conversely, partition-informed FSD has the highest HalfCheetah/Ant iso, but its Ant response is step-like and has larger composition error than SLC. A high isolation ratio does not guarantee controllability. The FSD/SUSD rows assess our adaptation and port; FSD is also tested outside its block-sparse training support. All methods use the same training budget, and DADS's intrinsic reward is still rising at epoch 8000, so these results do not establish its asymptotic failure (Figure 10).

5.3 Descriptor-Count Stress Test

At fixed $B = 8$, a single-seed $K \in \{2, 3, 4, 6\}$ ladder compares $\lambda_\perp = 0$ with fixed positive values $(1, 1, 1, 100)$ in platform order, without re-selecting them for each K (Figure 4; Tables 14–15). Positive-$\lambda_\perp$ SLC alone passes every tested K on HalfCheetah, Ant, and Humanoid. On Ant and Humanoid, reach remains 100% while error grows to 12.0% and 19.4%; Ant also becomes non-isolated at $K = 6$ (iso 0.68). Franka fails the gate at $K = 3, 4$ and has only 8% reach at $K = 6$. Thus validity, reach, isolation, and realized error must be interpreted jointly. A targeted three-seed SLC–DADS audit of four difficult cells leaves SLC majority-valid in all four (33/36 seed–reset cells overall) with 100% predicted reach for every seed; DADS reach is 0–67% and its validity varies by

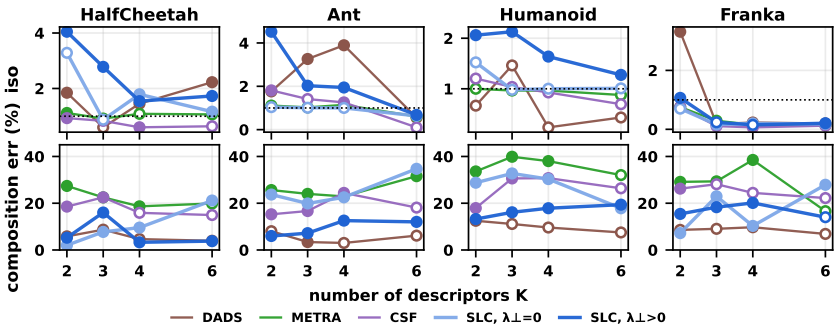

Figure 4: Descriptor-count stress test at B=8 with a single training seed. Top: isolation, with filled markers passing the R^2 gate in a majority of resets. Bottom: joint composition error, with open markers denoting reach below 50%. Full setup and values are in Appendix B.8.

setting and seed (Table 16). This audit does not replace the single-seed all-method ladder. Because the ladder changes B and the Humanoid configuration, it is a stress test, not a scaling law; varying B at fixed K also shows no monotone benefit (Table 17).

5.4 Ablating the Zero-Intercept Form and Stabilization

Across 12 reset-0 cells, adding a skill-independent intercept $q_0(s, a)$ causes 11 to fail, whereas zero-intercept SLC passes all and has higher success on every platform (Table 21). Table 1's 8/9 HalfCheetah count includes resets 1 and 2. This supports the zero-intercept form, but the intercept also changes capacity and optimization, so it does not isolate antisymmetry as the cause.

Selected-SLC critic losses remain bounded, whereas unregularized Ant and Humanoid are unstable and their embeddings grow to 10^7 and align with w; HalfCheetah succeeds and Franka fails without this growth. Magnitude-only, scalar-embedding, alternative-penalty, and shared-trunk controls succeed only in subsets of tested cells. The evidence therefore supports regularization as a stabilizer, not orthogonality or extra capacity as necessary or sufficient (Figures 10–11; Tables 18–20 and 22).

5.5 Scope and Failure Cases

Claims cover static-pose bounded joint angles. Geometric retraining passes for HalfCheetah and Ant torso-to-foot distances (with weaker Ant continuity) but fails for the tested Humanoid and Franka descriptors. All 15 combinations in an older non-partition unbounded batch fail the gate, but that batch uses METRA+SF rather than official InfoNCE CSF; neither result proves descriptor-invariant control or structural impossibility. These scope tests use one baseline seed versus three for selected geometric SLC. The default per-basis SLC critic has 3–9 times the parameters of a baseline critic, and the single-seed shared-trunk control does not separate capacity from parameter sharing (Appendix C). Composition is conditional on reset 0 and the first-SVD null fill, whose alternative shifts Humanoid descriptors by 17–26%; $\lambda_\perp$ is selected in-sample (Appendices B.5, C.1, and C.3).

6 Conclusion

Unsupervised skill discovery can produce diverse or affine-readable behaviors without yielding a usable control interface. Under SLC-only platform-specific regularization selection, the selected SLC pipeline is the only compared learned pipeline that is majority-valid across all four controlled static-pose joint-angle settings at the declared gate. On HalfCheetah, Ant, and Humanoid, it provides joint evidence for approximately controllable behavior under our three criteria, whereas Franka remains partial. The comparisons show that affine readability or a high isolation ratio alone is insufficient: responses may remain coupled, step-like, or inaccurate under composition. Ablations support the zero-intercept form and a stabilizing role for regularization, but not orthogonality as the unique mechanism. Geometric transfer is mixed, and generalization beyond static bounded descriptors to more descriptors and dynamic or contact-rich control remains open.

AI USE STATEMENT

The core idea of this work—the skill-linear critic and its use for controllable behavior as defined here—and the initial implementation of the critic were developed by the authors without AI assistance. We used generative AI assistants (Claude, Anthropic; ChatGPT and Codex, OpenAI; Gemini, Google) for research ideation and execution. This included feedback on methodology and experimental design, such as the great-circle isolation test, seed-fixed environment resets, and the joint reporting of predicted reachability and realized composition error and success; assistance with auxiliary code for baseline ports, environment variants, evaluation and analysis pipelines, debugging, figures, videos, and tables; and suggestions of candidate hypotheses and interpretations during analysis. We also used these tools to critique and refine mathematical claims and supporting arguments, including derivations, qualifications, and counterexamples. For retrieval and discovery, we used them to search for, retrieve, summarize, and compare related literature. They also assisted in drafting and restructuring portions of the manuscript, translating working drafts, and polishing the writing, including captions and keywords. Generative AI did not directly generate experimental or synthetic datasets; dataset curation and qualitative data analysis are not applicable to this work. The authors made all final decisions regarding the research questions, methodology, experimental configurations, interpretations, and claims. All AI-assisted code used for reported results was reviewed by the authors, and quantitative outputs were checked against stored evaluation records and raw logs. Claims about cited work were checked against the original sources. The authors independently verified the key derivations and reviewed every mathematical claim retained in the paper. Unsupported or inaccurate AI-proposed interpretations, arguments, and edits were rejected or corrected. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Hyperparameters shared by all platforms and methods are in Table 2, per-platform settings (skill dimension, descriptors, observation indices, $\lambda_\perp$) are in Table 3, and the environment configuration, including settings passed through environment variables that cannot be recovered from a run's argument log, is in Table 4 (Appendix A.2). Every quantitative result table is emitted by a dedicated `gen_*.py` script from stored evaluation records, without manual transcription; this includes the main table and the protocol, ablation, geometric-descriptor, and unbounded-descriptor appendices. The final generators select the official InfoNCE CSF runs and the task-RNG-corrected Humanoid evaluation records. Section 4 defines the evaluation protocol, and Appendices B.2–B.6 give all constants and aggregation conventions. The anonymous supplement supports audit and reuse by releasing the SLC-specific critic and training-integration excerpts, a framework-independent reference implementation of the control metrics, the R^2 numerical check and its record, `experiment_manifest.md`, and a rendered supplementary video. It does not provide an end-to-end executable snapshot of the common training fork and does not redistribute third-party baseline repositories, raw checkpoints, replay buffers, or cluster infrastructure.

REFERENCES

Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299*, 2018.

Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Vassil Atanassov, Wanming Yu, Alexander Luis Mitchell, Mark Nicholas Finean, and Ioannis Havoutis. Constrained skill discovery: Quadruped locomotion with unsupervised reinforcement learning. *arXiv preprint arXiv:2410.07877*, 2024.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Jonathan Hunt, Shibl Mourad, David Silver, and Doina Precup. The option keyboard: Combining skills in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. Universal successor features approximators. In *International Conference on Learning Representations*, 2019.

Wilka Carvalho, Angelos Filos, Richard Lewis, Honglak Lee, and Satinder Singh. Composing task knowledge with modular successor feature approximators. In *International Conference on Learning Representations*, 2023.

Rafael Cathomen, Mayank Mittal, Marin Vlastelica, and Marco Hutter. Divide, discover, deploy: Factorized skill learning with symmetry and style priors. In *Conference on Robot Learning*, pp. 750–768, 2025.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Jonathan Colaço Carr, Qinyi Sun, and Cameron Allen. Focused skill discovery: Learning to control specific state variables while minimizing side effects. *Reinforcement Learning Journal*, 6:2585–2599, 2025.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6637–6647, 2020.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.

Carlos Florensa, Yan Duan, and Pieter Abbeel. Stochastic neural networks for hierarchical reinforcement learning. In *International Conference on Learning Representations*, 2017.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *International Conference on Learning Representations, Workshop Track*, 2017.

Tuomas Haarnoja, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. Composable deep reinforcement learning for robotic manipulation. In *IEEE International Conference on Robotics and Automation*, pp. 6244–6251, 2018a.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018b.

Steven Hansen, Will Dabney, Andre Barreto, Tom Van de Wiele, David Warde-Farley, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Seyed Mohammad Hadi Hosseini and Mahdieh Soleymani Baghshah. SUSD: Structured unsupervised skill discovery through state factorization. In *International Conference on Learning Representations*, 2026.

Jiaheng Hu, Zizhao Wang, Peter Stone, and Roberto Martín-Martín. Disentangled unsupervised skill discovery for efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 76529–76552, 2024.

Jiaheng Hu, Peter Stone, and Roberto Martín-Martín. SLAC: Simulation-pretrained latent action space for whole-body real-world RL. In *Conference on Robot Learning*, 2025.

Yiming Huang, Zhiyang Dou, and Lingjie Liu. ModSkill: Physical character skill modularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12394–12404, 2025.

Paul Jansonnie, Bingbing Wu, Julien Perez, and Jan Peters. Unsupervised skill discovery for robotic manipulation through automatic task generation. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pp. 926–933. IEEE, 2024.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Hao Liu and Pieter Abbeel. APS: Active pretraining with successor features. In *International Conference on Machine Learning*, pp. 6736–6747, 2021.

Geraud Nangue Tasse, Steven James, and Benjamin Rosman. A boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9497–9507, 2020.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*, 2022.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. In *International Conference on Machine Learning*, pp. 27225–27245. PMLR, 2023.

Seohong Park, Oleh Rybkin, and Sergey Levine. METRA: Scalable unsupervised RL with metric-aware abstraction. In *International Conference on Learning Representations*, 2024.

Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. MCP: Learning composable hierarchical control with multiplicative compositional policies. *Advances in Neural Information Processing Systems*, 32, 2019.

Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pp. 188–204. PMLR, 2021.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320, 2015.

Archit Sharma, Michael Ahn, Sergey Levine, Vikash Kumar, Karol Hausman, and Shixiang Gu. Emergent real-world robotic skills via unsupervised off-policy reinforcement learning. In *Robotics: Science and Systems*, 2020a.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*, 2020b.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065. PMLR, 2019.

Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable factors. *arXiv preprint arXiv:1708.01289*, 2017.

Emanuel Todorov. Compositionality of optimal control laws. *Advances in neural information processing systems*, 22, 2009.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13–23, 2021.

Julen Urain, Puze Liu, Anqi Li, Carlo D'Eramo, and Jan Peters. Composable Energy Policies for Reactive Motion Generation and Reinforcement Learning. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021. doi: 10.15607/RSS.2021.XVII.052.

Zizhao Wang, Jiaheng Hu, Caleb Chuck, Stephen Chen, Roberto Martín-Martín, Amy Zhang, Scott Niekum, and Peter Stone. SkiLD: Unsupervised skill discovery guided by factor interactions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77748–77776, 2024.

Tom Zahavy, Brendan O'Donoghue, Andre Barreto, Volodymyr Mnih, Sebastian Flennerhag, and Satinder Singh. Discovering diverse nearly optimal policies with successor features. In *ICML Workshop on Unsupervised Reinforcement Learning*, 2021. arXiv:2106.00669.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a MISL fly? analysis and ingredients for mutual information skill learning. In *International Conference on Learning Representations*, 2025.

APPENDIX

A TRAINING, ENVIRONMENTS, AND STIFFNESS CONTROLS

Appendix A gives training, environment, and joint-response details. Appendix B specifies the evaluation protocol, its robustness checks, and the full tables. Appendix C collects additional ablations, descriptor scope, visualizations, and failures on unbounded axes.

A.1 TRAINING DETAILS

This section gives the SAC target and intrinsic-reward estimator underlying Algorithm 1, together with shared optimizer and network settings (Table 2) and per-platform settings (Table 3).

Table 2 separates shared settings from method-specific ones; FSD alone uses a block-sparse skill prior. In Table 3, B is both the skill dimension and, for SLC, the number of basis critics, whereas K is the descriptor count; $B > K$ throughout. The listed B is shared across methods except on Franka, where FSD and SUSD use $B = 4$ rather than 3 because their blocks require $K \mid B$. The platform-specific $\lambda_\perp$ comes from the seed-0 iso sweep (Section 5.1, Table 13), and descriptor names follow the MuJoCo model files.

Baseline implementations. All methods run in one codebase fork. DADS is ported from METRA's PyTorch implementation, SUSD uses a constructed partition and per-factor distance based on its public implementation, and FSD is our continuous block-sparse adaptation whose dense-sphere evaluation is off-support.

Clipped-double-Q SAC target. For a replay transition $(s, a, \alpha, s', \delta_t)$, each critic head is regressed toward the stop-gradient target

$$y_{\mathrm{SAC}} = r_t + \gamma(1-\delta_t)\,\mathbb{E}_{a'\sim\pi(\cdot|s',\alpha)}\Big[\min_{m=1,2} \bar{Q}_\alpha^{(m)}(s',a') - \beta^{-1}\log\pi(a' \mid s',\alpha)\Big]. \tag{9}$$

Here $r_t = r_{\mathrm{impl}}$ is the shared DADS-style intrinsic reward, δ_t is the termination indicator, $\beta > 0$ is the inverse SAC entropy temperature, and $\bar{Q}_\alpha^{(m)}$ are the two target critics. Each target head is exactly linear in α, while their pointwise minimum is piecewise linear.

All parameter groups use Adam with PyTorch defaults and learning rate 10^{-4}. The entropy coefficient β^{-1} is initialized to 0.01 and tuned automatically toward target entropy $-\dim(\mathcal{A})/2$ using the standard SAC temperature loss.

METRA skill-dimension check. To check for a more favorable baseline dimension, we compared $B \in \{4, 6, 8, 10\}$ on HalfCheetah with a single seed. No value passed the gate, and all ungated isolation estimates were below 1 (Table 17).

Intrinsic-reward estimator details. The implemented skill dynamics q_ξ is a unit-variance diagonal Gaussian whose mean is a 2×1024 MLP. It is trained by maximum likelihood to predict the normalized state change $\overline{\Delta s}$, where $\Delta s = s' - s$, from the normalized state input $\bar{s}$ and skill α. Both $\bar{s}$ and $\overline{\Delta s}$ are batch-normalized and then have their non-descriptor coordinates set to zero. The Gaussian log-density is nevertheless summed over every observation coordinate, so a masked coordinate j still contributes $-\mu_j^2/2$ plus a skill-independent constant, where μ_j is the corresponding predicted-mean coordinate. Let $\ell_0 = \log q_\xi(\overline{\Delta s} \mid \bar{s}, \alpha)$ and $\ell_u = \log q_\xi(\overline{\Delta s} \mid \bar{s}, \tilde{\alpha}_u)$ for N_{alt} alternative skills. The implemented intrinsic reward is

$$r_{\mathrm{impl}} = \log(N_{\mathrm{alt}}+1) - \log\left[1 + \sum_{u=1}^{N_{\mathrm{alt}}} \exp\big(\,\mathrm{clip}(\ell_u - \ell_0, -50, 50)\big)\right]. \tag{10}$$

DADS and SLC use this identical estimator with $N_{\mathrm{alt}} = 100$. The alternative skills are unnormalized standard-normal samples rather than draws from the sphere prior, so r_{impl} is an MI-motivated surrogate rather than a strict MI lower bound. The policy is the Gaussian policy of SAC with its output squashed by $\tanh$. The policy and critic receive the full observation except on HalfCheetah, where both mask the absolute forward position x. The same observation mask is applied to policy calls and to the critic and actor updates.

Table 2: Shared and method-specific optimizer, rollout, and network hyperparameters. Method-specific entries are labeled; defining flags are listed below.

Parameter	Value	Parameter	Value
Training epochs	8000	Learning rate (all networks)	1×10^{-4}
Episode length	200	Discount γ	0.99
Trajectories per epoch	8	Target smoothing (Polyak) τ_{tgt}	0.005
Inner steps N_{opt}	50	SAC target-entropy coef.	1.0
Minibatch size	256	Random warm-up trajectories	48
Replay buffer capacity	1×10^{6}	Reward scale	1.0
Replay warm-up	1×10^{4}	MLP hidden layers	2×1024
Basis embedding dim. n (SLC)	64	Observation normalizer	off
Alternative skills N_{alt} (DADS/SLC)	100	Spectral normalization	off
Skill prior (non-FSD)	$\text{Unif}(\mathbb{S}^{B-1})$	Skill-dynamics batch norm (DADS/SLC)	on

Table 3: Per-platform environment, skill, descriptor, regularization, observation, and worker settings.

	HalfCheetah	Ant	Humanoid	Franka
Environment	`half_cheetah`	`ant`	`dmc_humanoid_state`	`kitchen_state`
Basis critics B	6	4	8	3
Descriptors K	3	2	2	2
Orthogonality $\lambda_\perp$	0	1	1	100
Observation index	$\{6,7,8\}$	$\{8,12\}$	$\{15,18\}$	$\{1,3\}$
Descriptor	`fthigh`, `fshin`, `ffoot`	`ankle_1`, `ankle_3`	`right_shoulder1`, `left_shoulder1`	`panda0_joint2`, `panda0_joint4`
Parallel workers	1	1	1	8

Method-defining flags. METRA uses the dual-constrained inner-product objective (`inner=1`, `dual_reg=1`, `dual_lam=30`, `dual_slack=0.001`). CSF uses the official contrastive path (`log_sum_exp=1`, `sample_new_z=1`, 256 negatives, `infonce_lam=5`). Although the CSF logs retain the dual flags, the `log_sum_exp=1` branch bypasses both the dual penalty and the dual-λ update, so they are inactive on this path. SUSD uses `dual_dist=s2_from_s` and curiosity weighting, whereas FSD uses $\lambda = 2$. Appendix tables that explicitly say METRA+SF use the METRA objective with the SF critic and are not CSF.

SUSD port. With no explicit `susd_partition`, our port makes one factor from each unmasked descriptor and divides the B-dimensional skill into K equal blocks. For factor i, it uses $r_i = (\phi_i(s') - \phi_i(s))^\top \alpha_i$. Let the diagonal `s2_from_s` model predict $q(\Delta s \mid s) = \mathcal{N}(\mu(s), \operatorname{diag}\sigma^2(s))$, let D_{obs} be the observation dimension, and set $g = \exp[D_{\text{obs}}^{-1}\sum_j \log\sigma_j^{-1}]$. For the coordinate set I_i of factor i, our CSD distance is $\chi_i = D_{\text{obs}}^{-1}\sum_{j\in I_i}(\Delta s_j - \mu_j)^2(\sigma_j^{-1}/g)^2$. The predictor uses a 2×1024 MLP and diagonal standard deviations clipped to $[10^{-6}, 10^{6}]$. The SAC intrinsic reward is $\sum_i r_i\sqrt{\chi_i}$, whereas the representation objective uses the unweighted sum $\sum_i r_i$ plus METRA's dual constraint, matching the source implementation's separation of policy and encoder weighting. All reported SUSD runs use `susd_curiosity=1` and require $K \mid B$.

Single-head Boltzmann interpretation and twin critics. For a fixed critic, entropy-regularized policy improvement has the Boltzmann target $\pi^*(a \mid s) \propto \exp(\beta Q(s,a))$ (Haarnoja et al., 2018b), where $\beta > 0$ is the inverse temperature. Substituting Eq. (2) gives

$$\pi^*_\alpha(a \mid s) \propto \exp\big(\beta Q_\alpha(s,a)\big) \propto \prod_{k=1}^{B} \pi_k(a \mid s)^{\alpha_k}, \qquad \pi_k(a \mid s) \propto \exp\big(\beta c_k(s,a)\big). \tag{11}$$

Thus each head admits a signed weighted-product interpretation. In our $\tanh$-bounded action spaces, continuous basis critics are bounded over actions for each fixed state, so the partition functions remain finite even when $\alpha_k < 0$ and a factor is inverted. The product combines functions of

the same action. It is not an independence factorization across action components. SAC updates the actor with the minimum of two critics, so each head is linear in α but the clipped critic entering the actor update is piecewise linear in α: $\min(Q_\alpha^{(1)}, Q_\alpha^{(2)}) = Q_\alpha^{\text{avg}} - \frac{1}{2}|Q_\alpha^{(1)} - Q_\alpha^{(2)}|$. Equation (11) therefore interprets each head, not generally the actor's exact target, and neither structure guarantees graded or isolated behavior.

