JOINT-FORCE MATCHING: CONTACT MECHANICS AS SUPERVISION FOR RL CRITICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contact forces improve robot learning, and grasping demonstrations record them at every contact site and time step. In reinforcement learning (RL), the critic estimates the return of each action and guides the policy. Existing methods, however, predict each contact force or take it as an input, and they leave out how the forces load the hand's joints. Under elastic contact, these loads are the gradient of a single energy. Which structure in contact forces improves the critic's value estimates has remained unclear. We propose joint-force matching (JFM), which maps measured contact forces through the hand's Jacobians to joint loads and trains the negative joint-angle gradient of a scalar energy on the critic's representation to match them. JFM keeps force prediction and adds this gradient target. In simulated multi-finger grasping, JFM's critics have a lower return error than those of a force-prediction baseline at every pretraining seed, and JFM's offline-to-online policy succeeds on 33 of 158 held-out trajectories in closed loop against the baseline's 18. A free-vector head reaches a four to five times lower loss on the same joint loads, yet its critics have a higher mean return error than JFM's. We prove that JFM's energy targets the conservative component of frictional joint loads; the non-conservative remainder sets a floor on JFM's loss, and a free vector can also fit it. JFM keeps its advantage over the baseline when the energy gradient fed to the critic is zeroed at evaluation. JFM's gain in value estimates therefore comes from requiring the joint loads to be the gradient of one energy computed from the critic's representation.

1 INTRODUCTION

Contact forces improve robot learning across tasks, from reinforcement learning (RL) of contact-rich skills to policies and world models for manipulation (Lee et al., 2019; Liu et al., 2025; Higuera et al., 2026; He et al., 2026). Multi-finger grasping demonstrations record these forces at every contact site and time step (Figure 1a). In RL from demonstrations, the critic, an action-value function, estimates the return of each action and so guides the policy toward better actions. Contact forces can make these value estimates more accurate. Touch predicts whether a grasp succeeds (Calandra et al., 2017), and it adds a dense signal to the one scalar per step that the task reward gives the critic.

Existing methods, however, share a limitation for this use: they leave out how contact forces load the hand's joints. They treat each contact force as a separate measurement, predicting it as an output (Jang et al., 2026) or feeding it as an input to a policy or a world model (He et al., 2025; Zang et al., 2026). Contact mechanics relates every contact force to the loads on the hand's joints. Through the hand's kinematics, each force pushes on the joints between its contact site and the palm through a lever arm (Murray et al., 1994). For elastic contact, one energy produces all these joint loads, as the energy stored in a stretched spring produces its force. One energy thus determines the loads at every joint and hand pose. It remains unclear *what in contact forces improves the critic's value estimates*.

We propose joint-force matching (JFM), which trains the critic's representation to produce the joint loads as the gradient of one energy. JFM computes the joint loads from the measured contact forces through the hand's Jacobians (Murray et al., 1994; Prattichizzo & Trinkle, 2016). It computes a scalar energy from the representation and trains the energy's negative gradient with respect to the joint angles to match these loads (Figure 1c). JFM keeps force prediction, which supervises what the representation outputs at each state. JFM also supervises how the representation changes with the joint angles, because the energy's gradient passes through the representation's Jacobian with respect

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

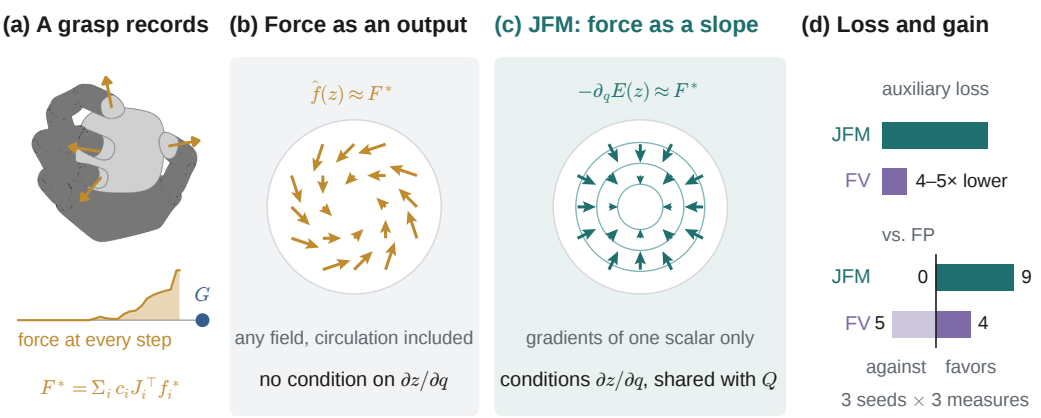

Figure 1: JFM improves on the force-prediction baseline (FP) at every seed and measure on the validation samples, whereas a free-vector head (FV) fits the same joint loads four to five times more closely yet improves on FP in four of nine comparisons. (a) One grasp, with a force at every step, here the summed site force of one demonstration, and one return G at the end. (b) Force as an output, as in FV. An output can follow any field, circulation included. (c) JFM, with force as the negative q-derivative of one scalar energy on the shared features z. (b) and (c) are schematic two-joint fields. (d) Auxiliary loss, and the comparisons with FP that favor or go against each method (three seeds by force error, return error and return rank; Table 1).

to these angles. The critic reads the same representation and still learns from task returns, as with other auxiliary objectives in RL (Jaderberg et al., 2017; Lyle et al., 2021; Schwarzer et al., 2021). JFM needs no energy labels, and the critic and the policy run without force input at evaluation.

The representation's Jacobian with respect to the joint angles links JFM's target to the critic. The critic learns to rank each action from the value of the state it reaches, and this value changes with the joint angles through the same Jacobian. Whether a grasp holds depends on how the contact loads respond to joint motion (Prattichizzo & Trinkle, 2016), so a Jacobian trained by these loads can help the critic separate grasps that hold from grasps that slip.

Force prediction and output regression each have a limitation for this link. Force prediction supervises each state separately and leaves this Jacobian without direct supervision. An output head that regresses the joint loads can fit any load field, including circulating loads that no energy produces, so a lower loss need not give the critic a better representation (Figure 1b). In practice, grasp contact is frictional, and friction adds such circulating loads. Yet an energy can still capture the spring-like part of how the loads respond to joint motion, so the energy view extends to frictional contact.

Contributions. Our experiments show that an auxiliary target improves the critic more consistently through its constraint on the representation than through the closeness of its fit.

The mechanics linking contact forces to the hand's joints gives the critic more accurate value estimates than force prediction alone. Against a force-prediction baseline (FP), JFM lowers the critic's return error at every pretraining seed, significantly across seeds on the validation samples, and predicts contact forces more accurately after offline RL training. In closed loop, JFM's offline-to-online policy completes the grasp on 33 of 158 held-out trajectories against 18 for FP, a significant gain despite less pretraining (Sections 5.2 and 5.5).

We trace JFM's gain in value estimates to the requirement that the joint loads be the gradient of an energy. A free-vector head (FV) fits the same joint loads with a four to five times lower loss. Yet FV's critics have a significantly higher mean return error than JFM's, their differences from FP change sign between seeds, and FV's offline-trained policy covers the fewest held-out trajectories of the three methods in closed loop. With the energy gradient zeroed in the critic's input at evaluation, JFM still leads FP, so the gain lies in the critic that JFM trains (Sections 5.3 and 5.4).

JFM's energy provably targets the conservative component of frictional joint loads. This component is what the gradient of a scalar energy can represent. Friction adds a remainder that no energy produces, and this remainder adds a constant to JFM's loss (Section 4). FV can also fit the

remainder and so reach a loss below this constant. Force matching, a standard method for fitting molecular energies, therefore extends to the contact data of robot hands.