FSD, continuous adaptation. The original FSD formulation permits a nonempty set of target variables. Its reported experiments use grid worlds with tabular Q-learning, 16 discrete skills, and one target variable per focused skill. Our implementation retains three principles: (i) per-axis representations ϕ_i that take only that axis as input, (ii) a side-effect penalty on changes in non-target variables, and (iii) the association of each training skill with a single target variable, instantiated here as a block-structured skill that is nonzero only on the target block. The target axis is sampled uniformly from the K axes. Its block has dimension B/K, is sampled by normalizing a standard Gaussian vector to the unit sphere, and all other blocks are set to zero. Let r_j denote the normalization scale of descriptor j. Per-axis normalization is available in the port, but `focused_ranges` was unset in every reported run, so $r_j = 1$ and the experiments penalize changes in raw descriptor units (radians for joint-angle descriptors). Our choices are (a) to restrict the penalty to the held descriptor axes and instantiate it at time t as $\lambda\|((d_j(s_t) - d_j(s_0))/r_j)_{j\neq i_{\text{tar}}}\|_2$ for target axis i_{tar}, (b) to give the representation term and the episode-start-relative penalty at every step, and (c) to use METRA's dual Lipschitz constraint. Thus it differs from our METRA baseline in all three retained ingredients. The penalty weight is $\lambda = 2$. All other hyperparameters equal METRA's. Because FSD trains on block-sparse skills, evaluation on the dense sphere is outside its training distribution.

METRA+SF. The SF-critic baseline of Tables 5 and 17, Figure 7, and Appendices C.1 and C.3 runs the `metra_sf` algorithm of the CSF codebase without the InfoNCE term: the representation ϕ is trained with the METRA objective (including its dual constraint) and only the critic takes the successor-feature form $\psi(s, a, z)^\top z$ of CSF. We label this variant METRA+SF to distinguish it from the official InfoNCE CSF used in Table 1 and the K ladder of Section 5.3.

A.2 Environment Configuration

Table 4 gives the per-platform configuration, including settings absent from argument logs. Rows below the rule are set through environment variables rather than command-line arguments and cannot be recovered from a run's argument log. The model entry points are `half_cheetah.xml`, `ant.xml`, DM Control's `humanoid_color.xml`, and Franka's `franka_kitchen_torque_ctrl.xml` include stack. HalfCheetah uses the springs and damping already specified in its XML. Placing it on its back removes leg–ground contact and friction from joint bending without pinning the free root. Gravity is reversed in the torso frame, but the native stiffness limits displacement. A separate upright-training check restores contact and is reported only as a robustness test (Figure 7). Main-text poses are rotated upright only for rendering. The stock Ant ankle angle lies outside its joint range, so stiffness 430 produces about 375 N·m at the first step and flips the body in 1 of 20 passive resets. Initializing all four ankles at their spring midpoints removes this transient (Figure 5). Its torso, hips, and all eight actuators remain free.

A Humanoid spawned upright topples under zero action in every reset (median 14 steps), so we fix the torso (Figure 6). In the main comparison, we also reset the waist/legs and elbows to their initial positions and zero velocities after every environment step: free elbows produced large non-target motion that contaminated the shoulder response. These are runtime snapbacks rather than XML constraints, and the policy retains all 21 actuator outputs, so reset coordinates can exert forces within a physics step. The four shoulder axes remain free and spring-loaded, although only the two `shoulder1` angles are descriptors. The Humanoid K ladder instead unlocks and spring-loads the elbows, and its hand-distance study frees all six arm joints. Franka uses a nine-action torque controller. Runtime stiffness 60 and damping 3 are applied to all seven arm joints. Nonrobot, non-floor contacts are disabled and object DOFs frozen so kitchen objects cannot perturb the static arm response, while the 60-dimensional observation layout is retained. The main Franka observation appends neither end-effector coordinates nor arm velocities. Within every main platform these choices are shared by all compared methods. HalfCheetah, Ant, and Franka also hold their ladder physics fixed across K. The geometric studies intentionally change the descriptor/observation definition,

and Ant additionally flips the body to remove foot contact. These cross-group differences are not pooled with the main comparison.

Table 4: Platform-specific environment configuration for the main comparison unless a row says otherwise.

Platform	Configuration
HalfCheetah	native XML springs; `spawn_flipped=1`, `policy_omit_idxs=[0]`; six actions
Ant	free torso; `ankle_stiffness=430`, `hip_stiffness=0`, `ant_ankle_init_at_spring=1`
Humanoid	torso and waist/legs reset; elbows reset; four shoulder axes free; 21 actions
Franka	`kitchen_control_mode=torque`, `kitchen_ee_in_obs=0`; nine actions
Humanoid (env. var.)	`HUMANOID_PIN=1`, `HUMANOID_PIN_Z=1.6`, `HUMANOID_LOCK_LEGS=1`, `HUMANOID_LOCK_ELBOWS=1`, `HUMANOID_SHOULDER_STIFF=16`, `HUMANOID_SHOULDER_DAMP=1.5`
Humanoid (K ladder)	`HUMANOID_LOCK_ELBOWS=0`, `HUMANOID_ELBOW_STIFF=32`, `HUMANOID_ELBOW_DAMP=5.66`
Franka (env. var.)	`FRANKA_ARM_STIFF=60`, `FRANKA_ARM_DAMP=3`, `FRANKA_NO_KITCHEN=1`, `FRANKA_HIDE_KITCHEN=1`

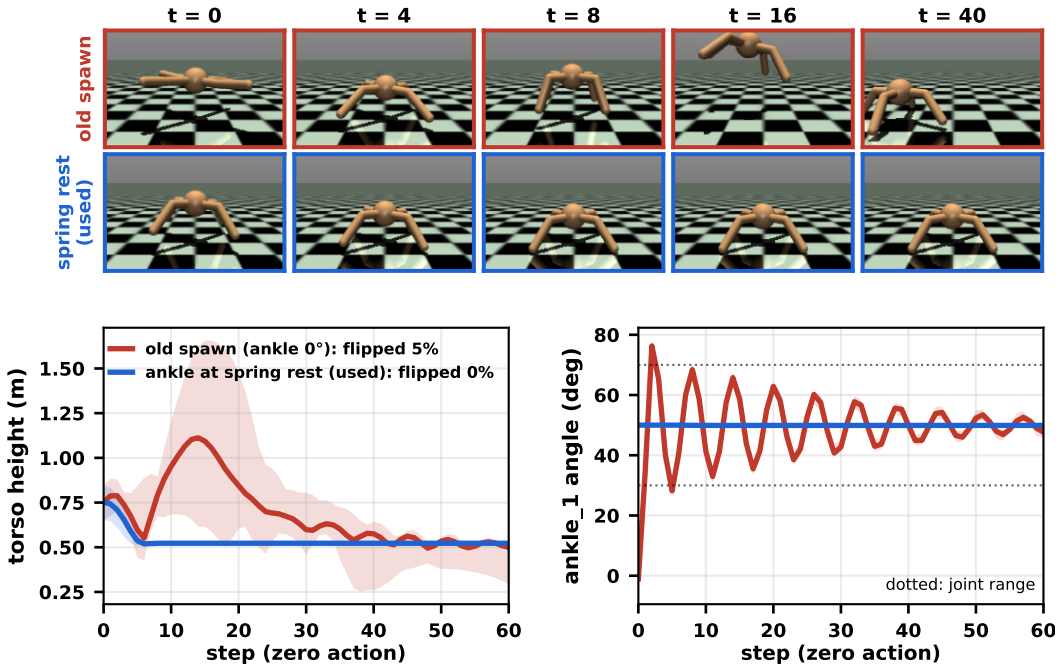

Figure 5: Ant zero-action reset diagnostic over 20 resets. With the stock ankle initialization (top), stiffness 430 creates an approximately 375 N·m first-step spring transient and flips the body in 1/20 resets. Initializing all four ankles at the spring-rest midpoints (bottom) removes this transient. The latter setting is shared by all methods in the main Ant comparison.

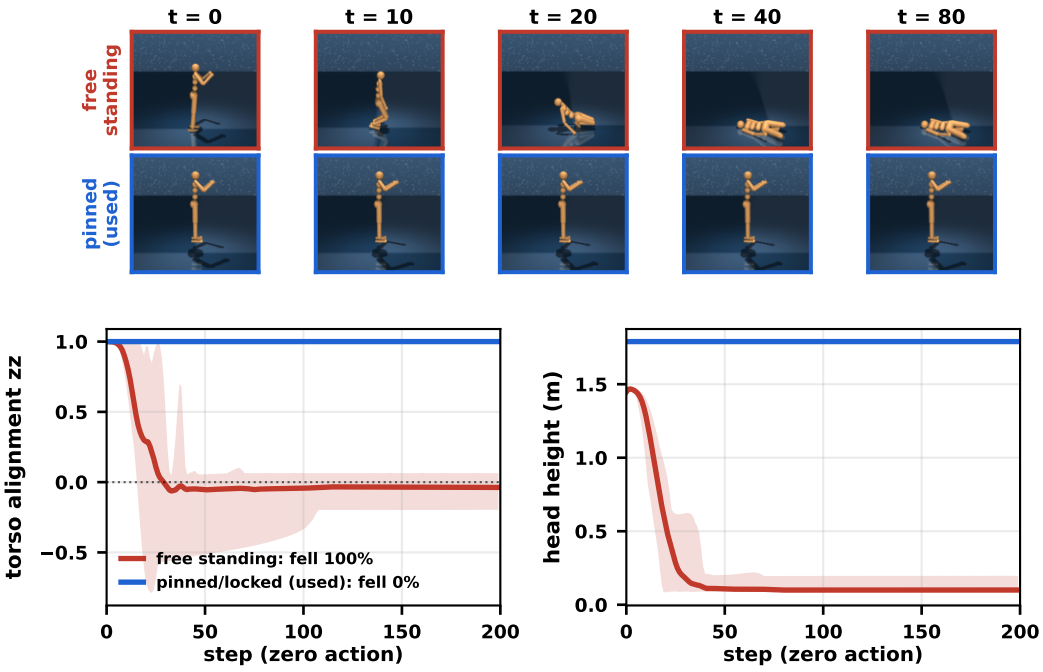

Figure 6: Humanoid zero-action reset diagnostic over 20 resets. The free-standing model (top) topples in every reset (median 14 steps), coupling shoulder motion to balance and falling. Torso pinning (bottom) keeps torso alignment and head height stable and is shared by all reported static-pose Humanoid configurations. Arm-joint freedom still differs across the main comparison, K ladder, and geometric-descriptor follow-up.

A.3 Joint Stiffness: Open-Loop Selection and Spring Control

Joint stiffness was set before training by an open-loop test. We chose approximate linearity of the settled angle after 200 steps of constant torque as a design criterion for this testbed. Without it, the default response often saturates into a step-like curve, although nonlinear responses need not be discontinuous in general. The joints of HalfCheetah are already approximately linear with their native stiffness. For Ant, Humanoid, and Franka, we use the smallest tested stiffness that gives an approximately linear response over the evaluated range. Larger values shrink the reachable range. Figure 8 shows the tested curves. The selected values are 430 for the Ant ankles, 16 for the Humanoid shoulders, and 60 for the Franka arm.

Table 5 compares ankle stiffness 0 and 480 in an earlier pinned-Ant setting. Both conditions are evaluated at epoch 8000; the stiffness-0 runs ended there, whereas the stiffness-480 runs continued to epoch 20,000. The torso is fixed, ankles 1–2 are the descriptors, $B = 3$, and the comparison uses one seed and reset with $\lambda_\perp = 0$ for SLC. This differs from the main Ant setting, which has a free torso, ankles 1 and 3 as descriptors, and $B = 4$, so only the effect of the spring within this historical setting should be read from the table.

Without the spring, the settled ankle response is step-like (Figure 8) and no method passes the gate; SLC is borderline with a gap of about $24\times$ the ideal. With the spring, SLC passes ($R^2_{\min} = 0.89$, iso 1.49, gap $3\times$), whereas the baselines remain below the gate with ungated iso estimates of 0.76–1.08. Thus this control provides gate-valid isolation and continuity evidence only for SLC, not a method-independent isolation effect of the spring, and it does not by itself explain the main-table method differences.

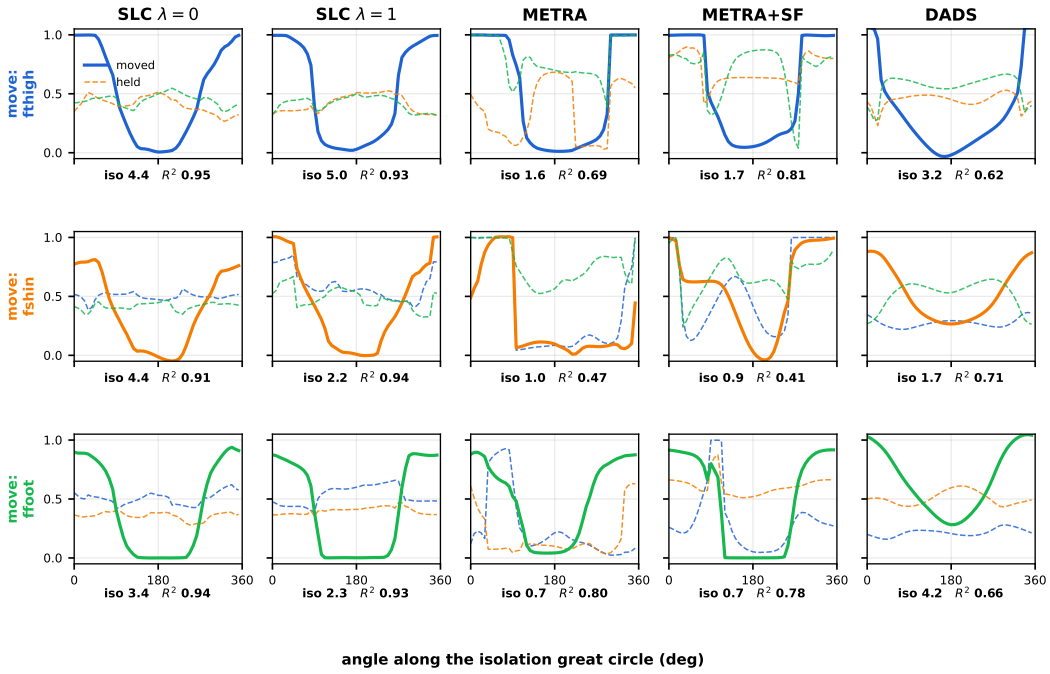

Figure 7: Upright-trained HalfCheetah robustness check with feet on the ground, one training seed, and one reset. Rows move each descriptor along its isolation great circle. Columns compare SLC with $\lambda_\perp \in \{0, 1\}$, METRA, METRA+SF, and DADS. Both SLC variants pass the gate ($R^2_{\min} = 0.91, 0.93$), whereas METRA, METRA+SF, and DADS do not ($0.47, 0.41, 0.62$). Panel labels give descriptor-wise iso_i and R^2_i.

Table 5: Effect of ankle stiffness in the historical pinned-Ant control at epoch 8000 (one seed and reset). Gray iso and gap entries fail the main $R^2_{\min} \geq 0.75$ gate; "border" denotes $0.60 \leq R^2_{\min} < 0.75$ as a diagnostic only.

	stiffness 0			stiffness 480		
Method	iso	$R^2_{\min}$ (gate)	gap	iso	$R^2_{\min}$ (gate)	gap
DADS	0.49	0.03 (fail)	45	0.76	0.65 (border)	13
METRA	0.79	0.71 (border)	53	0.96	0.51 (fail)	49
METRA+SF	1.00	0.68 (border)	57	1.08	0.58 (fail)	16
SLC ($\lambda_\perp$=0)	1.10	0.62 (border)	24	1.49	0.89 (pass)	3

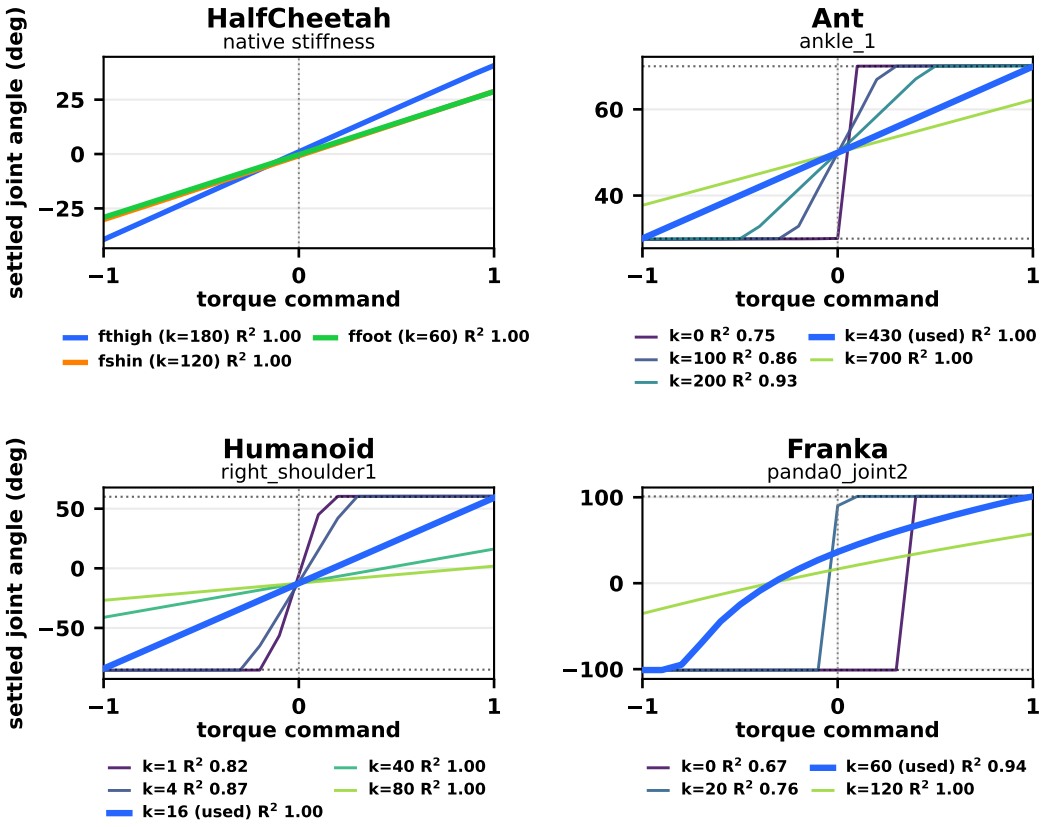

Figure 8: Open-loop descriptor-joint response after 200 steps of constant torque. HalfCheetah shows the native stiffness of each of its three joints. For Ant, Humanoid, and Franka, the thick blue curve is the selected stiffness (430, 16, and 60), and thinner curves are alternatives. Dotted horizontal lines mark joint limits.

B Evaluation Protocol, Robustness Checks, and Full Results

This section adds the sample sizes, window lengths, reset convention, and aggregation to the definitions of Section 4. Every table is emitted directly as LaTeX by a script that reads the evaluation logs and npz files. No number was transcribed by hand.

B.1 Decoder Scope and Robustness

Definitions and caveats. We deliberately use an affine model as the primary decoder rather than selecting the decoder class that maximizes predictive accuracy. The property under test is a simple global control interface: one map should relate skills to realized descriptor responses over the sampled sphere. For descriptor i, the affine model $\hat{D}_i(\alpha) \approx g_i^\top \alpha + b_i$ has the constant gradient g_i. These global directions make the downstream tests well-defined: projecting g_i onto the subspace orthogonal to the held-descriptor rows constructs an isolation path, while the pseudoinverse G^+ supplies a closed-form inverse for joint targets. The affine model also matches the training-free linear action reference.

A more flexible decoder asks the different question of whether behavior is predictably nonlinear. For example, a quadratic model $\hat{D}_i(\alpha) \approx b_i + g_i^\top \alpha + \alpha^\top H_i \alpha$ has gradient $g_i + (H_i + H_i^\top)\alpha$, which depends on the current skill. It therefore provides neither the single global descriptor directions used by the isolation test nor the global pseudoinverse used by the composition test. Failure of the affine gate consequently does not rule out control under a nonlinear decoder; it means that the response does not satisfy the affine interface evaluated here.

Table 6 checks decoder flexibility on the main-table runs using three seeds, reset 0, and the common target scale. The quadratic decoder fits $\hat{D}$ on $[\alpha, \alpha_i\alpha_j, 1]$, and both decoders use the same dense search over 2×10^4 sphere candidates, so only the decoder differs. This search-based inversion differs from the main protocol's closed-form inversion, making absolute values incomparable across the two tables. $R^2_{\min}$ is in-sample, while $\mathrm{err}_{\mathrm{in}}$ and succ follow Table 1; affine-gate failures remain diagnostic. Quadratic fits improve several baselines, but their effects on realized composition are method-dependent, and selected SLC remains competitive.

$\mathrm{iso} = \min_i \mathrm{iso}_i$ is sensitive to one noisy axis. A cell is invalid if a descriptor has zero sampled variance, the held gradients are dependent, or g_i lies in their span. None of these geometry failures occurred in the 210 main-comparison cells. For composition, the evaluator computes rank with cutoff $10^{-6}\sigma_{\max}(G)$ and aborts if $\mathrm{rank}\, G < K$. Thus every reported reach value has full row rank under this cutoff. The realized descriptor response is reset-conditioned and averaged over the late window; it does not test settling or within-window variation.