2 Related Work

Contact forces in manipulation. Contact forces improve manipulation as measurements that a model reads or predicts. Policies feed measured force to the actor (Liu et al., 2025; He et al., 2025; Li et al., 2026) or build it into the action targets of imitation (Chen et al., 2025; Ablett et al., 2025). Predictive and world models learn to forecast forces, wrenches or touch (Jang et al., 2026; He et al., 2026; Zang et al., 2026; Jin et al., 2026; Higuera et al., 2026). These methods lack a model of how the forces load the hand's joints. JFM builds on force prediction and supervises the critic's representation with these joint loads through the gradient of an energy.

Auxiliary objectives in reinforcement learning. Auxiliary tasks and self-predictive objectives add training signal to the representations of RL agents (Jaderberg et al., 2017; Schwarzer et al., 2021; Lyle et al., 2021). Latent world models learn their representations from reconstruction, reward or value targets (Hafner et al., 2025; Hansen et al., 2024; Cheng et al., 2025). Offline methods such as implicit Q-learning (IQL) learn action values from fixed demonstrations (Kostrikov et al., 2022). These objectives supervise what the representation predicts at each state and leave its derivatives without direct supervision. JFM, in contrast, supervises a derivative. It serves as an auxiliary objective for the critic of such a learner and matches the joint-angle gradient of an energy to the measured joint loads.

Energy gradients as supervision. Gradient supervision trains a function through its derivatives. Score matching (Hyvärinen, 2005), Sobolev training (Czarnecki et al., 2017) and Sobolev RL (Schramm et al., 2025) are examples. Sobolev RL matches the derivatives of a critic through differentiable dynamics, and JFM takes its derivative target from measured contact forces. Networks for physical systems build a potential into the model, from conservative force fields (Chmiela et al., 2017) to Lagrangian and Hamiltonian networks (Lutter et al., 2019; Greydanus et al., 2019). Closest to JFM is force matching, which fits the gradient of a potential to reference forces (Ercolessi & Adams, 1994) and underlies coarse-grained molecular models (Izvekov & Voth, 2005; Noid et al., 2008; Wang et al., 2019). These reference forces are gradients of a molecular potential energy. Soft Q-learning and implicit behavior cloning use an energy to score actions (Haarnoja et al., 2017; Levine, 2018; Florence et al., 2022), whereas JFM's energy models joint loads. JFM brings force matching to frictional joint loads and to the critic's representation.

3 Joint-Force Matching

JFM trains one shared representation with two kinds of supervision. Measured contact forces supervise an auxiliary energy through its joint-angle gradient, and task returns supervise the critic and the policy.

3.1 Joint-load target and energy

$$F^*(q, c) = \sum_{i=1}^{N_c} c_i J_i(q)^\top f_i^*, \tag{1}$$

$$\widehat{\mathcal{L}}_{\mathrm{JFM}} = \frac{1}{|\mathcal{B}_J|} \sum_{x \in \mathcal{B}_J} \| -\partial_q E(x) - F^*(x) \|_2^2 . \tag{2}$$

Hand kinematics converts contact-site forces into joint loads. For each contact site $i = 1, \ldots, N_c$, let $f_i^* \in \mathbb{R}^3$ denote the force on the hand, $c_i \in \{0, 1\}$ its contact flag, and $J_i(q)$ its linear Jacobian, with force and Jacobian expressed in the same Cartesian frame. Equation (1) gives the vector of joint loads F^*. Under any small joint motion δq, F^* does the same work as all contact forces together, since $f_i^{*\top} J_i \delta q = (J_i^\top f_i^*)^\top \delta q$ (Figure 1a).

The learner receives transitions $(s_t, a_t, r_t, s_{t+1}, d_t)$ with termination flag d_t. We write $x = (q, u)$ for the model input, where q holds the hand joint angles and u collects the other recorded inputs:

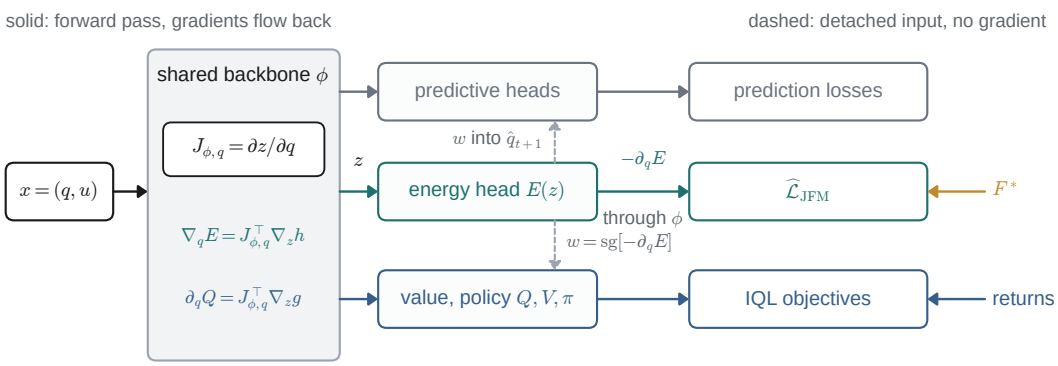

Figure 2: The joint-load target and the return objectives train one shared backbone. The forward computation runs left to right, and gradients flow back along solid arrows, so each objective updates its own head and the shared backbone, and $\widehat{\mathcal{L}}_{\mathrm{JFM}}$ alone reaches the energy head. The energy's joint derivative is taken through ϕ and feeds the value and policy heads and the transition predictor of Equation (3) as the detached input w (dashed), so no return gradient travels back along that path. Both joint derivatives factor through $J_{\phi,q}$ (Section 4.2).

joint velocities, contact flags, object pose, environment descriptors and the commanded joint target. Derivatives in q hold u fixed (Appendix A.4).

JFM matches the joint loads with the gradient of an energy computed from the shared representation. A transformer ϕ maps the input x to tokens z, with a summary token for the whole input and a site token for each contact. The energy combines a summary term and contact-gated site terms, $E(x) = h_0(z) + \sum_i c_i h_i(z)$. Each flag c_i switches its site's term on while that site touches the object, so the energy changes as contacts form and break. JFM's loss, Equation (2), is the squared error between the energy's negative joint derivative and the joint loads over a batch $\mathcal{B}_J$ of states. The loss needs no energy labels, since the measured loads supervise the energy's gradient.

Second-order backpropagation passes the JFM loss to the energy head and the shared backbone, so each measured contact force trains the critic's representation through the energy's derivative. The loads at neighboring hand poses all come from one energy's derivative, so this target constrains how the representation changes as the hand moves. An output head, in contrast, can match the loads at each state separately. State, contact and force heads train alongside JFM, and each objective updates its own head and the shared backbone (Figure 2, Appendix B).

3.2 Critic and policy training

$$\hat{q}_{t+1} = q_t + \Delta_{\mathrm{dyn}}(x_t) + \eta w_t, \tag{3}$$

$$\mathcal{L} = \lambda_s \mathcal{L}_{\mathrm{state}} + \lambda_c \mathcal{L}_{\mathrm{contact}} + \lambda_f \mathcal{L}_{\mathrm{force}} + \lambda_{\mathrm{jfm}} \widehat{\mathcal{L}}_{\mathrm{JFM}} + \lambda_q (\mathcal{L}_Q + \mathcal{L}_V) + \lambda_\pi \mathcal{L}_\pi. \tag{4}$$

We reserve the energy-head parameters for force supervision and train action values from task returns. A detached derivative $w_t = \mathrm{sg}[-\partial_q E(x_t)]$ conditions the value and policy heads. The operator sg passes its input forward unchanged and blocks its gradient. The transition predictor of Equation (3) also uses this derivative, where Δ_{dyn} is a learned increment and η sets the residual scale. Because both inputs are detached, the JFM loss alone trains the energy head. FP replaces ηw_t with a learned correction from the summary token and gives its value and policy heads a zero auxiliary input.