Table 6: Affine versus quadratic decoders on the main-table runs (three-seed means at reset 0). $R^2_{\min}$ is in-sample; $\mathrm{err}_{\mathrm{in}}$ and succ follow Table 1.

Platform	Method	$R^2_{\min}$		$\mathrm{err}_{\mathrm{in}}$ (%)		succ (%)	
		affine	quad	affine	quad	affine	quad
HalfCheetah	DADS	0.26	0.58	11.6	8.7	67	86
	METRA	0.70	0.82	23.1	20.7	61	64
	CSF	0.80	0.83	24.0	25.2	50	44
	FSD	0.89	0.95	4.5	3.8	100	100
	SUSD	0.41	0.57	7.5	7.0	94	83
	SLC, $\lambda_\perp$=0	0.94	0.97	4.0	3.8	100	100
	SLC, $\lambda_\perp$>0	0.94	0.96	5.0	4.9	100	100
Ant	DADS	0.31	0.66	12.8	7.7	61	78
	METRA	0.46	0.69	35.2	31.4	53	58
	CSF	0.79	0.81	24.4	25.7	81	86
	FSD	0.89	0.91	13.0	13.6	97	94
	SUSD	0.21	0.33	10.2	10.4	6	8
	SLC, $\lambda_\perp$=0	0.36	0.45	27.9	27.5	44	50
	SLC, $\lambda_\perp$>0	0.89	0.93	5.6	5.3	100	100
Humanoid	DADS	0.38	0.66	15.3	13.2	44	56
	METRA	0.58	0.82	25.4	26.0	58	64
	CSF	0.73	0.83	23.4	27.0	78	56
	FSD	0.64	0.83	13.1	14.2	78	83
	SUSD	0.48	0.71	9.4	10.9	50	78
	SLC, $\lambda_\perp$=0	0.64	0.79	25.1	25.2	72	75
	SLC, $\lambda_\perp$>0	0.86	0.94	12.4	8.0	100	100
Franka	DADS	0.18	0.52	14.0	11.7	8	33
	METRA	0.68	0.77	27.4	25.9	50	61
	CSF	0.49	0.66	24.5	22.4	53	58
	FSD	0.05	0.73	34.8	27.1	25	56
	SUSD	0.48	0.65	16.8	15.8	69	69
	SLC, $\lambda_\perp$=0	0.38	0.58	19.8	18.2	53	69
	SLC, $\lambda_\perp$>0	0.79	0.88	14.4	12.6	72	86

B.2 Rollouts, Fits, and the Affine Gate

Rollouts. Policies are the checkpoints at epoch 8000, and evaluation uses the deterministic action obtained by applying $\tanh$ to the pre-$\tanh$ Gaussian mean. This is the implementation's deterministic-action path, not generally the density mode of the squashed policy, and it applies the same observation mask used during training. Episodes are $T = 200$ steps and the late window is the last $T_{\mathrm{eval}} = 40$ steps. We use the final fifth of every episode as a fixed averaging window across methods and platforms; its length was not tuned or subjected to a sensitivity analysis. Both NumPy's global RNG and the `dm_control` task RNG are restored before every rollout so that all skills within one test see the same reset realization. Fit and isolation-circle sensitivity are measured by repeating with different reset seeds, whereas composition uses reset 0 only. Franka's reset does not depend on the RNG, so repeated resets are identical and only reset 0 is evaluated.

Humanoid evaluation correction. An earlier evaluator reset NumPy's RNG but allowed the internal `dm_control` task RNG to advance as skills were evaluated in a reused environment. This made one Humanoid reset depend on evaluation order. The corrected evaluator recreates the environment for each run and restores both RNG states before every rollout. The main comparison, descriptor-count ladder, magnitude/scalar controls, qualitative panels, and their common composition scale use the corrected records. Table 13 is the exception: it preserves the historical selection sweep, so its Humanoid spread and the choice $\lambda_\perp = 1$ precede the correction. We did not repeat that sweep; all headline Humanoid measurements at the fixed value $\lambda_\perp = 1$ use the corrected evaluator.

Affine fit and gate. We draw $N = 150$ skills on the sphere by normalizing standard normal vectors, and all methods with the same B use the same sample (sphere seed 0). A least-squares affine fit with an intercept gives a per-descriptor R^2. A cell is valid only if every descriptor has nonzero sampled variance, G has full row rank, and $R^2_{\min} = \min_i R^2_i \geq 0.75$. These conditions exclude frozen descriptors, linearly dependent control directions, and responses poorly described by one affine map, respectively. Below this prespecified fit-quality gate, metrics that depend on G—iso, gap, gradient angles, and the condition number—are shown in gray and excluded from comparisons. The composition error and success are realized observations and are reported, but likewise in gray and excluded from the bold comparison. $\angle_{\min}$ is the smallest angle between the fitted gradients (90° is orthogonal) and κ is the condition number of G (1 is isotropic). Passing the gate establishes only

that the sampled affine-fit criterion is met, while failing it does not rule out smooth control under a nonlinear decoder.

$$\{\alpha^{(m)}\}_{m=1}^{N} \sim \mathrm{Unif}(\mathbb{S}^{B-1}), \qquad N = 150, \tag{12}$$

$$(G, b) = \arg\min_{G,b} \sum_{m=1}^{N} \left\| \hat{D}(\alpha^{(m)}) - G\alpha^{(m)} - b \right\|^2,$$

$$\text{valid} \iff \min_i R_i^2 \geq 0.75, \quad \mathrm{rank}_{10^{-6}\sigma_{\max}} G = K, \quad \min_i \sum_m \left(\hat{D}_i(\alpha^{(m)}) - \bar{D}_i\right)^2 > 10^{-12}. \tag{13}$$

Here $\bar{D}_i$ is the sample mean of realized descriptor-response coordinate i, and $\sigma_{\max}(G)$ and $\sigma_{\min}(G)$ below are the largest and smallest singular values of G. After the stated nonzero-variance check, the evaluator computes the ordinary coefficient of determination $R_i^2 = 1 - \mathrm{SSE}_i/\mathrm{SS}_i$, where SSE_i and SS_i are the residual and total sums of squares. An earlier implementation added 10^{-9} to SS_i. A read-only numerical check of all 3,643 reconstructible archived fits found a maximum per-descriptor difference of 4.20×10^{-8} and no gate change at any threshold reported in Table 7.

$$\angle_{\min} = \min_{i<j} \arccos \frac{|g_i^\top g_j|}{\|g_i\| \, \|g_j\|}, \qquad \kappa = \frac{\sigma_{\max}(G)}{\sigma_{\min}(G)}. \tag{14}$$

B.3 Gate Robustness and Realized Success

Gate threshold. The gate is needed because iso, $\angle_{\min}$, κ, and composition all use the same affine fit. Reading iso below the gate produces spurious wins: 29 of 210 cells have $R^2_{\min} < 0.60$ yet iso at least 1.5 or $\angle_{\min} \geq 75^\circ$. The threshold 0.75 has two motivations. In synthetic great-circle response classes—sinusoidal (graded) versus sawtooth (discontinuous), each with 4000 noise-perturbed draws—the equal-error threshold is 0.706, and 0.75 gives 0% false positives and 0.2% false negatives. We transfer 0.75 heuristically, without independent validation, to the random-sphere affine fit. Separately, across 400 synthetic draws, the mean R^2 of a perfect step response fitted from 150 sphere samples is $0.76, 0.73, 0.71, 0.70$ for $B = 3, 4, 6, 8$. Thus it passes on average at $B = 3$, the dimension used by Franka except for FSD and SUSD ($B = 4$), and lies near the gate at $B = 4$, so gap remains necessary. Table 7 varies the threshold from 0.60 to 0.90. The selected SLC verdict is unchanged through 0.80 on HalfCheetah, Ant, and Humanoid, while Franka falls from 3/3 to 1/3 at 0.80. SLC's all-platform majority pass is unique among the compared learning pipelines for $0.65 \leq \tau \leq 0.75$. At $\tau = 0.60$, CSF also passes in a majority of cells on all four platforms, and at $\tau \geq 0.80$ SLC no longer has a majority on Franka. Thus the selected SLC rows' own gate verdict is stable through 0.80 on three platforms but threshold-sensitive on Franka, and the all-platform uniqueness claim holds only over the stated interval. Figure 9 shows the empirical cell distribution around these thresholds.

Realized success and held-out fit. The reach of Table 12 is the feasibility predicted by the affine fit. Table 8 counts, for the same requests, the fraction whose realized descriptor lies within ϵ of the target. At $\epsilon = 20\%$ the selected SLC realizes 97–100% of the requests on HalfCheetah, Ant, and Humanoid and only 75% on Franka. FSD, which receives the partition, ties SLC on HalfCheetah (100%) and approaches it on Ant (94%) but falls to 86% on Humanoid and 28% on Franka. Table 9 tests whether the $R^2_{\min}$ of the gate is in-sample overfitting, using 5-fold cross-validation (R^2 computed on the pooled out-of-fold predictions, minimum over descriptors, mean over seeds). The center-only row always outputs b and has geometry-only error $\rho\|u\|_1/(2K)$. At $K = 2$ this is at least $\rho/(2K) = 22.5\%$ for $\rho = 0.9$, so it cannot pass the 20% axis-mean threshold. At $K = 3$, its reported success is specific to the fixed request directions. The held-out $R^2_{\min}$ of the selected SLC is within 0.02 of the in-sample value and the gate verdict holds for every seed. No baseline newly passes the gate under held-out evaluation. CSF loses one passing seed on Ant and its passing seed on Humanoid.

Axis-mean success can nevertheless hide a failed descriptor. At $\rho = 0.9$ and $\epsilon = 20\%$, an axis-aligned request can pass after realizing only about 11% (one ninth) of its displacement for $K = 2$, or even no displacement for $K = 3$. We therefore report the 10% and every-axis criteria in Table 8.

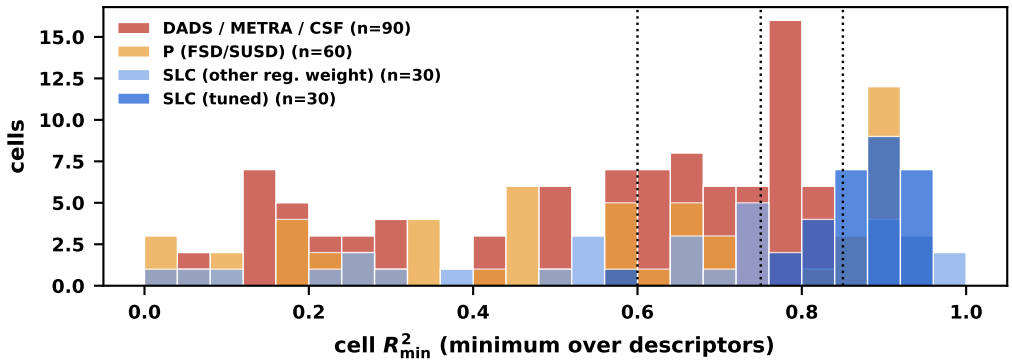

Figure 9: Distribution of cell-level R^2_{min} for the 210 learned-method cells behind Table 1, grouped by method; dotted lines mark $\tau = 0.60$, 0.75, and 0.85.

Table 7: Number of (seed, reset) cells whose minimum descriptor R^2 is at least τ, for every learned method and platform of Table 1 (9 cells per method, 3 on Franka). The main results use $\tau = 0.75$.

Platform	Method	τ=0.60	τ=0.65	τ=0.70	τ=0.75	τ=0.80	τ=0.85	τ=0.90
HalfCheetah	DADS	0/9	0/9	0/9	0/9	0/9	0/9	0/9
	METRA	9/9	9/9	6/9	0/9	0/9	0/9	0/9
	CSF	9/9	9/9	9/9	9/9	3/9	0/9	0/9
	SLC, $\lambda_\perp$=0†	8/9	8/9	8/9	8/9	8/9	8/9	8/9
	SLC, $\lambda_\perp$>0	9/9	9/9	9/9	9/9	9/9	9/9	8/9
	FSD	9/9	9/9	9/9	9/9	9/9	6/9	6/9
	SUSD	1/9	0/9	0/9	0/9	0/9	0/9	0/9
Ant	DADS	0/9	0/9	0/9	0/9	0/9	0/9	0/9
	METRA	3/9	0/9	0/9	0/9	0/9	0/9	0/9
	CSF	9/9	9/9	9/9	9/9	3/9	0/9	0/9
	SLC, $\lambda_\perp$=0	3/9	3/9	3/9	0/9	0/9	0/9	0/9
	SLC, $\lambda_\perp$>0†	9/9	9/9	9/9	9/9	9/9	9/9	3/9
	FSD	9/9	9/9	9/9	9/9	9/9	9/9	4/9
	SUSD	0/9	0/9	0/9	0/9	0/9	0/9	0/9
Humanoid	DADS	3/9	3/9	0/9	0/9	0/9	0/9	0/9
	METRA	3/9	0/9	0/9	0/9	0/9	0/9	0/9
	CSF	9/9	9/9	6/9	3/9	0/9	0/9	0/9
	SLC, $\lambda_\perp$=0	6/9	6/9	3/9	0/9	0/9	0/9	0/9
	SLC, $\lambda_\perp$>0†	9/9	9/9	9/9	9/9	9/9	6/9	0/9
	FSD	6/9	6/9	1/9	0/9	0/9	0/9	0/9
	SUSD	2/9	0/9	0/9	0/9	0/9	0/9	0/9
Franka	DADS	0/3	0/3	0/3	0/3	0/3	0/3	0/3
	METRA	2/3	2/3	2/3	0/3	0/3	0/3	0/3
	CSF	2/3	1/3	1/3	1/3	0/3	0/3	0/3
	SLC, $\lambda_\perp$=0	0/3	0/3	0/3	0/3	0/3	0/3	0/3
	SLC, $\lambda_\perp$>0†	3/3	3/3	3/3	3/3	1/3	0/3	0/3
	FSD	0/3	0/3	0/3	0/3	0/3	0/3	0/3
	SUSD	1/3	1/3	1/3	1/3	1/3	0/3	0/3

Thresholded success is restricted to the main comparison with $K \leq 3$. At $K = 6$, a center-only response passes the axis-mean 20% criterion for every unit request direction, so the K ladder reports continuous error and reach instead.

Table 8 evaluates 12 requests at $\rho = 0.9$ for each of three training seeds; the linear torque reference uses three random matrices, and the geometry-only center reference uses the common request directions once. Per-axis normalized error is $|\hat{D}_i(\alpha^*) - d_i^*|/(2h_i)$. The "mean" and "all" columns threshold the axis-mean error and every axis separately, respectively, while $|e|$ reports mean absolute

Table 8: Realized composition success at $\rho = 0.9$. "mean" thresholds axis-mean normalized error, "all" thresholds every axis, and $|e|$ is mean absolute joint-angle error. Gray learned rows are majority gate-invalid diagnostics.

Method	HalfCheetah					Ant					Humanoid					Franka				
	ϵ=10%		ϵ=20%		$\lvert e\rvert$	ϵ=10%		ϵ=20%		$\lvert e\rvert$	ϵ=10%		ϵ=20%		$\lvert e\rvert$	ϵ=10%		ϵ=20%		$\lvert e\rvert$
	mean	all	mean	all	(deg)	mean	all	mean	all	(deg)	mean	all	mean	all	(deg)	mean	all	mean	all	(deg)
Center-only (ref.)	0	0	$0_{\pm 0}$	0	14.0	0	0	$0_{\pm 0}$	0	11.7	0	0	$0_{\pm 0}$	0	43.0	0	0	$0_{\pm 0}$	0	52.0
Linear torque (ref.)	100	89	$100_{\pm 0}$	100	0.7	100	100	$100_{\pm 0}$	100	0.1	94	81	$100_{\pm 0}$	100	7.0	64	36	$97_{\pm 5}$	89	16.1
DADS	28	11	$58_{\pm 46}$	44	10.7	14	11	$53_{\pm 34}$	33	7.6	17	11	$53_{\pm 13}$	39	29.0	0	0	$14_{\pm 5}$	6	48.5
METRA	11	3	$64_{\pm 5}$	17	10.7	3	0	$42_{\pm 30}$	31	9.1	17	8	$47_{\pm 10}$	33	28.2	14	6	$53_{\pm 5}$	22	34.2
CSF	22	6	$64_{\pm 17}$	28	10.3	11	3	$89_{\pm 5}$	50	6.4	33	8	$75_{\pm 8}$	50	20.2	25	14	$44_{\pm 24}$	36	48.0
FSD	89	47	$100_{\pm 0}$	100	4.2	72	44	$94_{\pm 10}$	94	3.5	31	25	$86_{\pm 17}$	67	19.4	8	3	$28_{\pm 5}$	14	73.4
SUSD	8	3	$78_{\pm 10}$	14	10.2	0	0	$3_{\pm 5}$	0	10.8	17	8	$33_{\pm 22}$	25	31.6	22	8	$64_{\pm 34}$	47	34.5
SLC, $\lambda_\perp$=0	100	89	$100_{\pm 0}$	100	2.6	11	8	$44_{\pm 13}$	17	9.5	22	11	$75_{\pm 8}$	47	23.4	14	3	$58_{\pm 25}$	33	34.3
SLC, $\lambda_\perp$>0	81	58	$100_{\pm 0}$	100	3.9	75	61	$100_{\pm 0}$	100	2.7	72	58	$97_{\pm 5}$	94	11.0	25	17	$75_{\pm 8}$	53	28.1

Table 9: Five-fold held-out affine fit on the $N = 150$ reset-0 probes. Values are three-seed means of the minimum descriptor R^2; gate counts held-out passes at 0.75.

Method	HalfCheetah			Ant			Humanoid			Franka		
	in	held-out	gate	in	held-out	gate	in	held-out	gate	in	held-out	gate
Center-only (ref.)	–	–	–	–	–	–	–	–	–	–	–	–
Linear torque (ref.)	–	–	–	–	–	–	–	–	–	–	–	–
DADS	0.26	0.18	0/3	0.31	0.28	0/3	0.38	0.25	0/3	0.18	0.12	0/3
METRA	0.70	0.66	0/3	0.46	0.42	0/3	0.58	0.52	0/3	0.68	0.65	0/3
CSF	0.80	0.77	3/3	0.79	0.78	2/3	0.73	0.68	0/3	0.49	0.45	1/3
FSD	0.89	0.88	3/3	0.89	0.89	3/3	0.64	0.58	0/3	0.05	-0.02	0/3
SUSD	0.41	0.33	0/3	0.21	0.13	0/3	0.48	0.36	0/3	0.48	0.44	1/3
SLC, $\lambda_\perp$=0	0.94	0.94	3/3	0.36	0.31	0/3	0.64	0.57	0/3	0.38	0.33	0/3
SLC, $\lambda_\perp$>0	0.94	0.94	3/3	0.89	0.88	3/3	0.86	0.84	3/3	0.79	0.77	3/3

joint-angle error in degrees. Subscripts in the $\epsilon = 20\%$ mean column are between-seed standard deviations (across random matrices for the linear reference); other success columns are means. Majority gate-invalid learned rows are gray diagnostics. These are realized outcomes, unlike the predicted reach values in Tables 12 and 15.

A training-free linear torque reference. The control $a = M\alpha$ sends the skill directly to a constant actuator-torque command without learning (unit-norm rows, no intercept, three random M). It uses one reset and the common target scale. For SLC, iso, gap, $R^2_{\min}$, and valid repeat the main-table aggregates over nine cells (three on Franka), whereas error and reach use reset 0 for the same three training seeds. At spring-loaded joints, constant torque specifies a static pose, so descriptors are nearly linear in the skill. Table 10 gives iso in the hundreds, $R^2_{\min} = 1.00$, and 0.1% error on HalfCheetah and Ant. On Humanoid and Franka it has higher iso and lower gap than SLC. Its narrower pose range lowers reach to 78, 89, 61, and 33%. This characterizes the linear controllability already afforded by the static-pose body. With no exploration or feedback and only constant torques, it is not a controller for dynamic or contact-dominated regimes.

B.4 Isolation and Continuity

Isolation great circle. One circle is built per descriptor i. Under the full-row-rank gate, all evaluated configurations satisfy $B \geq K + 1$, so the subspace orthogonal to the $K - 1$ held gradients has dimension $B - K + 1 \geq 2$. Within this subspace, e_i is the projection of the moved gradient and f_i is orthogonal to all gradients. We roll out the $N_G = 180$ circle points (2° apart) and record all K descriptors. The numerator and denominator of iso come from the same circle, and the platform value is the minimum over axes. In the main comparison, all descriptors are joint angles in degrees, so the ratio has common units.

$$\alpha^{(i)}(\theta_g) = \cos\theta_g\, e_i + \sin\theta_g\, f_i, \qquad \theta_g = \frac{2\pi g}{N_G},\ g = 0, \ldots, N_G - 1,\ N_G = 180. \tag{15}$$

Substituting this circle into the fitted affine map gives

$$G\alpha^{(i)}(\theta) + b = b + \cos\theta\, Ge_i + \sin\theta\, Gf_i = b + \cos\theta\, Ge_i, \tag{16}$$

Table 10: Training-free linear torque reference versus selected SLC under the main protocol. Reference values are mean±sd over three random unit-row-norm matrices M.