Return fitting follows implicit Q-learning (IQL) (Kostrikov et al., 2022), which fits action values on the demonstrated actions. The state-value head V fits an expectile of the target critic, the action-value head Q fits a bootstrapped return, and the policy π imitates demonstrated actions with an advantage-dependent weight. The task reward credits contact, lift and stability, penalizes changes in contact force, and adds a bonus for a successful grasp (Appendix B). Pretraining initializes the predictive backbone and energy, and offline learning then adds the value and policy objectives of Equation (4) (Algorithm 1 in Appendix B). The JFM weight rises from 0.2 to 1.0 when offline RL training begins, so the joint-load supervision stays active while the critic trains.

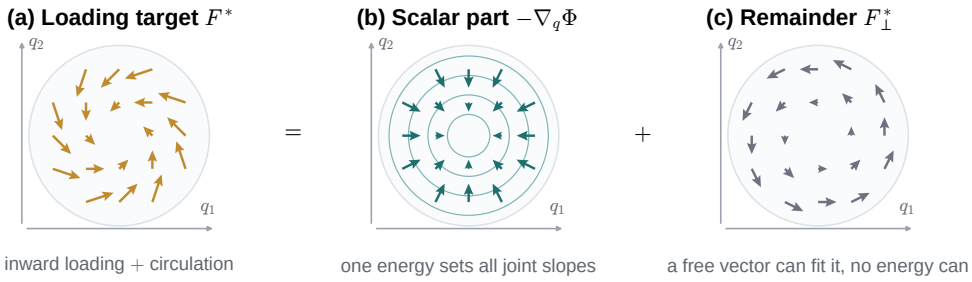

Figure 3: An energy gradient reproduces the conservative part of a load field, and a free vector can also fit the circulating remainder. (a) The two-joint load target $F^* = (-q_1 - q_2, -q_2 + q_1)$. (b) The component a scalar energy can produce, for $\Phi = (q_1^2 + q_2^2)/2$, with contours of equal energy. (c) The orthogonal remainder, which no scalar gradient reproduces. The uniform disk measure makes the two parts orthogonal; arrows are load vectors in configuration space.

4 JFM UNDER FRICTIONAL CONTACT

4.1 CONSERVATIVE COMPONENT OF THE JOINT LOADS

$$\mathcal{L}_{\mathrm{JFM}}(E) = \|-\nabla_q E - F^*\|^2_{L^2(\rho)}\,. \tag{5}$$

Proposition 1 (Projection identity)**.** *Under the projection assumption of Appendix A.1, the joint loads split as $F^* = -\nabla_q \Phi + F^*_\perp$ with $\Phi \in \mathcal{V}$, where $F^*_\perp$ is orthogonal in $L^2(\rho)$ to the joint gradient of every admissible potential. For every admissible $E \in \mathcal{V}$ and every differentiable energy $E_\theta \in \mathcal{V}$,*

$$\mathcal{L}_{\mathrm{JFM}}(E) = \|E - \Phi\|^2_{\dot{H}^1(\rho)} + \|F^*_\perp\|^2_{L^2(\rho)}, \tag{6}$$

$$\nabla_\theta \mathcal{L}_{\mathrm{JFM}}(E_\theta) = \nabla_\theta \|E_\theta - \Phi\|^2_{\dot{H}^1(\rho)}, \tag{7}$$

where $\|v\|^2_{\dot{H}^1(\rho)} = \|\nabla_q v\|^2_{L^2(\rho)}$.

JFM's energy targets the conservative component of frictional joint loads, the part of the loads given by the gradient of an energy: at fixed conditioning u, one scalar energy sets the loads on all joints at once. JFM minimizes the population objective of Equation (5), where ρ is the input measure and E ranges over a linear space $\mathcal{V}$ of potentials that contains every energy the network represents. The loss depends on the energy through its joint gradient alone and leaves its additive level free. Proposition 1 splits the loads into the conservative component $-\nabla_q \Phi$ and a remainder $F^*_\perp$ that no admissible energy produces (Figure 3). Physically, $F^*_\perp$ is the part of the loads whose work depends on the path of the joints, as the work of friction does. The remainder adds a constant to JFM's loss, so the population gradient is the gradient of the fit to the conservative component, where θ collects the energy and backbone parameters.