	Linear torque reference $a{=}M\alpha$					SLC (tuned $\lambda_\perp$)				
Platform	iso↑	gap↓	$R^2_{\min}$ (valid)	err↓ (%)	reach (%)	iso↑	gap↓	$R^2_{\min}$ (valid)	err↓ (%)	reach (%)
HalfCheetah	381.4±57.8	1.1	1.00 (3/3)	0.1	78	4.1±0.8	3.6	0.90 (8/9)	3.5	100
Ant	144.5±27.7	1.1	1.00 (3/3)	0.1	89	2.1±0.3	6.7	0.89 (9/9)	6.5	100
Humanoid	8.0±2.2	1.2	0.96 (3/3)	3.3	61	2.7±0.8	6.3	0.86 (9/9)	7.4	100
Franka	3.7±1.4	1.4	0.93 (3/3)	7.5	33	1.6±0.3	25.1	0.79 (3/3)	16.3	100

because $f_i \in \ker G$. Moreover, $g_j^\top e_i = 0$ for every held descriptor $j \neq i$ by construction. The fitted response is therefore constant at b_j for each held descriptor and has the form $b_i + A_i \cos\theta$ for the moved descriptor, where $A_i = g_i^\top e_i$. A complete cosine plotted from 0 to 2π appears U-shaped even though the underlying map is affine in α. A flat cap or floor caused by a joint limit is instead a clipped sinusoid, approximately $\mathrm{clip}(b_i + A_i \cos\theta, \ell_i, u_i)$, and is not modeled by the main affine fit.

The gate R^2 is computed from the random sphere probes, not from these circle points. Conversely, iso and gap use new realized circle rollouts rather than the fitted predictions. Thus a high fit R^2 and visible saturation on the circle can coexist. Gap detects a large adjacent jump but does not explicitly penalize a long plateau; we therefore interpret it as a continuity diagnostic rather than a measure of nonzero sensitivity throughout the circle.

$$\mathrm{iso}_i = \frac{\mathrm{range}_g\, \hat{D}_i(\alpha^{(i)}(\theta_g))}{\max_{j\neq i} \mathrm{range}_g\, \hat{D}_j(\alpha^{(i)}(\theta_g)) + \varepsilon_0}, \qquad \mathrm{iso} = \min_i \mathrm{iso}_i. \tag{17}$$

Here $\varepsilon_0 = 10^{-9}$ degrees is a numerical denominator floor. Because iso compares absolute degree ranges rather than normalizing each descriptor by its own reachable range, a nearly stationary held descriptor can make the ratio large. We therefore read iso only with the fit gate and gap.

Continuity. maxgap is the largest jump between adjacent points on the circle divided by the range of that axis. The circle is closed, so the last point is also joined to the first. The reference $\mathrm{maxgap}_{\mathrm{ideal}}$ is that of a perfect sine sampled uniformly at N_G points, $\max_g \frac{1}{2}|\sin\theta_{(g+1) \bmod N_G} - \sin\theta_g| \approx 0.01745$. Table generation uses the four-decimal value 0.0174. It is not a lower bound (a triangle wave gives $2/N_G \approx 0.011$). At the other extreme, an ideal binary step has normalized maxgap 1, or about $57\times$ the sine reference. Thus the reported multiples are a graded diagnostic between smooth and abrupt responses rather than a binary classification. In the main comparison, the per-axis multiples are averaged into one value per (seed, reset) cell, then over gate-valid resets within each seed, and finally over contributing seeds. The K ladder instead reports the median over available resets, as stated in Table 14.

$$\begin{aligned}
\mathrm{maxgap}_{eri} &= \frac{\max_g \left|\hat{D}_i(\alpha^{(i)}(\theta_{(g+1) \bmod N_G})) - \hat{D}_i(\alpha^{(i)}(\theta_g))\right|}{\mathrm{range}_g\, \hat{D}_i(\alpha^{(i)}(\theta_g)) + \varepsilon_0}, \\
\gamma_{er} &= \frac{1}{K}\sum_i \frac{\mathrm{maxgap}_{eri}}{\mathrm{maxgap}_{\mathrm{ideal}}}, \\
\mathrm{gap} &= \frac{1}{|\mathcal{E}_v|}\sum_{e\in\mathcal{E}_v} \frac{1}{|\mathcal{R}_e^v|}\sum_{r\in\mathcal{R}_e^v} \gamma_{er}.
\end{aligned} \tag{18}$$

The subscripts e, r, i index the training seed, reset, and descriptor. Here $\mathcal{E}_v$ contains the seed indices e with at least one gate-valid reset, and $\mathcal{R}_e^v$ is the valid-reset set for seed e.

B.5 Composition and Null-Space Conventions

Composition accuracy. For each $\rho \in \{0.3, 0.5, 0.7, 0.9\}$, fixed seed 0 supplies an independently drawn set of $N_T = 12$ directions, giving every method the same 48 displacements $\rho(h \odot u)$ from its own fitted center b, not the same absolute target poses. The common scale h is the median half-range over the runs in that evaluation batch: 21 runs in the main comparison and five runs for each (platform, K) cell of the K ladder. After the Humanoid task-RNG correction, its h was recomputed over all 21 corrected runs, giving the full-range vector $2h = (145.732°, 145.764°)$. A common h

imposes the same absolute displacement. Normalizing by each method's own range would favor narrow movers. The radius ρ is measured in these half-range-normalized coordinates, not relative to the reachable-set boundary. Inner error averages the 24 requests with $\rho \leq 0.5$. Outer error averages the 12 requests at $\rho = 0.9$ and is reported separately in the full table. Reach uses those same outer requests. An inversion inside the sphere is filled through $\ker G$. One outside is normalized to the sphere, rolled out, and counted as unreached. Err is the axis-mean error normalized by the common full range $2h_i$, and composition uses reset 0. The affine map used to form G, b and the requested composition rollouts share reset 0. Transfer across reset realizations is not tested. Reach is predicted feasibility, not realized success. Parenthesized err at reach below 50% indicates that most outer requests were clipped. Succ independently counts axis-mean err $\leq \epsilon$, with the all-axis criterion also reported in Table 8.

Equation (8) defines the joint request and its minimum-norm inverse. The sampled targets and executed sphere skills are

$$d^* = b + \rho\,(h \odot u), \quad u \sim \mathrm{Unif}(\mathbb{S}^{K-1}), \quad \rho \in \{0.3, 0.5, 0.7, 0.9\}, \quad N_T = 12 \text{ per } \rho, \tag{19}$$

$$\alpha^* = \begin{cases} \alpha_{\mathrm{ls}} + \sqrt{1 - \|\alpha_{\mathrm{ls}}\|^2}\,\nu, & \|\alpha_{\mathrm{ls}}\| \leq 1 \;\; (\nu \in \ker G,\ \|\nu\| = 1), \\ \alpha_{\mathrm{ls}}/\|\alpha_{\mathrm{ls}}\|, & \text{otherwise (clipped, not reached).} \end{cases} \tag{20}$$

$$\mathrm{err} = \frac{1}{K}\sum_{i=1}^{K} \frac{|\hat{D}_i(\alpha^*) - d_i^*|}{2h_i}, \qquad \mathrm{reach} = \frac{1}{N_T}\sum_{u} \mathbb{1}\big[\|G^+(d^* - b)\| \leq 1\big]\Big|_{\rho=0.9}, \tag{21}$$

where $\mathbb{1}[\cdot]$ is the indicator function.

$$h = \mathrm{median}_{\text{runs in batch}}\ \tfrac{1}{2}\,\mathrm{range}_m\,\hat{D}(\alpha^{(m)}). \tag{22}$$

Null-space fill and its sensitivity. For the isolation circle, f_i is the first SVD vector of the subspace orthogonal to all gradients. This choice need not be equivariant to an orthogonal change of skill coordinates, so iso and gap are not strictly rotation-invariant when that null space has dimension at least two. We did not measure their sensitivity to f_i. Composition uses the first SVD vector ν of $\ker G$, with its sign fixed by making the largest-magnitude component positive. For the selected seed-0 SLC checkpoints, the first two available null fills change realized descriptors by about 5% of the common range $2h_i$ on HalfCheetah and Ant and 17–26% on Humanoid, depending on ρ. Aggregate error and success were not computed for the second Humanoid fill because the evaluator stores only the per-ρ mean absolute between-fill difference over requests and axes, not the second fill's realized descriptors. For the $B = 3$ Franka SLC row, whose null space is one-dimensional, the alternative sign was not evaluated. No null-fill sensitivity is reported for the $B = 4$ FSD and SUSD rows. Thus the reported composition results are conditional on the stated first-SVD and sign convention. Clipped requests have no null component and contribute zero to this deviation. The quoted selected SLC runs have no clipped requests.

For reproducibility, let Q_{-i} be the reduced-QR basis of the held-gradient rows. The moved axis is $e_i \propto (I - Q_{-i}Q_{-i}^\top)g_i$. For the second circle axis, Q_{all} is the reduced-QR basis of all gradient rows. The vector f_i is the first left-singular vector of $I - Q_{\mathrm{all}}Q_{\mathrm{all}}^\top$ with singular value above 10^{-6}, after subtracting its numerical component along e_i and renormalizing. The retained singular values are tied, so this ordering is numerically non-canonical. Reversing the sign of f_i only reparameterizes the same circle.

B.6 Seeds, Resets, and Aggregation

Seeds, resets, and aggregation. Training seeds are 0, 1, and 2. The main comparison uses resets 0, 1, and 2 (only reset 0 on Franka because its reset is RNG-independent), while single-seed ladders use up to three available resets and report the denominator explicitly. A reset is a repeated measurement of one policy, not an independent sample. In the main comparison, iso and gap are first averaged over gate-valid resets within each seed and then averaged across seeds that have a valid reset. Iso additionally reports between-seed sd when at least two seeds contribute. Gradient angles and condition numbers use gate-valid cells. R^2 averages all scheduled cells, valid counts the passing cells, realized err is a seed mean of reset-0 outcomes, and predicted reach is averaged over the same seed runs. Graying is a row-level majority rule. The reset-0 fits of all three selected SLC seeds

pass the gate. Gate-invalid rows retain their realized reset-0 composition values as gray diagnostics but are excluded from ranking. The single-seed K ladder reports its separate available-reset aggregation and gate denominator in its caption. Table 8 reports seed-level sd of success. Range ratios use a 10^{-9} denominator floor, whereas R^2 uses its ordinary denominator after the stated nonzero-variance guard. The main inventory contains 84 learned runs. The supplementary archive includes `experiment_manifest.md`, which maps run stems, evaluation batches, generator scripts, and historical/corrected Humanoid outputs.

B.7 Standard-Metric-Inspired Diagnostics

Relation to standard disentanglement metrics. DCI (Eastwood & Williams, 2018) characterizes whether latent coordinates specialize to individual factors (disentanglement), whether each factor is captured compactly (completeness), and how well the factors can be predicted from the latents (informativeness). MIG (Chen et al., 2018) measures, for each factor, the normalized gap between its two most informative latent coordinates. IRS (Suter et al., 2019) measures the robustness of inferred factors under interventions on other factors. These metrics address latent–factor correspondence or intervention robustness rather than whether a behavioral response varies continuously or realizes several targets together.

Table 11 reports them as an exploratory, seed-0 diagnostic computed without re-simulation from 150 random skills and the isolation-circle sweeps. Because MIG and DCI assume latent coordinates, grad uses the fitted coordinates $\hat{g}_i^\top \alpha$, where $\hat{g}_i = g_i/\|g_i\|$, and raw uses the original skill axes. Raw scores depend on arbitrary sphere rotation. DCI uses Lasso coefficients (regularization 0.02, standardized inputs) and a 30% hold-out, whereas MIG uses kNN(5) mutual information and 20-bin entropy. Our behavioral IRS analogue moves one descriptor along its isolation circle and measures robustness of the held descriptors. Grad coordinates are themselves defined by G, so results below the fit gate are not interpretable. The results expose this coordinate dependence: DCI-D and DCI-C are often near one in the fitted-gradient basis, while raw-axis DCI and MIG are usually lower and can reorder the methods. Fitted-basis MIG and IRS often identify the same strong isolation cells but do not give an identical ranking. FSD's high raw-axis DCI-D reflects its supplied partition. These diagnostics do not measure continuity or composition and are included as a check and a bridge to standard metrics, not as substitutes for the behavioral-control protocol.

B.8 Full Comparison Tables

Table 12 contains the standard deviations, R^2, gradient angles, condition numbers, and outer errors behind Table 1. It follows the aggregation of Appendix B.6: iso and gap first average valid resets within seed, $R^2_{\min}$ averages all scheduled cells, and composition uses reset 0. Parenthesized errors have reach below 50%; objective tags and graying follow Table 1. Table 13 gives the full $\lambda_\perp$ sweep, Tables 14 and 15 give the numbers behind Figure 4, and Table 16 reports the separate targeted multi-seed audit.

Orthogonality sweep. Table 13 reports the seed-0 sweep used to select $\lambda_\perp$ for each platform. Selection maximizes mean iso over the tested values without consulting the gate or composition; all selected points nevertheless pass the gate. Here $\angle_{\min}$ is the smallest angle between fitted descriptor gradients, and κ is the condition number of G. Reset variability is platform-dependent: Franka has an RNG-independent reset, while Ant varies little across resets. This sweep predates the final per-run environment recreation, so the Humanoid range includes an order-contaminated reset. The sweep and its selection of $\lambda_\perp = 1$ were not repeated, whereas the main Humanoid evaluation at that fixed value uses the corrected protocol.

Descriptor ladder manifest. Within each platform the $K = 2, 3, 4, 6$ sets are nested. HalfCheetah uses observation indices $\{6, 8\}$ (`fthigh`, `ffoot`), then adds 7 (`fshin`), 5 (`bfoot`), and finally 3,4 (`bthigh`, `bshin`). Ant uses $\{8, 12\}$ (`ankle_1`, `ankle_3`), then adds 10 (`ankle_2`), 14 (`ankle_4`), and finally 7,11 (`hip_1`, `hip_3`). Humanoid uses $\{15, 18\}$ (right/left `shoulder1`), then adds 17 (right elbow), 20 (left elbow), and finally 16,19 (right/left `shoulder2`). Franka uses $\{1, 3\}$ (joints 2 and 4), then adds 5 (joint 6), 0 (joint 1), and finally 2,4 (joints 3 and 5). Thus each step adds coordinates without replacing an earlier descriptor.

Table 11: Standard-metric-inspired diagnostics on the main-table data (seed 0, mean over resets). Fitted-coordinate DCI/MIG and IRS are protocol adaptations; gray rows are majority gate-invalid, and † marks selected $\lambda_\perp$.

		grad basis (K coords)				raw axes (B)				
Platform	Method	DCI-D	DCI-C	DCI-I	MIG	DCI-D	MIG	IRS	iso	R^2
HalfCheetah	SLC, $\lambda_\perp{=}0^\dagger$	1.00	1.00	0.96	0.62	0.10	0.07	0.80	4.89	0.95
	SLC, $\lambda_\perp{>}0$	1.00	1.00	0.96	0.61	0.09	0.04	0.77	3.63	0.96
	METRA	0.95	0.95	0.74	0.36	0.32	0.03	0.10	0.90	0.72
	CSF	0.95	0.95	0.77	0.30	0.08	0.03	0.17	0.88	0.79
	DADS	0.86	0.83	0.34	0.18	0.15	0.02	0.07	1.00	0.14
	FSD	0.99	0.99	0.90	0.50	0.80	0.27	0.84	5.23	0.90
	SUSD	0.91	0.91	0.70	0.08	0.18	0.04	0.33	0.53	0.56
Ant	SLC, $\lambda_\perp{=}0$	0.82	0.81	0.10	0.02	0.57	0.01	0.01	1.01	0.06
	SLC, $\lambda_\perp{>}0^\dagger$	0.97	0.97	0.93	0.53	0.18	0.24	0.58	2.37	0.92
	METRA	0.98	0.98	0.61	0.25	0.08	0.06	0.01	1.01	0.61
	CSF	1.00	1.00	0.79	0.76	0.20	0.14	0.47	1.90	0.80
	DADS	0.68	0.66	0.52	0.07	0.19	0.08	0.24	1.05	0.27
	FSD	1.00	1.00	0.89	0.42	0.96	0.10	0.70	10.19	0.89
	SUSD	0.93	0.92	0.36	0.08	0.04	0.08	0.02	0.73	0.30
Humanoid	SLC, $\lambda_\perp{=}0$	1.00	1.00	0.74	0.29	0.23	0.08	0.14	1.16	0.72
	SLC, $\lambda_\perp{>}0^\dagger$	0.96	0.96	0.88	0.38	0.15	0.07	0.70	3.46	0.88
	METRA	0.94	0.94	0.62	0.21	0.11	0.01	0.05	1.06	0.51
	CSF	0.80	0.81	0.66	0.20	0.14	0.07	0.00	0.98	0.66
	DADS	0.69	0.69	0.22	0.05	0.23	0.02	0.02	0.59	0.19
	FSD	0.84	0.85	0.73	0.36	0.46	0.16	0.37	1.74	0.66
	SUSD	0.97	0.97	0.61	0.23	0.15	0.04	0.31	1.83	0.37
Franka	SLC, $\lambda_\perp{=}0$	0.65	0.64	0.38	0.13	0.13	0.02	0.04	1.11	0.27
	SLC, $\lambda_\perp{>}0^\dagger$	0.95	0.95	0.80	0.24	0.00	0.11	0.41	1.90	0.81
	METRA	1.00	1.00	0.77	0.11	0.08	0.27	0.00	0.76	0.74
	CSF	0.88	0.88	0.82	0.34	0.78	0.22	0.49	0.57	0.80
	DADS	0.76	0.76	0.20	0.01	0.17	0.20	0.00	0.91	0.16
	FSD	1.00	1.00	0.17	0.04	0.13	0.04	0.00	0.64	0.10
	SUSD	1.00	1.00	0.80	0.27	0.31	0.15	0.15	1.00	0.80

Ladder reporting. All ladders use $B = 8$, seed 0, and fixed physics within each platform. Relative to the main comparison, Humanoid unlocks its elbows with stiffness 32 and damping 5.66; its rows use per-run environment recreation and fixed task RNG. Table 14 reports mean iso over up to three available resets, the valid-reset count, and median gap. A majority gate failure makes iso and gap diagnostic. Table 15 uses seed 0 and reset 0 and reports inner error / predicted reach. Its positive-SLC values are $\lambda_\perp = 1$ on HalfCheetah, Ant, and Humanoid and 10^2 on Franka. For each platform and K, the common h is the median half-range over these five ladder runs rather than the 21-run main batch, so errors are not comparable to Table 1; parentheses mark reach below 50%.

The targeted audit in Table 16 selects four cells where seed-0 DADS was competitive on at least one metric or positive-$\lambda_\perp$ SLC was most stressed. This three-seed SLC–DADS audit is separate from the all-method single-seed ladder. Entries are ordered by seeds 0/1/2; valid gives the number of three resets passing the gate, while $\mathrm{err}_{\mathrm{in}}$ and reach use reset 0 and the fixed common h from the five-method ladder.

Table 12: Full metrics behind Table 1, including $R^2_{\min}$, gradient geometry, outer composition error, and predicted reach.

Platform	Method	Objective	iso	R^2	valid	$\angle_{\min}$	κ	gap	err. inner / outer (%)	reach
HalfCheetah	DADS	MI	–	0.27	0/9	–	–	–	(11.4±4.1 / 17.9)	0%
	METRA	T-Dist.	–	0.70	0/9	–	–	–	21.5±6.3 / 17.4	81%
	CSF	NCE	0.95±0.14	0.80	9/9	24°	6.96	8×	24.1±2.5 / 16.9	50%
	FSD (adapt.)	F	5.08±0.39	0.89	9/9	83°	2.34	5×	6.0±0.9 / 6.7	75%
	SUSD (port)	F	–	0.40	0/9	–	–	–	(8.0±2.4 / 17.0)	0%
	SLC, $\lambda_\perp{=}0^\dagger$	MI	4.09±0.76	0.90	8/9	84°	1.54	4×	3.5±0.8 / 4.2	100%
	SLC, $\lambda_\perp{>}0$	MI	2.81±1.11	0.94	9/9	85°	1.37	9×	4.8±0.7 / 6.3	100%
Ant	DADS	MI	–	0.29	0/9	–	–	–	(13.7±1.5 / 19.3)	6%
	METRA	T-Dist.	–	0.45	0/9	–	–	–	34.2±6.8 / 23.0	67%
	CSF	NCE	1.59±0.30	0.79	9/9	72°	1.45	11×	23.2±3.0 / 16.1	100%
	FSD (adapt.)	F	6.93±4.50	0.89	9/9	88°	1.14	14×	13.0±2.0 / 8.8	100%
	SUSD (port)	F	–	0.19	0/9	–	–	–	(11.3±1.1 / 27.4)	0%
	SLC, $\lambda_\perp{=}0$	MI	–	0.35	0/9	–	–	–	(26.8±5.9 / 24.0)	33%
	SLC, $\lambda_\perp{>}0^\dagger$	MI	2.14±0.25	0.89	9/9	83°	1.20	7×	6.5±2.4 / 6.9	100%
Humanoid	DADS	MI	–	0.38	0/9	–	–	–	(14.5±2.5 / 19.9)	0%
	METRA	T-Dist.	–	0.57	0/9	–	–	–	24.3±5.7 / 19.4	100%
	CSF	NCE	1.89	0.73	3/9	35°	3.26	12×	22.1±4.7 / 13.9	78%
	FSD (adapt.)	F	–	0.65	0/9	–	–	–	10.9±1.6 / 13.3	100%
	SUSD (port)	F	–	0.49	0/9	–	–	–	(11.9±5.7 / 21.7)	11%
	SLC, $\lambda_\perp{=}0$	MI	–	0.64	0/9	–	–	–	29.4±3.8 / 16.1	100%
	SLC, $\lambda_\perp{>}0^\dagger$	MI	2.70±0.79	0.86	9/9	86°	1.10	6×	7.4±5.8 / 7.6	100%
Franka	DADS	MI	–	0.18	0/3	–	–	–	(13.8±0.4 / 27.6)	0%
	METRA	T-Dist.	–	0.68	0/3	–	–	–	25.5±6.0 / 18.8	50%
	CSF	NCE	0.57	0.50	1/3	85°	4.25	17×	(29.3±14.3 / 28.0)	25%
	FSD (adapt.)	F	–	0.05	0/3	–	–	–	(39.3±2.0 / 39.4)	0%
	SUSD (port)	F	1.00	0.48	1/3	47°	2.92	19×	(15.3±4.4 / 19.3)	28%
	SLC, $\lambda_\perp{=}0$	MI	–	0.38	0/3	–	–	–	(15.7±7.3 / 18.8)	25%
	SLC, $\lambda_\perp{>}0^\dagger$	MI	1.58±0.32	0.79	3/3	83°	1.45	25×	16.3±7.5 / 15.2	100%

Table 13: Full per-platform sweep of $\lambda_\perp$. Iso is mean [min, max] over three resets. Below-gate iso, angle, and condition-number entries are gray diagnostics; † marks the selected value.

Platform	$\lambda_\perp$	iso [min, max]	R^2	$\angle_{\min}$	κ
HalfCheetah	$0^\dagger$	4.89 [4.42, 5.25]	0.95	89°	1.35
	10^{-2}	3.82 [2.48, 4.54]	0.95	87°	1.39
	10^{-1}	3.01 [2.53, 3.65]	0.95	87°	1.37
	10^{0}	3.63 [3.12, 4.03]	0.96	87°	1.40
	10^{1}	3.26 [1.59, 5.19]	0.90	84°	1.54
	10^{2}	1.58 [1.30, 2.06]	0.87	84°	1.42
Ant	0	1.01 [1.00, 1.01]	0.06	80°	2.17
	10^{-1}	2.04 [1.92, 2.13]	0.92	89°	1.15
	$10^{0\dagger}$	2.36 [2.27, 2.45]	0.92	86°	1.24
	10^{1}	0.94 [0.93, 0.94]	0.76	83°	1.13
	10^{2}	0.49 [0.47, 0.51]	0.27	15°	25.54
Humanoid	0	1.12 [1.00, 1.18]	0.72	78°	1.24
	10^{-2}	2.04 [1.65, 2.83]	0.83	82°	1.16
	$10^{0\dagger}$	3.09 [2.28, 4.12]	0.88	87°	1.09
	10^{1}	1.73 [1.26, 2.10]	0.85	86°	1.10
	10^{2}	2.28 [1.32, 4.17]	0.87	80°	1.19
Franka	0	1.11 [1.11, 1.11]	0.27	80°	1.50
	10^{-1}	1.33 [1.33, 1.33]	0.52	78°	1.45
	10^{0}	0.92 [0.92, 0.92]	0.39	84°	1.67
	10^{1}	1.18 [1.18, 1.18]	0.62	89°	1.36
	$10^{2\dagger}$	1.90 [1.90, 1.90]	0.81	83°	1.43
	10^{3}	1.39 [1.39, 1.39]	0.89	75°	1.47

Table 14: Descriptor-count stress test at $B = 8$ (seed 0): mean iso, valid-reset count, and median gap. Gray entries are majority gate-invalid; values underlie Figure 4.

		DADS		METRA		CSF		SLC, $\lambda_\perp{=}0$		SLC, $\lambda_\perp{>}0$	
Platform	K	iso (valid)	gap	iso (valid)	gap	iso (valid)	gap	iso (valid)	gap	iso (valid)	gap
HalfCheetah	2	1.85 (3/3)	2	1.12 (3/3)	17	0.93 (3/3)	15	3.29 (0/3)	28	4.05 (3/3)	11
	3	0.60 (0/3)	36	0.91 (0/3)	28	0.83 (3/3)	8	0.87 (0/3)	17	2.78 (3/3)	8
	4	1.45 (0/3)	10	1.09 (0/3)	12	0.60 (3/3)	2	1.79 (3/3)	9	1.54 (3/3)	5
	6	2.23 (3/3)	4	1.07 (0/3)	5	0.64 (0/3)	2	1.16 (3/3)	10	1.73 (3/3)	9
Ant	2	1.76 (0/3)	3	1.11 (0/3)	35	1.81 (3/3)	10	1.04 (0/3)	40	4.52 (3/3)	5
	3	3.26 (3/3)	2	1.03 (0/3)	44	1.41 (0/3)	5	1.01 (0/3)	44	2.03 (3/3)	4
	4	3.89 (3/3)	2	1.14 (0/3)	13	1.26 (3/3)	8	1.00 (0/3)	39	1.94 (3/3)	3
	6	0.55 (2/3)	10	0.61 (0/3)	32	0.10 (1/3)	15	0.63 (0/3)	44	0.68 (3/3)	6
Humanoid	2	0.67 (0/3)	31	1.00 (0/3)	52	1.20 (0/3)	31	1.52 (0/3)	13	2.06 (3/3)	3
	3	1.46 (0/3)	14	0.97 (0/3)	52	1.04 (0/3)	50	0.98 (0/3)	13	2.13 (3/3)	3
	4	0.24 (0/3)	22	0.96 (0/3)	43	0.93 (0/3)	39	1.00 (0/3)	25	1.64 (3/3)	4
	6	0.43 (0/3)	13	0.88 (0/3)	40	0.70 (0/3)	38	1.01 (0/3)	22	1.27 (3/3)	5
Franka	2	3.31 (1/1)	1	0.75 (0/1)	39	0.76 (0/1)	30	0.70 (0/1)	8	1.07 (1/1)	18
	3	0.18 (0/1)	18	0.29 (0/1)	33	0.11 (0/1)	47	0.16 (0/1)	18	0.24 (0/1)	25
	4	0.23 (0/1)	28	0.16 (0/1)	42	0.07 (0/1)	35	0.20 (0/1)	6	0.15 (0/1)	19
	6	0.20 (0/1)	24	0.14 (0/1)	16	0.12 (0/1)	26	0.18 (0/1)	28	0.21 (1/1)	4

Table 15: Joint composition on the K ladder at $B = 8$ (seed 0, reset 0): inner error / predicted reach. Parentheses mark reach below 50%.

Platform	K	DADS	METRA	CSF	SLC, $\lambda_\perp{=}0$	SLC, $\lambda_\perp{>}0$
HalfCheetah	2	(5.8) / 0	27.4 / 75	18.6 / 100	2.0 / 100	5.3 / 100
	3	(8.7) / 0	(22.5) / 42	22.5 / 58	7.7 / 100	16.0 / 100
	4	(4.7) / 33	18.7 / 83	(15.9) / 8	9.6 / 100	3.3 / 100
	6	3.9 / 58	(19.9) / 33	(14.9) / 8	21.1 / 100	3.8 / 100
Ant	2	(8.0) / 0	25.6 / 100	15.3 / 100	23.7 / 100	6.0 / 100
	3	3.4 / 50	24.0 / 100	16.7 / 50	19.9 / 75	7.2 / 100
	4	(3.0) / 8	22.9 / 50	24.4 / 50	22.5 / 67	12.6 / 100
	6	(6.1) / 0	31.6 / 75	(18.1) / 0	34.7 / 83	12.0 / 100
Humanoid	2	(12.5) / 0	33.6 / 100	17.9 / 100	28.8 / 100	13.2 / 100
	3	(11.1) / 0	39.9 / 100	30.6 / 50	32.7 / 100	16.1 / 100
	4	(9.6) / 0	38.0 / 100	(30.7) / 17	30.3 / 100	17.9 / 100
	6	(7.5) / 0	(32.0) / 42	(26.4) / 0	17.8 / 100	19.4 / 100
Franka	2	(8.6) / 0	29.1 / 100	26.2 / 100	7.2 / 100	15.4 / 100
	3	(9.1) / 8	29.4 / 67	(28.1) / 8	23.0 / 100	18.3 / 100
	4	(9.7) / 0	38.5 / 67	(24.4) / 17	10.3 / 83	20.1 / 75
	6	(6.9) / 0	(16.6) / 0	(22.2) / 0	27.9 / 100	(14.0) / 8

Table 16: Targeted three-seed SLC–DADS audit of four difficult K-ladder cells. Entries are ordered by seeds 0/1/2.

Platform	K	Method	valid resets	$\mathrm{err}_{\mathrm{in}}$ (%)	reach (%)
HalfCheetah	6	SLC, $\lambda_\perp{=}1$	3/3/3	3.8/9.1/5.6	100/100/100
	6	DADS	3/2/0	3.9/2.7/4.0	58/33/67
Ant	3	SLC, $\lambda_\perp{=}1$	3/3/3	7.2/5.1/6.6	100/100/100
	3	DADS	3/3/3	3.4/3.1/3.4	50/33/0
	4	SLC, $\lambda_\perp{=}1$	3/3/3	12.6/11.1/4.2	100/100/100
	4	DADS	3/0/1	3.0/4.4/3.3	8/42/50
Humanoid	6	SLC, $\lambda_\perp{=}1$	3/3/0	19.4/10.4/12.2	100/100/100
	6	DADS	0/0/0	7.5/14.7/12.4	0/0/0

C Additional Ablations and Scope

Table 17 varies B with K fixed and is distinct from the K ladder of Section 5.3. It collects independently trained historical checks rather than one common cross-platform batch: upright HalfCheetah SLC averages three resets, whereas the separate supine HalfCheetah METRA and Ant rows use one reset. The table therefore tests whether larger B helps monotonically within each listed setting, not a common scaling law. Figures 10 and 11 show main-run curves on method-specific scales. Selected SLC, FSD, METRA, and CSF are flat late in training. The DADS intrinsic reward is still rising slowly at epoch 8000, so the fixed-budget comparison does not establish its convergence. Unregularized SLC is unstable on Ant and Humanoid, whereas the selected $\lambda_\perp$ keeps critic loss bounded on all platforms. Table 18 decomposes the basis embeddings parallel and orthogonal to w.

Table 17: Historical skill-dimension checks at fixed K: iso, mean $R^2_{\min}$, and valid-reset count. HalfCheetah SLC reports mean±sd over three resets; other rows use one.

Platform	Method	B	iso	R^2	valid
HalfCheetah (upright), K=3	SLC, $\lambda_\perp$=0	6	3.04±0.12	0.89	3/3
	SLC, $\lambda_\perp$=1	6	4.22±0.79	0.93	3/3
	SLC, $\lambda_\perp$=0	10	1.46±0.06	0.70	0/3
	SLC, $\lambda_\perp$=1	10	1.07±0.11	0.67	0/3
	SLC, $\lambda_\perp$=0	16	1.01±0.10	0.74	1/3
	SLC, $\lambda_\perp$=1	16	1.13±0.41	0.80	3/3
HalfCheetah, K=3	METRA	4	0.70	0.71	0/1
	METRA	6	0.77	0.73	0/1
	METRA	8	0.76	0.58	0/1
	METRA	10	0.93	0.55	0/1
Ant, K=2	SLC, $\lambda_\perp$=1	4	2.45	0.92	1/1
	SLC, $\lambda_\perp$=1	8	2.90	0.93	1/1
	METRA	4	1.01	0.61	0/1
	METRA	8	1.12	0.75	1/1
	METRA+SF	4	1.00	0.42	0/1
	METRA+SF	8	1.05	0.66	0/1
	DADS	4	1.11	0.31	0/1
	DADS	8	1.95	0.74	0/1

Basis-decomposition diagnostic. We draw $N_B = 1024$ transitions from the replay buffer, split each basis embedding into the components parallel and perpendicular to the unit projection vector $\hat{w}$, and average norms and angles over bases and samples:

$$\hat{w} = \frac{w}{\|w\|}, \qquad \|\psi^{\parallel}\| = \mathbb{E}\big|\langle \psi_k, \hat{w}\rangle\big|, \qquad \|\psi^{\perp}\| = \mathbb{E}\big\|\psi_k - \langle \psi_k, \hat{w}\rangle \hat{w}\big\|, \tag{23}$$

$$\angle(\psi, w) = \mathbb{E}\arccos\frac{|\langle \psi_k, \hat{w}\rangle|}{\|\psi_k\|}, \qquad N_B = 1024. \tag{24}$$

For unregularized HalfCheetah, the measured component parallel to w grows by $46.5/25.6 = 1.82\times$ from epoch 1000 to 8000, whereas the orthogonal component grows by about $1.05\times$ using the unrounded values ($5.167/4.901$). Ant and Humanoid instead reach parallel magnitudes of order 10^7 and angles below 1.5°. Franka shows no such growth, although the separate seed-0 selection evaluation is still gate-invalid ($R^2 = 0.27$; Table 13). These distinct failure modes motivate reporting, rather than equating, the platform-specific selected weights $0, 1, 1, 100$.

Relation to Fisher orthogonality. The original motivation for this penalty was to remove the off-diagonal entries of the Fisher information matrix. For the single-head Boltzmann family $\{\pi^*_\alpha\}$ defined by Eq. (11), $\partial \log \pi^*_\alpha / \partial \alpha_i = \beta\,(c_i - \mathbb{E}_{\pi^*_\alpha}[c_i])$, so for a fixed state distribution μ and with α

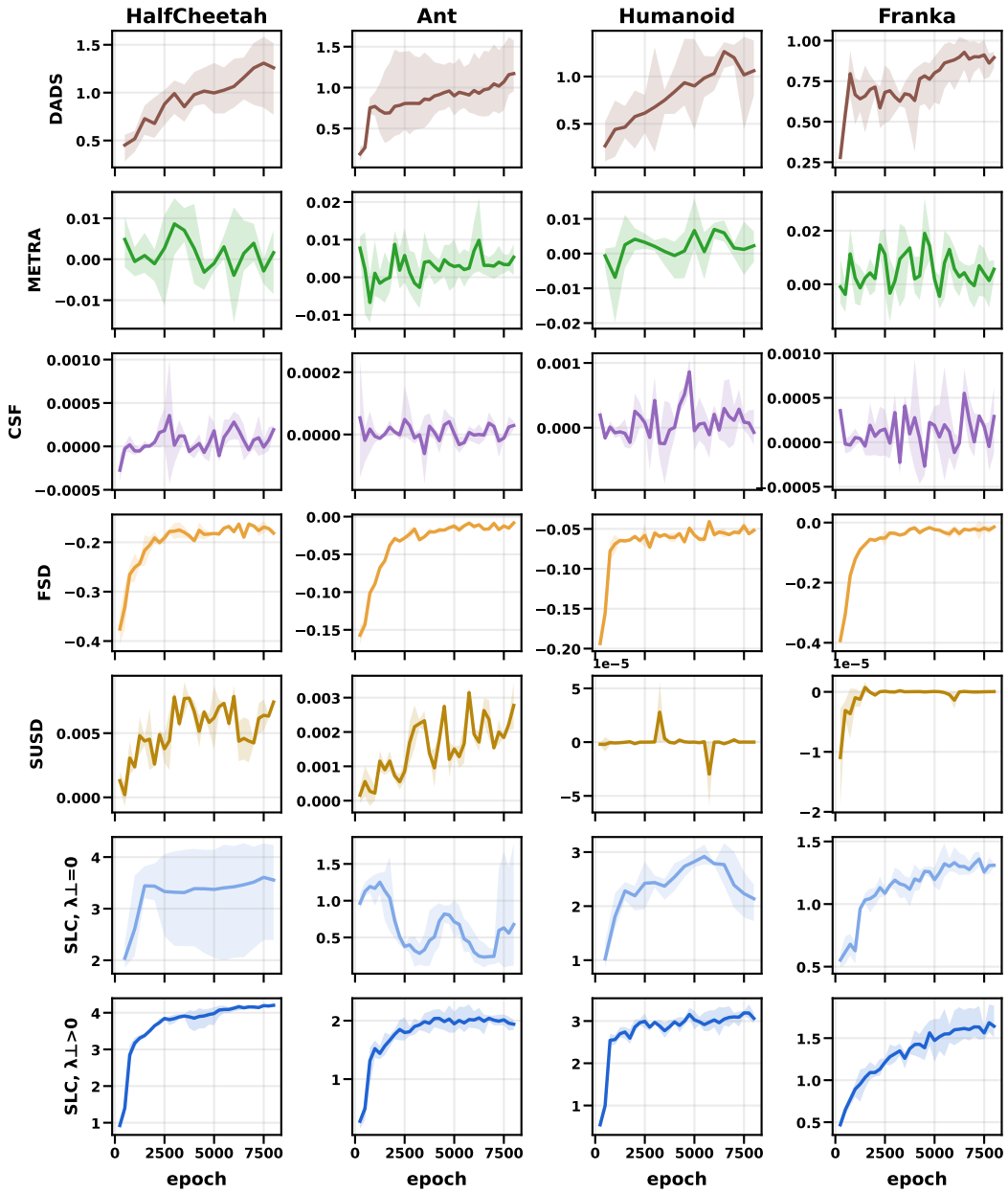

Figure 10: Intrinsic reward during the main training runs (rows: methods; columns: platforms; mean and min–max band over three seeds, one point per 250 epochs). Because skill-discovery objectives and vertical scales differ across panels, the curves diagnose within-method training progress rather than compare reward magnitude across methods.

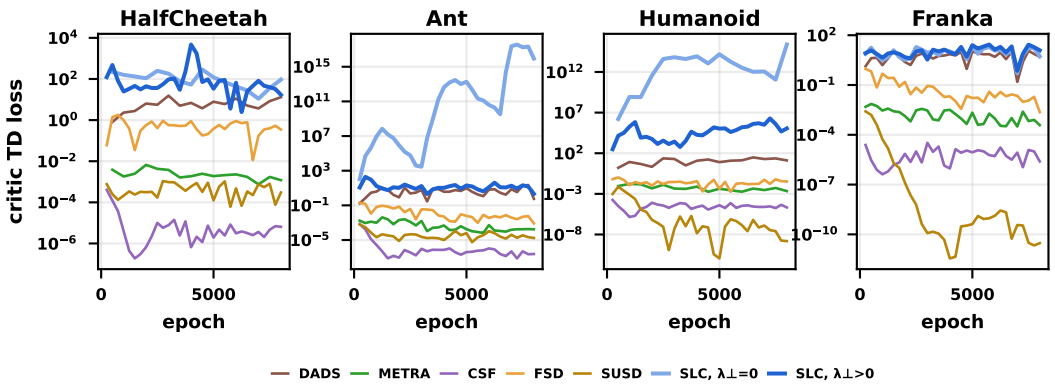

Figure 11: Critic TD loss during the main training runs (columns: platforms; curves: methods; log scale, mean over three seeds). The plot diagnoses optimization stability. Absolute loss magnitudes are not directly comparable across objectives or critic parameterizations.

Table 18: Basis embeddings decomposed parallel and orthogonal to w on the main runs (seed 1, epoch 8000 unless noted, 1024 replay transitions). ratio is $\|\psi^{\parallel}\|/\|\psi^{\perp}\|$, and the final column is the angle to w.

Platform	$\lambda_\perp$	$\|\psi^{\parallel}\|$	$\|\psi^{\perp}\|$	ratio	$\angle(\psi, w)$
HalfCheetah (ep 1000)	0	25.6	4.9	5.2	16.7°
HalfCheetah	0	46.5	5.17	9.0	7.6°
HalfCheetah	1	12.2	27.4	0.4	66.3°
Ant	0	8.65×10^{7}	1.81×10^{6}	47.8	1.2°
Ant	1	4.78	7.91	0.6	60.6°
Humanoid	0	1.24×10^{7}	5.64×10^{4}	220.4	0.8°
Humanoid	1	398	1066	0.4	70.0°
Franka	0	32.1	3.62	8.9	7.3°
Franka	100	6.97	10.4	0.7	56.6°

read in ambient coordinates, the Fisher information of the conditional policy is

$$\begin{aligned} F_{ij}(\alpha) &= \beta^2\,\mathbb{E}_{s\sim\mu}\Big[\mathrm{Cov}_{a\sim\pi^*_\alpha}\big(c_i(s,a), c_j(s,a)\big)\Big] \\ &= \beta^2\,\mathbb{E}_{s\sim\mu}\Big[w^\top \mathrm{Cov}_{a\sim\pi^*_\alpha}\big(\psi_i(s,a), \psi_j(s,a)\big)\,w\Big]. \end{aligned} \tag{25}$$

The off-diagonal entries are therefore the per-state, on-policy covariances between the basis-critic values c_k. This is not the Fisher information of whole trajectories, and since α lies on the unit sphere the intrinsic Fisher information would further require a projection onto the tangent space.

The penalty of Eq. (4) differs from this quantity in three ways. First, it orthogonalizes ψ before the projection onto w, which is neither sufficient nor necessary for removing the covariance of c_i and c_j: the centered $\tilde{\psi}_i = (x, 0)$ and $\tilde{\psi}_j = (0, x)$ are always orthogonal yet $\mathrm{Cov}(c_i, c_j) = w_1 w_2\,\mathrm{Var}(x)$, whereas $\tilde{\psi}_i = (x, y)$, $\tilde{\psi}_j = (z, y)$ with $w = (1, 0)$ and independent x, z give $\mathrm{Cov}(c_i, c_j) = 0$ while $\mathbb{E}\langle\tilde{\psi}_i, \tilde{\psi}_j\rangle = \mathrm{Var}(y) > 0$, so the penalty of Eq. (4) is also positive, $\mathbb{E}[\langle\tilde{\psi}_i, \tilde{\psi}_j\rangle^2] \geq (\mathbb{E}\langle\tilde{\psi}_i, \tilde{\psi}_j\rangle)^2 = \mathrm{Var}(y)^2 > 0$. Second, the expectation is taken over the replay buffer rather than under π^*_α. Third, it does not depend on α. The penalty is thus not a reduction of Fisher orthogonality but a surrogate that penalizes a different quantity, and we do not justify it by Fisher orthogonality.

The empirical proxy closest to Eq. (25) is the c-level correlation penalty, which penalizes the marginal correlation of c over replay transitions (not per state, not on-policy, not conditioned on α, and a normalized correlation rather than a covariance). As Table 19 shows, it lowers $|\mathrm{corr}(c_i, c_j)|$

from 0.155 to 0.008 and empirically suppresses $\|\tilde{\psi}\|$ in this run, but it does not orthogonalize the bases (27.4°) and it lowers the effective rank of c (3.23 of 4). Since even the diverged baseline has an effective rank of 3.75 of 4 for the replay c matrix, there was no rank collapse as far as this statistic shows. Accordingly, we describe the observed role of regularization as suppressing embedding growth rather than as guaranteeing orthogonalization or bounded norms. The right-hand columns, which re-measure the same checkpoints under the protocol of Section 4, point the same way: of the two c-level penalties that remove the marginal correlation of c equally well, OLE passes the gate (iso 1.46) and the correlation penalty fails it (R^2 0.07), so in this single-seed comparison lowering the replay correlation of c alone did not separate control success from failure.

For completeness, let $C \in \mathbb{R}^{M\times B}$ contain minibatch-centered scalar values c_k in its columns. The correlation control uses $\mathcal{L}_{\text{corr}} = \frac{1}{2}\|\operatorname{offdiag}(\bar{C}^\top \bar{C})\|_F^2$, where each column of $\bar{C}$ has unit Euclidean norm. The OLE control uses $\mathcal{L}_{\text{OLE}} = \sum_k \max(1, \|C_{:k}\|_2) - \|C\|_*$, with nuclear norm $\|\cdot\|_*$. Both Table 19 runs use loss weight 1, $B = 4$, seed 0, and no ψ-level penalty. The table compares critic statistics from 4096 replay transitions with one-reset control metrics for the same epoch-8000 checkpoints. Its angle is RMS-cosine, $\arccos\sqrt{\mathbb{E}\cos^2}$, and erank is the exponential entropy of the normalized squared singular values of centered c. These historical checkpoints predate spring-centered ankle initialization and evaluate all rollouts rather than filtering to upright ones, so they are not comparable to Table 1. Across them, all three penalties reduce $\|\tilde{\psi}\|$ from 10^7 to 10^1–10^2; only the ψ penalty makes the bases nearly orthogonal, while OLE passes the gate without doing so and the correlation penalty fails.

Table 19: Penalty location and effect in the historical Ant check ($B = 4$, seed 0, epoch 8000). Left: critic statistics; right: one-reset control metrics. Gray entries fail the gate.

Penalty	$\angle(\tilde{\psi}_i, \tilde{\psi}_j)$	$\|\tilde{\psi}\|$	$\|w\|$	$\lvert\text{corr}(c_i, c_j)\rvert$	erank$(c)/B$	iso	R^2_{min}	gap
none ($\lambda_\perp$=0, diverged)	0.3°	5.31e+07	42.74	0.155	3.75/4	1.00	0.64	46
ψ-level pairwise (ours, $\lambda_\perp$=1)	89.8°	8.84	6.54	0.149	3.77/4	1.37	0.79	11
c-level OLE (nuclear-norm gap)	28.2°	103	0.75	0.006	3.65/4	1.46	0.80	10
c-level correlation (scale-invariant)	27.4°	12.4	1.42	0.008	3.23/4	0.64	0.07	45

Magnitude-only and scalar-embedding controls. Table 20 separates magnitude control from a multidimensional embedding at seed 0. The magnitude-only variant uses $\mathcal{L}_{\text{mag}} = \mathbb{E}_{\mathcal{B},k}\|\tilde{\psi}_k\|_2^2$ without regularizing w, with weights $10^{-3}, 10^{-1}, 10^1$ on Ant and Humanoid and $10^{-1}, 10^1$ on Franka. The scalar variant sets $n = 1$ and retains the selected $\lambda_\perp$. R^2_{min} is the mean of the per-reset minima; isolation metrics average gate-valid evaluations among three resets, except for one reset on Franka. Humanoid uses corrected per-run recreation and fixed task RNG rather than the historical selection sweep. Composition uses reset 0 and the final main-table ranges. Min range is the smallest realized single-axis range relative to its common range and exposes nearly frozen policies.

A nondegenerate magnitude-only weight passes every reset gate on Ant, Humanoid, and Franka, but does not recover the selected behavior uniformly: Ant retains a 32× gap and 67% success, and Humanoid has 20.1% inner error. At weight 10, Ant and Franka move through only 0.0% and 3.1% of the common range and reach no outer request, so their high R^2 and iso reflect nearly frozen responses. The scalar embedding is gate-valid only on HalfCheetah, where iso falls from 4.89 to 1.96. These controls show that the pairwise penalty is not necessary for gate validity and that magnitude control explains part of its effect, not that either ingredient is sufficient across platforms.

Number of basis critics B. Increasing B does not improve the independently trained runs of Table 17 monotonically. On HalfCheetah trained upright, the smallest tested $B \in \{6, 10, 16\}$, $B = 6$, is best (iso 4.22 at $\lambda_\perp > 0$), $B = 10$ fails the gate in all three resets, and $B = 16$ has an iso near 1. On Ant, increasing B from 4 to 8 raises the iso of SLC from 2.45 to 2.90. METRA reaches the gate at $B = 8$ with only weak isolation (iso 1.12), while DADS and METRA+SF remain below it. These Ant values come from the historical B-dimension check rather than the later K-ladder batch.

Zero-intercept ablation. Table 21 tests whether the pure linear dependence on skill matters by adding a skill-independent critic head $q_0(s, a)$. This removes the antisymmetry $Q_{-\alpha} = -Q_\alpha$ while matching the intrinsic reward, platform, and selected $\lambda_\perp$; $\mathcal{L}_\perp$ applies only to the basis embeddings, not q_0. The comparison pairs seeds 0–2 at reset 0 and reports mean±sd. Valid counts seeds passing

Table 20: Magnitude-only and scalar-embedding controls (seed 0). Composition uses reset 0; min range measures the smallest axis range relative to the common range. Gray entries fail the gate.

Platform	Variant	valid	$R^2_{\min}$	iso	gap	min range (%)	$\mathrm{err}_{\mathrm{in}}$ (%)	succ / reach (%)
HalfCheetah	selected $\lambda_\perp$=0, n=64	3/3	0.95	4.89	3.5	100.0	2.6	100 / 100
	scalar n=1	3/3	0.91	1.96	5.4	97.1	6.7	100 / 100
Ant	selected $\lambda_\perp$=1	3/3	0.92	2.37	6.4	99.1	3.9	100 / 100
	no penalty	0/3	0.06	1.01	–	99.6	24.7	33 / 0
	magnitude 10^{-3}	0/3	0.70	1.00	–	100.9	34.8	42 / 100
	magnitude 10^{-1}	3/3	0.78	1.20	32.3	99.7	10.6	67 / 75
	magnitude 10^{1}	3/3	1.00	40.90	1.3	0.0	13.2	0 / 0
	scalar n=1	0/3	0.69	0.93	–	89.7	29.1	33 / 0
Humanoid	selected $\lambda_\perp$=1	3/3	0.88	3.46	3.0	99.9	14.0	92 / 100
	no penalty	0/3	0.72	1.16	–	100.2	32.4	75 / 100
	magnitude 10^{-3}	0/3	0.62	0.99	–	100.0	24.9	75 / 100
	magnitude 10^{-1}	0/3	0.68	1.10	–	100.0	12.3	92 / 100
	magnitude 10^{1}	3/3	0.82	1.96	3.6	98.4	20.1	100 / 100
	scalar n=1	0/3	0.56	0.92	–	58.0	26.1	25 / 0
Franka	selected $\lambda_\perp$=100	1/1	0.81	1.90	21.4	100.0	12.9	83 / 100
	no penalty	0/1	0.27	1.11	–	100.0	24.1	58 / 25
	magnitude 10^{-1}	1/1	0.82	1.82	30.7	100.0	11.1	92 / 100
	magnitude 10^{1}	1/1	1.00	9.13	1.1	3.1	13.1	0 / 0
	scalar n=1	0/1	0.64	0.97	–	69.4	26.9	25 / 0

the gate, iso is gray when fewer than two pass, and succ is realized success at $\rho = 0.9$ under the 20% axis-mean tolerance. Because adding q_0 also changes capacity and optimization, this supports the zero-intercept design as a package rather than identifying antisymmetry alone.

Table 21: Zero-intercept versus affine critic over three paired seeds at reset 0. Rows report mean±sd; iso is gray when fewer than two seeds pass the gate.

Platform	Critic	iso	$R^2_{\min}$	valid	gap	$\mathrm{err}_{\mathrm{in}}$ (%)	succ (%)
HalfCheetah	SLC (pure linear)	4.37±0.81	0.94±0.02	3/3	3	3.5±0.8	100±0
	SLC + q_0 (affine)	1.03±0.32	0.50±0.13	0/3	26	10.2±2.6	78±21
Ant	SLC (pure linear)	2.23±0.36	0.89±0.03	3/3	6	6.5±2.4	100±0
	SLC + q_0 (affine)	0.66±0.21	0.15±0.12	0/3	36	13.0±2.8	17±8
Humanoid	SLC (pure linear)	2.68±1.14	0.86±0.04	3/3	7	7.4±5.8	97±5
	SLC + q_0 (affine)	1.44±0.67	0.68±0.12	1/3	26	10.2±4.0	81±21
Franka	SLC (pure linear)	1.58±0.32	0.79±0.03	3/3	25	16.3±7.5	75±8
	SLC + q_0 (affine)	0.39±0.42	0.32±0.14	0/3	14	16.3±6.3	22±13

Critic capacity: shared-trunk control. The default SLC critic gives each basis its own two-layer width-1024 MLP with output dimension $n = 64$ and has 3–9 times the parameters of a baseline critic (Table 22). A shared-trunk variant reduces this to 1.2–1.5 times by splitting one trunk's nB outputs into bases, but it also changes parameter sharing and representation structure, so it is not a capacity-only control. Intrinsic reward, platform, seed 0, reset 0, selected $\lambda_\perp$, and the paired batch's composition scale are matched. That scale predates the final Humanoid reevaluation, so the displayed 13.9% control value can differ from the 14.0% in Table 20. Params counts one critic head including w; the baseline row is one DADS/METRA critic with skill appended to its input. Gray iso fails the gate, and succ uses the 20% axis-mean tolerance at $\rho = 0.9$.

The shared trunk remains gate-valid on HalfCheetah and Humanoid, although Humanoid iso and error worsen. Thus the large parameter count is not necessary for gate validity on those two cells. Its Ant and Franka failures do not distinguish capacity from the changed sharing structure. For the two gate-valid cells, HalfCheetah iso improves from 4.99 to 9.15, whereas Humanoid success rises from 92% to 100% while iso falls from 2.87 to 1.43 and inner error rises from 13.9% to 21.5%. These paired single-seed outcomes are descriptive rather than a capacity-only causal estimate.

Table 22: Shared-trunk control for critic capacity (seed 0, reset 0). Params counts one critic head; gray iso fails the gate, and succ is realized composition success.

Platform	Critic	iso	$R^2_{\min}$	gap	$\mathrm{err}_{\mathrm{in}}$ (%)	succ (%)	params (M)
HalfCheetah	SLC (per-basis MLPs)	4.99	0.96	3	2.6	100	6.84
	SLC (shared trunk)	9.15	0.97	2	2.1	100	1.47
	baseline single critic						1.08
Ant	SLC (per-basis MLPs)	2.45	0.92	5	3.9	100	4.62
	SLC (shared trunk)	4.00	0.74	28	4.2	100	1.35
	baseline single critic						1.09
Humanoid	SLC (per-basis MLPs)	2.87	0.88	3	13.9	92	9.66
	SLC (shared trunk)	1.43	0.86	7	21.5	100	1.67
	baseline single critic						1.15
Franka	SLC (per-basis MLPs)	1.90	0.81	21	12.9	83	3.56
	SLC (shared trunk)	2.54	0.63	31	15.6	58	1.32
	baseline single critic						1.13

C.1 Geometric Descriptors

We also tried geometric descriptors other than joint angles: the distances from the torso to the front and rear feet on HalfCheetah, from the torso to the four foot tips on Ant, from the torso to both hands and the vertical height from the head to the lowest foot on Humanoid, and the end-effector coordinates, base-relative radius and height, and inter-link distances on Franka. All are bounded relative quantities. Absolute height above the ground was not used because it mixes jumping with handstands.

We then ran a controlled follow-up with two descriptors per platform: torso-to-front/rear-foot distance (appended observation indices 18,19) on HalfCheetah, torso-to-two-opposed-foot-tip distance (29,31) on Ant, and torso-to-right/left-hand distance (68,69) on Humanoid. HalfCheetah retains the supine main physics. Ant is additionally flipped to remove foot contact, and Humanoid frees all six arm joints so hand distance reflects full-arm kinematics. We call the setting that inherits each platform's main-comparison $\lambda_\perp$ the selected SLC setting; its $(B, \lambda_\perp)$ values are $(3, 0), (4, 1), (8, 1)$, respectively. On HalfCheetah, SLC, DADS, METRA, and CSF use $B = 3$, while FSD and SUSD use $B = 4$. Every method uses $B = 4$ on Ant and $B = 8$ on Humanoid. The alternative SLC setting keeps the same B and uses $\lambda_\perp = 1, 0, 0$, respectively. Thus this tests applicability after retraining on geometric descriptors under the stated configurations, not a descriptor-only substitution in every main environment. Each platform has nine training runs: the selected SLC at three seeds, its alternative regularization setting at seed 0, and DADS, METRA, official InfoNCE CSF, FSD, and SUSD at seed 0 (27 runs total). Fit, iso, and gap use three resets. Composition uses reset 0 and a platform-common range derived from the seven seed-0 configurations. Table 23 reports the selected SLC and five comparison methods. The selected SLC row aggregates nine fit cells over three training seeds, whereas comparison rows contain three cells from seed 0. Iso and gap aggregate only gate-valid cells and become gray when a row is majority-invalid; gate-invalid composition values are also gray diagnostics. Gap is relative to the sampled-sine reference, $\mathrm{err}_{\mathrm{in}}$ uses $\rho \leq 0.5$, and succ/every apply the 20% axis-mean/every-axis criteria at $\rho = 0.9$. The baseline rows are single-seed diagnostics, so this is a scope test rather than a new ranked main comparison. As elsewhere, FSD receives the descriptor partition and its dense-sphere probes are outside its block-sparse training support. The separate Franka comparison below uses no-kitchen physics but an end-effector action mode and appended end-effector observations.

SLC therefore extends after retraining beyond joint angles on HalfCheetah and Ant, although Ant is less continuous than the joint-angle main result. It does not succeed on Humanoid hand distances under the tested settings: the apparently low composition error is below the fit gate and is excluded from comparison. FSD passes the affine gate on Humanoid, but its response remains strongly step-like. Thus nonlinear geometric control is platform- and method-dependent rather than uniformly successful or uniformly impossible.

The separate Franka end-effector-coordinate comparison uses one training seed under its RNG-independent reset. All rows use $B = 3$, end-effector action control, and world-frame end-effector x, y coordinates at appended indices 60,61. The fixed robot base makes these equivalent to base-relative coordinates up to a constant translation. Both SLC settings, $\lambda_\perp \in \{0, 100\}$, failed the affine-fit gate ($R^2_{\min} = 0.24$ and 0.25), while METRA and METRA+SF (the METRA representation objective with the SF critic of CSF) passed with $R^2_{\min} = 0.81$ and 0.82 but their responses along

Table 23: Controlled geometric-descriptor follow-up. Selected SLC uses three seeds and $\lambda_\perp = 0, 1, 1$ in platform order; comparisons use seed 0. Gray entries are majority gate-invalid diagnostics.

Platform	Method	valid	iso ↑	gap ↓	$\mathrm{err}_{\mathrm{in}}$ ↓ (%)	succ/every ↑ (%)	reach (%)
HalfCheetah	DADS	0/3	–	–	5.5	50/17	0
	METRA	3/3	1.04	12.7	18.7	100/67	83
	CSF	0/3	–	–	13.5	83/50	33
	FSD	3/3	6.44	9.9	11.8	100/92	100
	SUSD	0/3	–	–	7.7	42/0	0
	SLC (3 seeds)	9/9	3.04	5.6	7.1±3.4	100/97	100
Ant	DADS	0/3	–	–	16.0	75/33	0
	METRA	0/3	–	–	25.5	33/25	100
	CSF	3/3	1.50	7.0	23.9	67/58	67
	FSD	3/3	10.78	16.5	18.0	100/83	100
	SUSD	0/3	–	–	8.9	8/0	0
	SLC (3 seeds)	9/9	2.12	9.3	9.8±3.8	97/89	100
Humanoid	DADS	0/3	–	–	12.2	33/8	25
	METRA	0/3	–	–	35.9	25/0	100
	CSF	0/3	–	–	29.1	92/17	100
	FSD	3/3	1.69	33.1	20.6	100/75	100
	SUSD	1/3	2.28	15.1	6.3	42/17	0
	SLC (3 seeds)	0/9	–	–	10.5±1.7	89/61	100

the great circle were square waves. An affine relation observed at joint angles need not survive nonlinear forward kinematics. These experiments show partial but nonuniform applicability after retraining. They do not establish a structural impossibility, and the main comparison remains on joint descriptors.

Other earlier single-seed pilots used a protocol preceding the current gate. The supplementary run inventory lists them. We treat them as hypothesis-generating and do not mix them with the controlled follow-up above.

C.2 Great-Circle Sweeps, Composition Grids, and 48-Skill Contact Sheets

All panels below use training seed 0 and reset 0. For the great-circle plots, each descriptor is normalized using the minimum and maximum over the random skill probes used for that run's fit. Circle values can therefore lie outside $[0, 1]$. The normalized curves show response shape, while the iso values labeled beneath them are computed from the unnormalized angles. We first contrast CSF and SLC, then expand the comparison to every learned row, composition requests, and the full 48-skill sweeps.

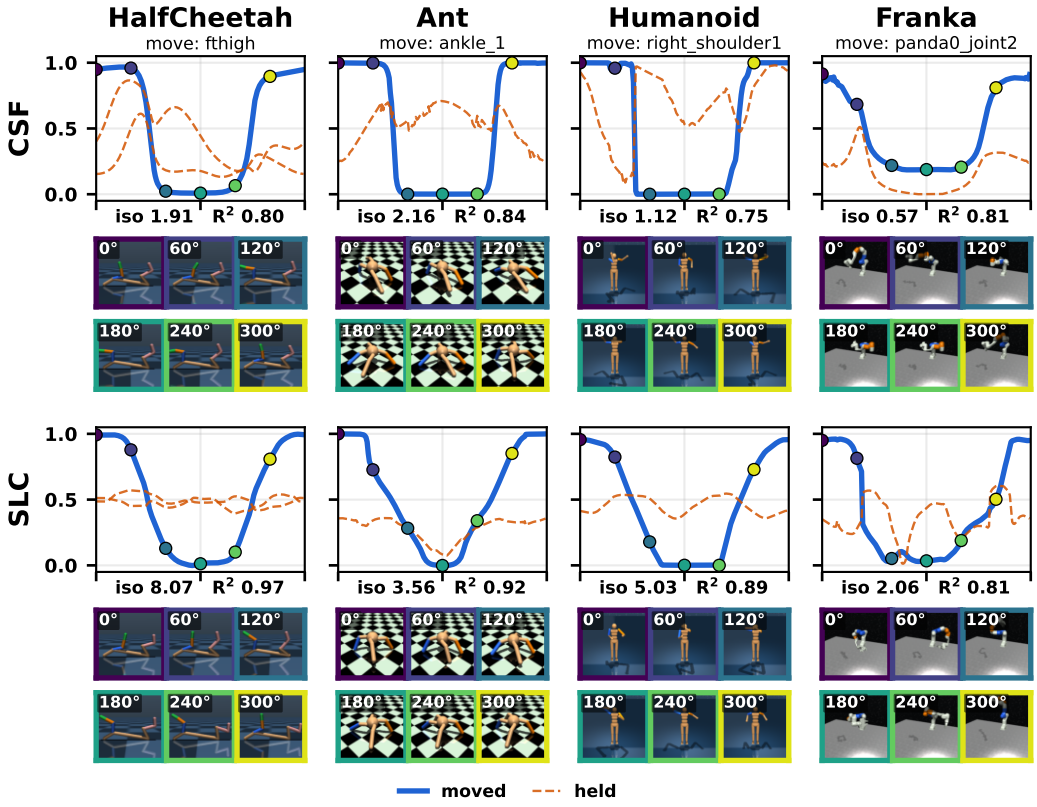

Figure 12: Great-circle sweeps of the first descriptor on each platform (moved, thick blue; held, dashed orange) with the final-frame pose at the six marked angles. Top: CSF; bottom: SLC with the per-platform $\lambda_\perp$. Labels below each curve report iso_i and descriptor-wise R_i^2 for this seed and reset. Normalization is defined in the text.

Figure 13 expands this sweep to all seven learned rows. The $\lambda_\perp = 0$ row shows three failure modes: on Ant the critic diverges and the sweep is noise, while on Humanoid and Franka the moved descriptor saturates and the held descriptor leaks. METRA and CSF largely drive descriptors to their limits and produce step-like responses, and DADS does not separate the axes. FSD obtains high iso on Ant by freezing the held descriptor while turning the moved one into a staircase, its fit collapses on Franka (R^2 0.10), and SUSD also exhibits held-descriptor leakage in these sweeps.

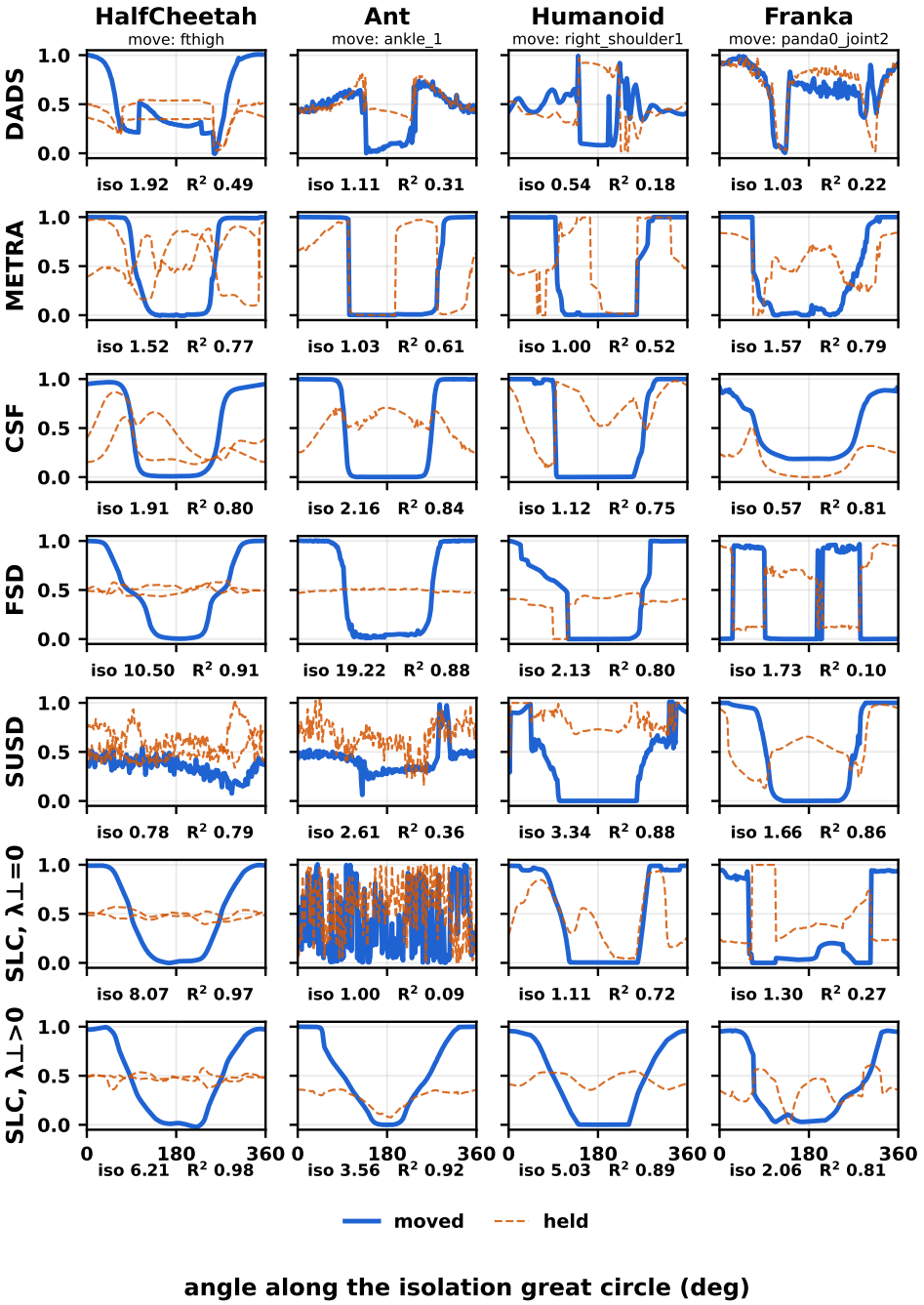

Figure 13: Great-circle sweeps for all seven learned rows of Table 1 (training seed 0, reset 0; normalization as in Figure 12). Rows are DADS, METRA, CSF, FSD, SUSD, SLC with $\lambda_\perp{=}0$, and SLC with $\lambda_\perp{>}0$. Each panel varies the first descriptor while holding the others; labels below each curve report iso_i and descriptor-wise R_i^2 for this cell.

Figure 14 shows every learned row in the requested–realized plane. In these seed-0 plots, DADS has a contracted and only weakly tracking response, especially on Franka. METRA jumps among a few corner poses, CSF remains coupled and misses many requests, and FSD follows requests on HalfCheetah and Ant but not on Franka.

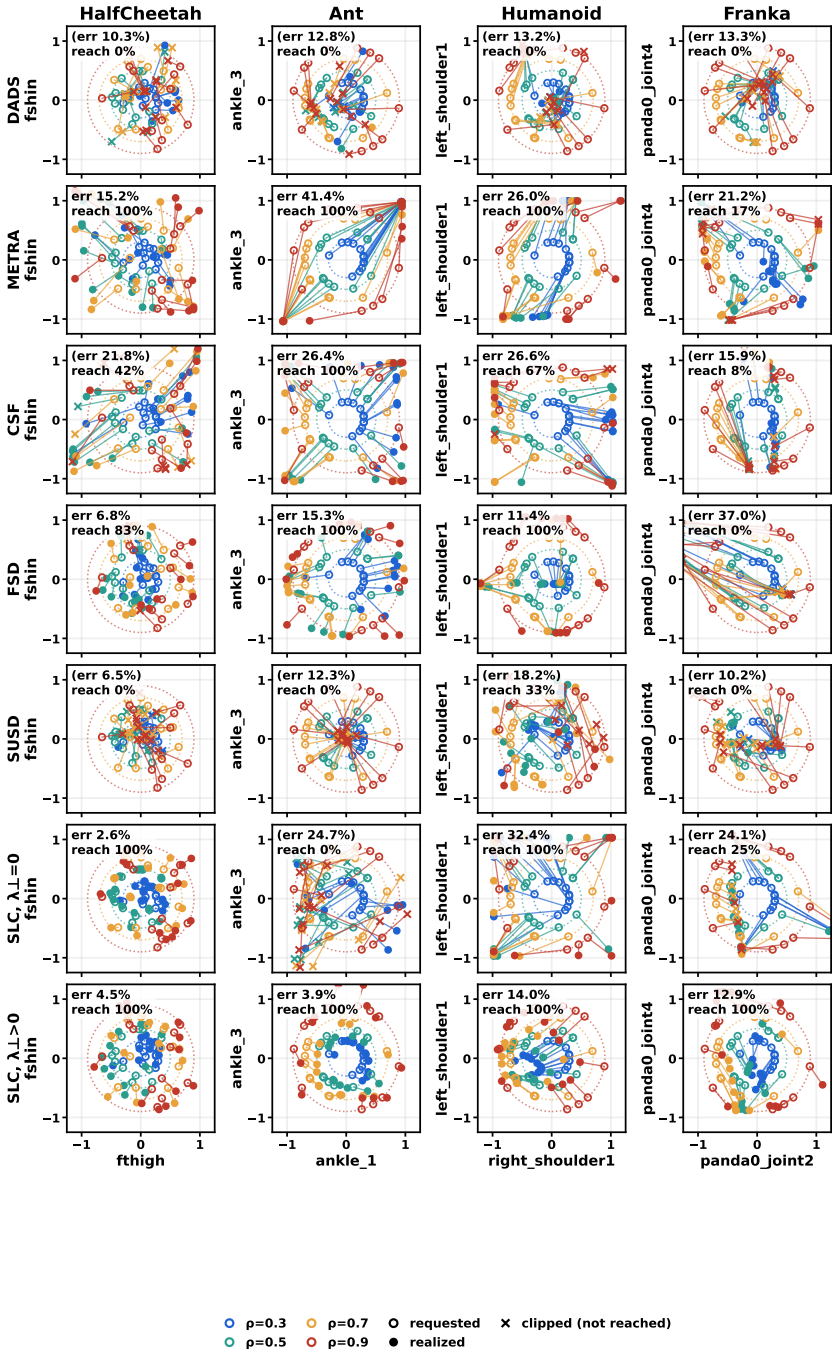

Figure 14: Requested (open) and realized (filled) descriptor pairs in the plane $(d-b)/h$ for all seven learned rows of Table 1 (training seed 0, reset 0). Rows follow the order of Figure 13. Circles mark the request radius ρ, crosses mark clipped requests, and insets give inner error and reach. HalfCheetah is projected onto its first two descriptors.

Figures 15 and 16 show a separate qualitative 3×3 grid of requests at axis coordinates $\{-0.5, 0, +0.5\}h$, rather than the 48 radial requests used by the aggregate composition metric. Under the main protocol's full-range, axis-mean normalization, the mean error over the nine cells is 3.4% versus 16.2% on HalfCheetah, 4.7% versus 19.6% on Ant, 13.0% versus 24.5% on Humanoid, and 16.9% versus 17.6% on Franka for SLC versus CSF. The CSF grids have 2, 0, 0, and 6 clipped requests, respectively. HalfCheetah requests its first two descriptors while holding the third, `ffoot`, at the fitted center. All grids use training seed 0 and reset 0.

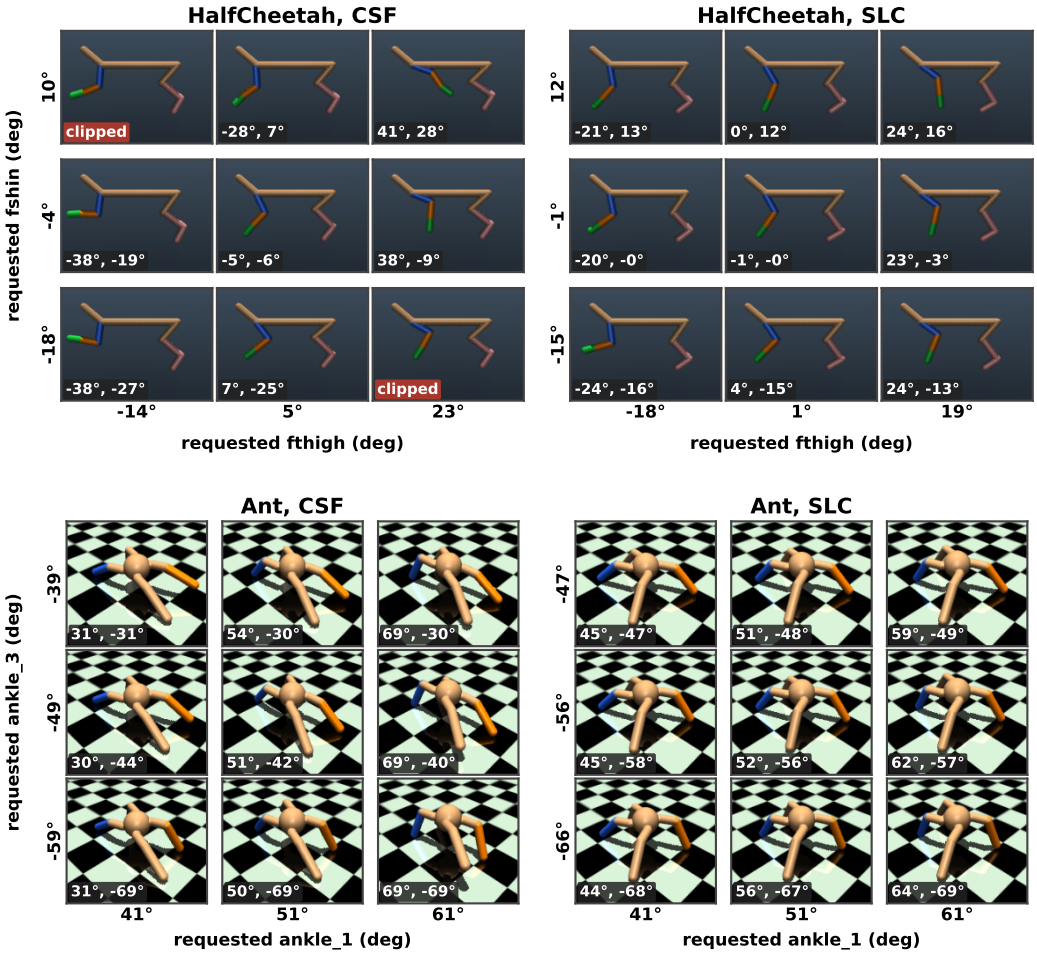

Figure 15: Requested-combination grids for HalfCheetah and Ant (CSF left, SLC right; seed 0, one reset). Axes: requested angles at $\{-0.5, 0, +0.5\}h$ from each method's own center; tile labels: realized angles. In the SLC panel, the tiles A, B, and A+B of Figure 3 are the cells $(+0.5h, 0)$, $(0, +0.5h)$, and $(+0.5h, +0.5h)$ of this grid. The seed-0 SLC cells on Humanoid and Franka (Figure 16) still visibly miss some requests.

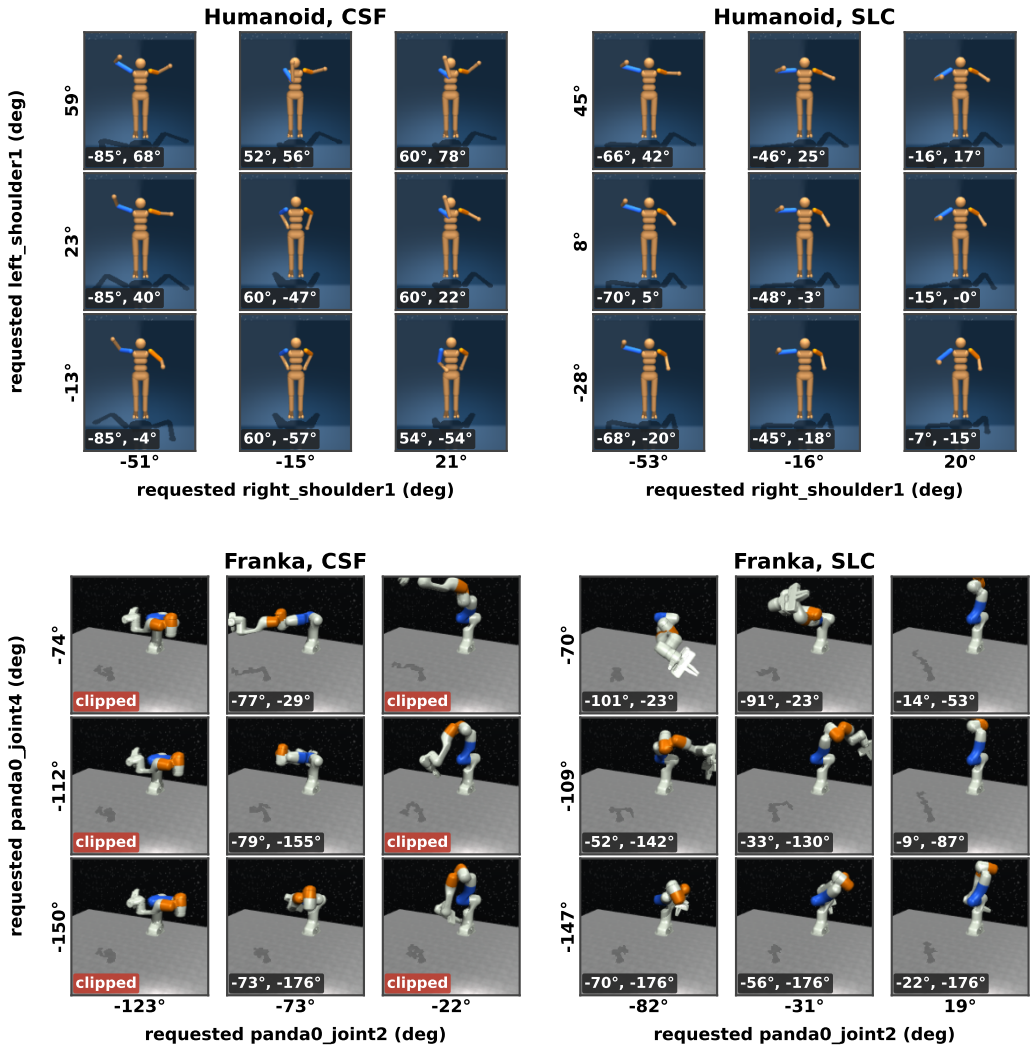

Figure 16: Requested-combination grids for Humanoid (`right_shoulder1`, `left_shoulder1`) and Franka (`panda0_joint2`, `panda0_joint4`), same layout as Figure 15; red labels are clipped requests.

Figure 17 shows the three HalfCheetah control paths together.

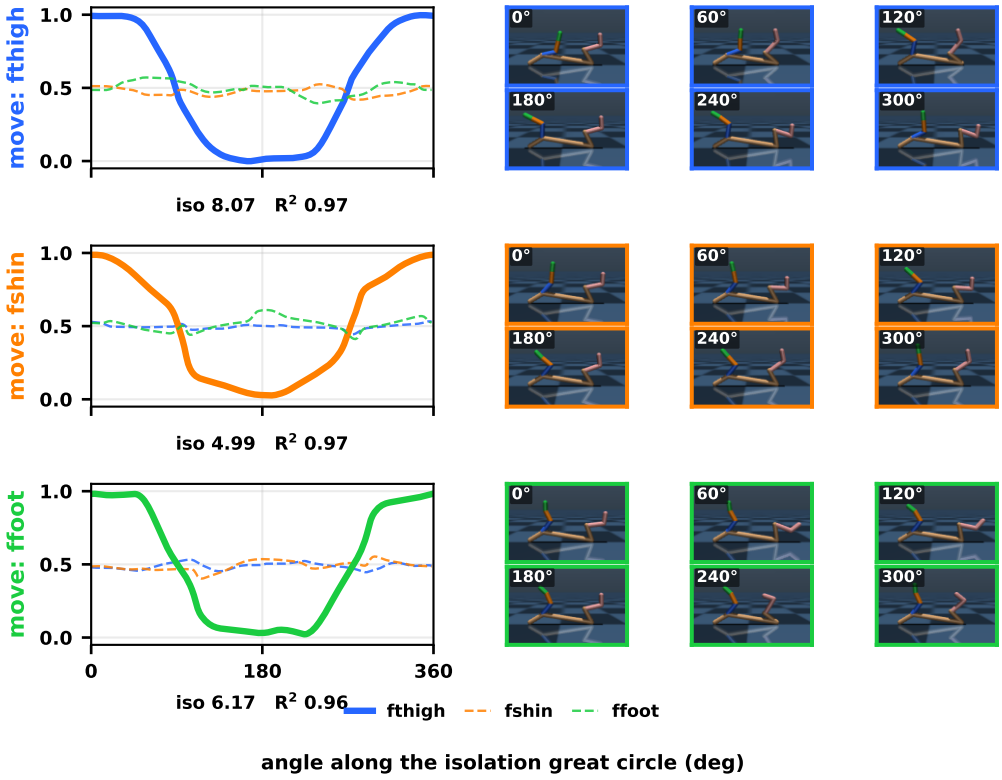

Figure 17: HalfCheetah, SLC: each descriptor is moved along its own isolation great circle while the other two are held (normalization as defined above; poses at the marked angles, moved part painted in the row color). Labels below each curve report iso_i and descriptor-wise R_i^2.

Full 48-skill contact sheets. Figures 18 and 19 show every sampled skill on HalfCheetah and Ant. These are qualitative rollout views. The reported metrics use the late-window descriptors rather than the rendered frames. The moved part is blue and the held parts are orange and green.



Figure 18: HalfCheetah, SLC: final-frame poses at 48 equally spaced angles along the isolation great circle of `fthigh` (same cell as Figure 12); each tile is the last frame of a 200-step rollout, annotated with the angle and the late-window-mean descriptors (moved marked with an asterisk). The policy is trained and rendered lying on its back.

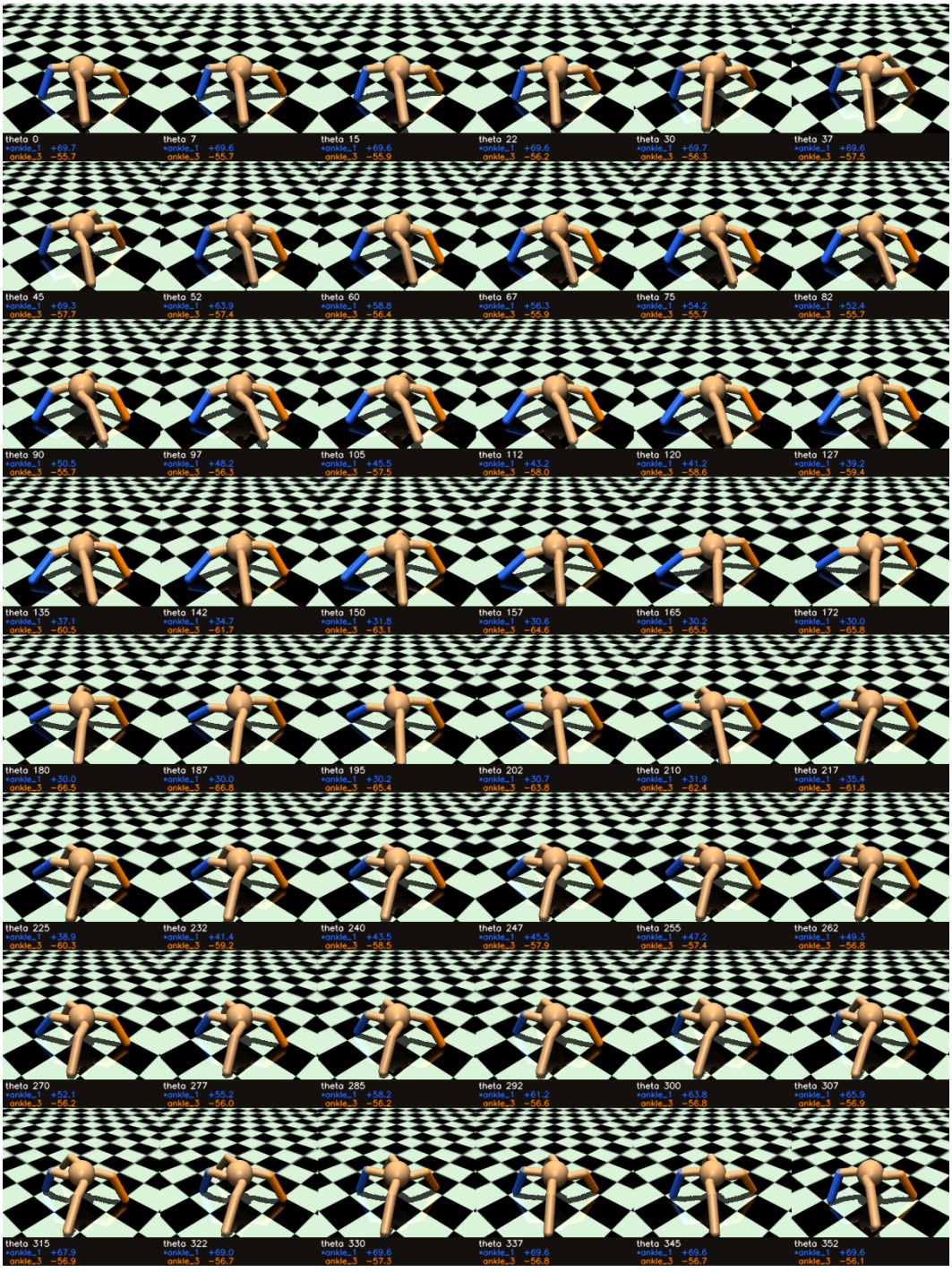

Figure 19: Ant, SLC: 48-skill sweep along the isolation great circle of `ankle_1` (blue leg), with `ankle_3` (orange leg) held. Layout as in Figure 18.

Figures 20 and 21 complete the same contact-sheet view for the pinned Humanoid and Franka.



Figure 20: Humanoid, SLC: 48-skill sweep along the isolation great circle of `right_shoulder1` (blue arm), with `left_shoulder1` (orange arm) held. Layout as in Figure 18.

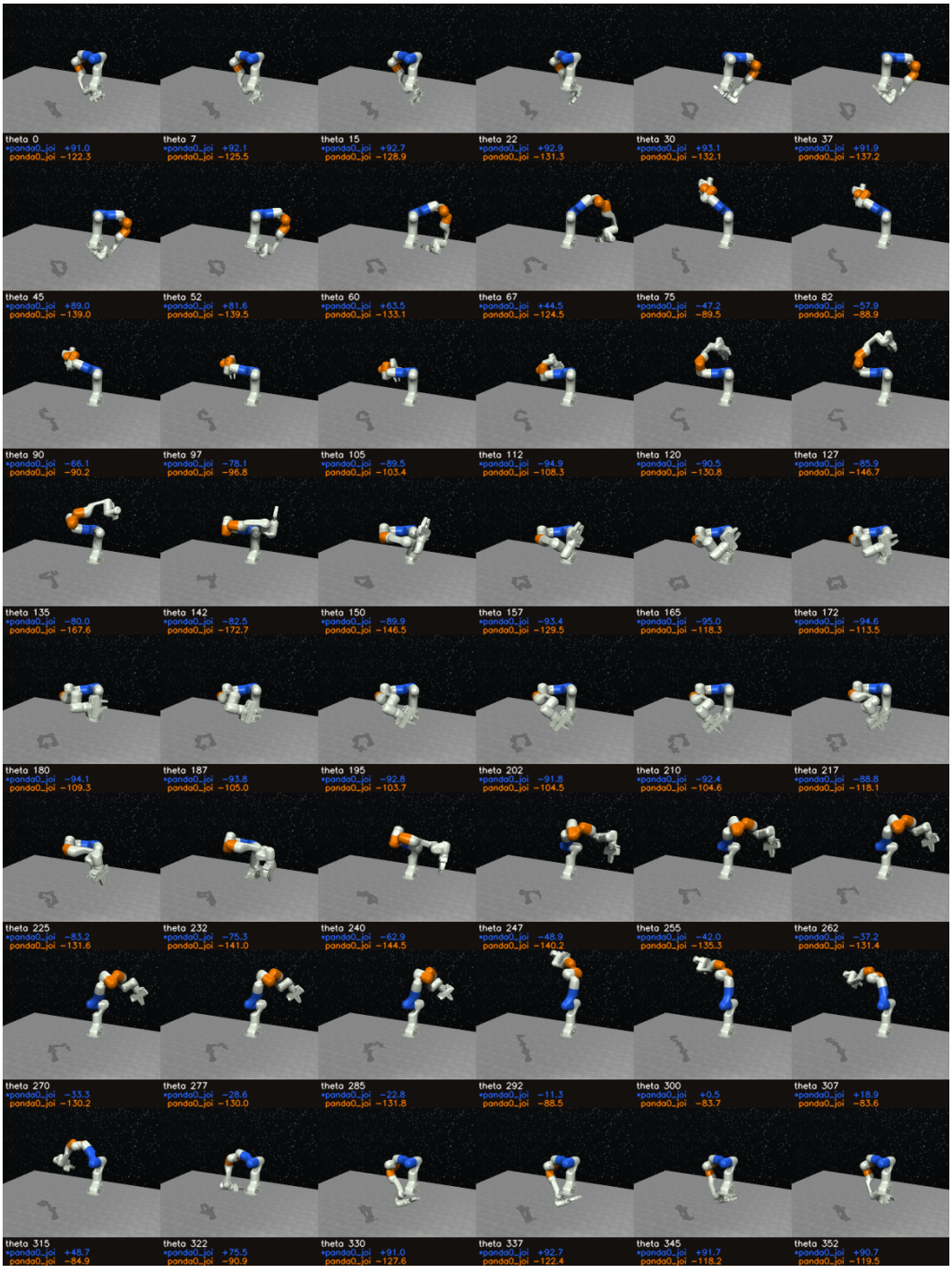

Figure 21: Franka, SLC: 48-skill sweep along the isolation great circle of `panda0_joint2` (blue link), with `panda0_joint4` (orange link) held. Layout as in Figure 18.

C.3 Unbounded Descriptors

This section supplements the unbounded-axis results of Section 5.5. All runs use seed 0 and three evaluation resets. HalfCheetah $\{x, z_{\text{body}}\}$ uses observation indices $\{0, 20\}$ and $B = 6$; Ant $\{x, \text{head_slide}\}$ uses $\{0, 29\}$, $B = 3$, the custom `ant_head.xml` slide joint (spring reference 0.2, stiffness 50, damping 3.5), and the main Ant ankle setting (stiffness 430 with spring-centered initialization). HalfCheetah $\{x, \text{neck}\}$ uses $\{0, 9\}$, $B = 3$, and `half_cheetah_neck_spring.xml`, whose neck hinge has range $[0, 0.8]$, spring reference 0.4, stiffness 50, and damping 10. Each set compares DADS, METRA, METRA+SF, and SLC with $\lambda_\perp \in \{0, 1\}$. These older scope runs do not use the official InfoNCE CSF. Table 24 reports the final-checkpoint gate results; each cell gives mean±sd of reset-level R^2_{min} followed by the number of resets passing 0.75. Every cell has nonzero descriptor variance and full-row-rank G, so all failures are due to fit quality. Figures 22 and 23 show the corresponding response curves.

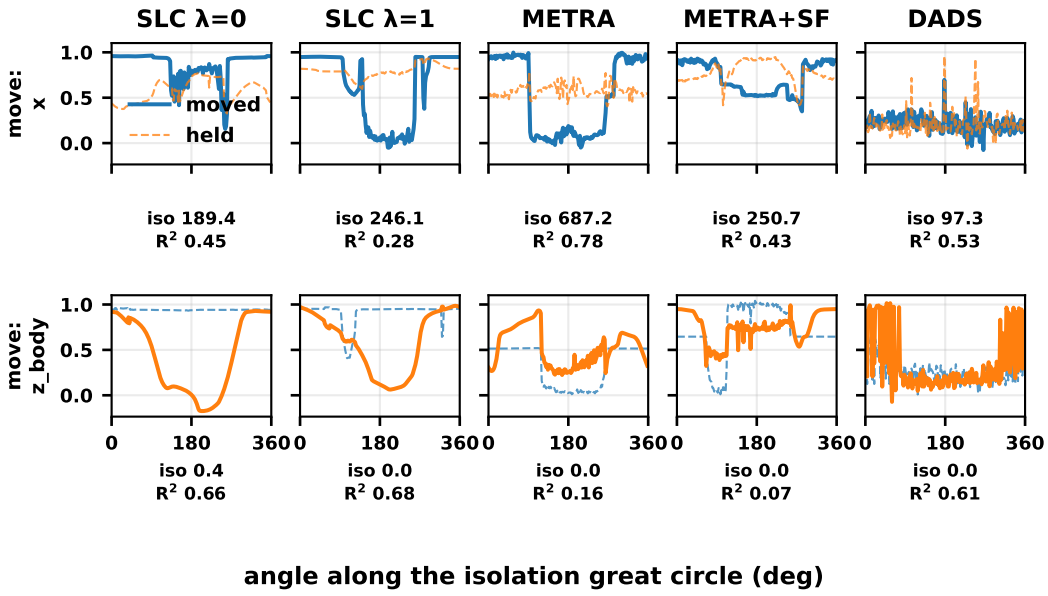

Figure 22: HalfCheetah with the unbounded pair $\{x, z_{\text{body}}\}$: great-circle sweeps of x (top) and z_{body} (bottom). Columns are the five configurations without a supplied descriptor partition (DADS, METRA, METRA+SF, and two SLC settings). The other pairs and the epoch ladder are shown below.

Table 24: Affine-fit gate for the unbounded descriptor pairs at epoch 8000 (seed 0, three resets). Cells report mean±sd $R^2_{\min}$ and passing-reset count.

Descriptor pair	DADS	METRA	METRA+SF	SLC, $\lambda_\perp=0$	SLC, $\lambda_\perp=1$
HalfCheetah $\{x, z_{\text{body}}\}$	0.542±0.028 (0/3)	0.166±0.002 (0/3)	0.077±0.005 (0/3)	0.484±0.039 (0/3)	0.290±0.016 (0/3)
Ant $\{x, \text{head_slide}\}$	0.246±0.021 (0/3)	0.601±0.028 (0/3)	0.013±0.005 (0/3)	0.621±0.027 (0/3)	0.657±0.014 (0/3)
HalfCheetah $\{x, \text{neck}\}$	0.622±0.010 (0/3)	0.559±0.005 (0/3)	0.519±0.006 (0/3)	0.585±0.014 (0/3)	0.540±0.005 (0/3)

All 15 descriptor-pair–configuration combinations therefore fail the gate in all three resets. Figure 23 shows the remaining two pairs, $\{x, \text{head_slide}\}$ on Ant and $\{x, \text{neck}\}$ on a HalfCheetah with a spring-loaded neck.

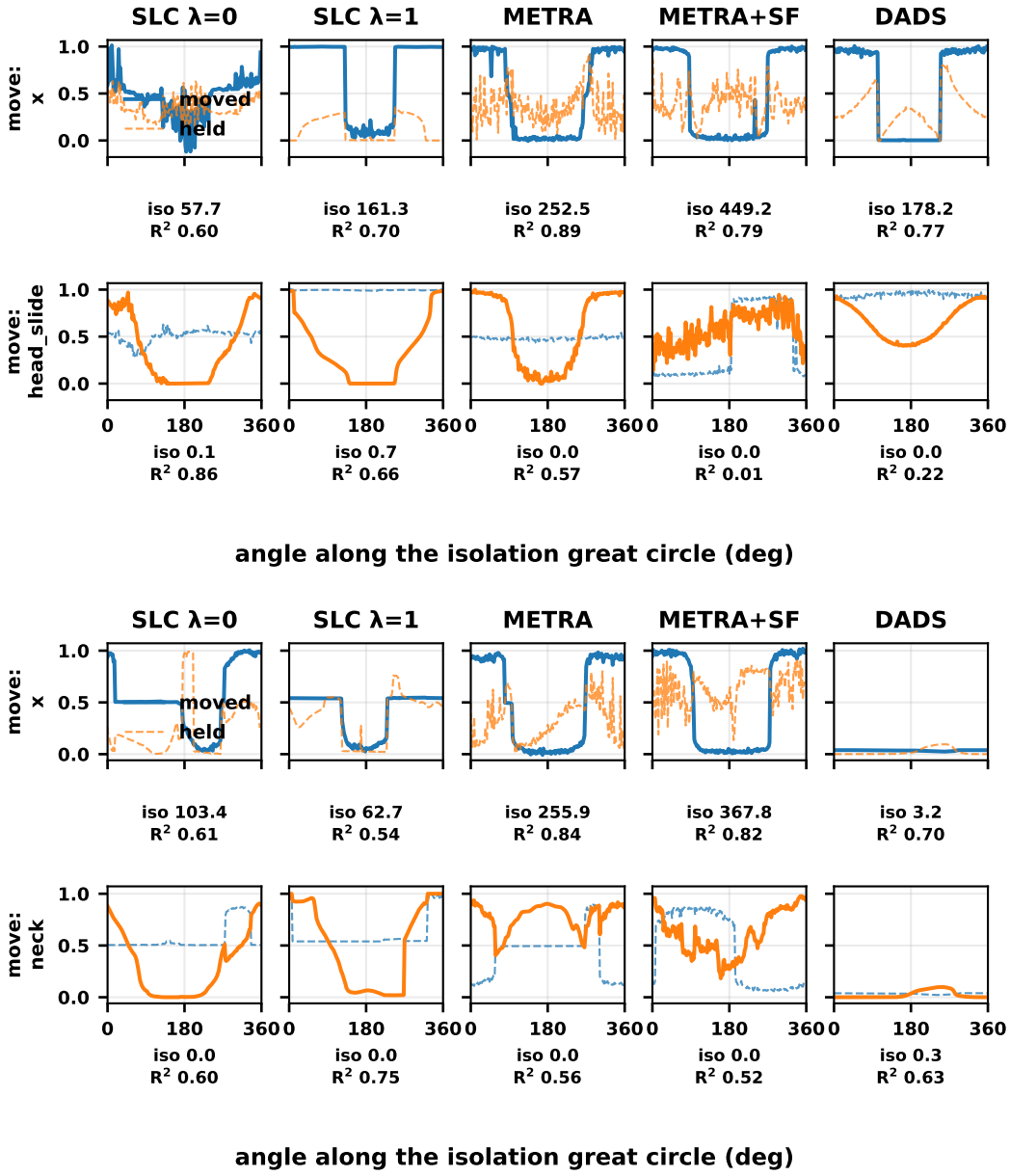

Figure 23: The two remaining unbounded pairs, five configurations each: Ant $\{x, \text{head_slide}\}$ (top) and HalfCheetah $\{x, \text{neck}\}$ with a spring-loaded neck (bottom). The layout follows Figure 22.

Figure 24 sweeps SLC with $\lambda_\perp = 0$ at intermediate checkpoints. On HalfCheetah a graded structure in x appears at epoch 4000 and disappears by 8000, and on Ant it never appears. Thus this training setup did not maintain the transient unbounded-axis structure. We do not isolate the skill-discovery objective as its cause.

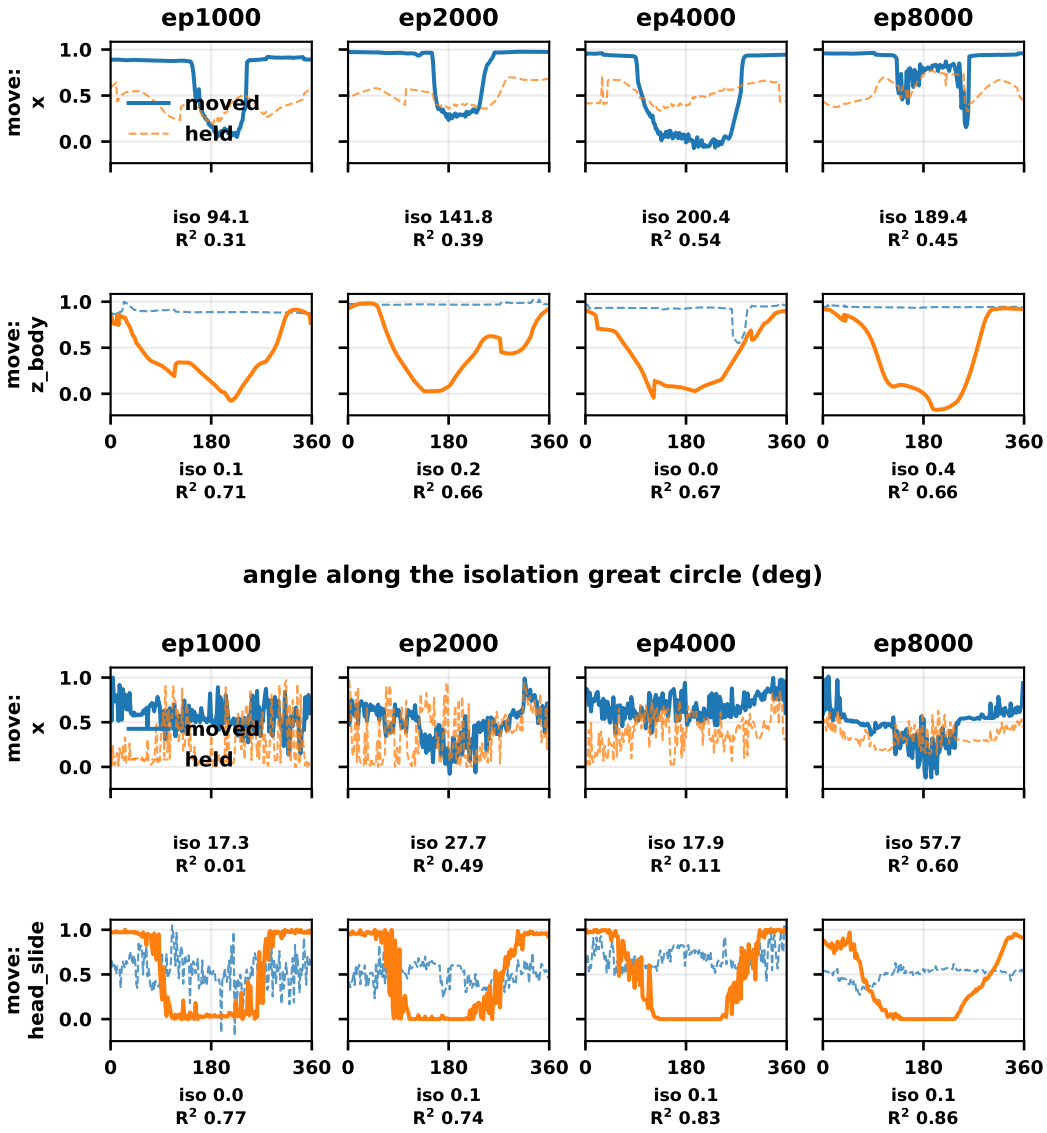

Figure 24: SLC ($\lambda_\perp$=0) on the unbounded pairs at epochs 1000, 2000, 4000, 8000: HalfCheetah $\{x, z_{\text{body}}\}$ (top) and Ant $\{x, \text{head_slide}\}$ (bottom).