The remainder separates JFM from a free-vector head. The Hessian of Φ is symmetric, so an energy represents the loads with a symmetric stiffness, and the remainder contains the antisymmetric part of the loads' Jacobian, the circulation of Figure 3c. In the two-joint example of Figure 3, the target combines an inward pull with a circulation. JFM's energy recovers the inward pull, and a free vector can fit both parts. FV can likewise fit the remainder of the grasp loads, so its loss can fall below the constant that bounds JFM's loss, and Section 5.3 tests whether this closer fit improves the critic. The identity is a weighted Helmholtz–Hodge decomposition (Bhatia et al., 2013), the projection that underlies force matching in molecular coarse-graining (Noid et al., 2008); JFM applies it to frictional joint loads and so extends force matching to frictional contact (proof in Appendix A.2).

4.2 SHARED BACKBONE JACOBIAN

$$\nabla_q E = J^\top_{\phi,q} \nabla_z h, \qquad \partial_q Q|_{u,a,w} = J^\top_{\phi,q} \nabla_z g. \tag{8}$$

The energy gradient and the critic's joint-angle derivative pass through one backbone Jacobian. The energy head reads the summary and contact tokens, and the critic reads the summary token of the

same representation $z = \phi(x)$. Write $E(x) = h(\phi(x), c)$ for the energy head and $Q(x, a; w) = g(\phi(x), a, w)$ for the critic, where g selects the summary token and w is the auxiliary input. With $J_{\phi,q} = \partial\phi/\partial q$ taken at fixed u, candidate action a and w, the chain rule gives Equation (8), so the JFM loss trains the sensitivities through which the critic's values change with the joint angles. Each action reaches the critic's bootstrap target through the predicted next joint angles (Appendix B), so these sensitivities also shape how the critic ranks actions. Where contacts load the hand, JFM trains the energy's joint gradient to follow the loads. The representation then varies with the joint motions that these loads resist, and the critic reads this same variation. JFM supervises the combination of these sensitivities that the energy head selects, and task returns train the combination that the critic uses (Appendix A.3). The critic therefore evaluates actions with a representation trained by the joint loads.

5 Experiments

5.1 Experimental setup

Human operators teleoperate a four-finger LEAP hand (Shaw et al., 2023) with 16 actuated joints in simulation to grasp 71 rigid objects. The demonstrations have a median length of 79 steps, and simulator sensors report the contact force at 17 sites on the hand at every step (Figure 4). Each object–pose–variant tuple contributes four trajectories for training and one for validation (Appendix B). In the main study, the three methods share a six-layer transformer of width 256, the state, contact and force prediction objectives, and budgets of 20 pretraining and 15 offline epochs (Appendix B). FP uses these shared objectives. JFM adds the joint loads of Equation (1) through its energy. FV regresses the same joint loads with a vector head that matches JFM's energy head in capacity and feeds the critic and the transition predictor as JFM's w does. FV and JFM thus differ in whether they match the loads through an energy derivative.

Three measures read the critic's value estimates and the force predictions on held-out data. Return error is the root-mean-square error between the critic's action values and the recorded discounted returns, return rank is their Spearman correlation, and force error is the mean squared error of the force predictions on held-out states. Lower return error and higher return rank mean more accurate value estimates. The expectile bootstrap of IQL places each method's action values above the recorded returns on average, and centered return error removes this mean offset.

Two studies compare the methods. The main study trains each method from three pretraining seeds and evaluates the nine models twice, on 8,000 validation samples and on every contact state of the 157 successful demonstrations among the 158 held-out ones. The replication repeats offline RL training for JFM and FP over eight seeds from one pretrained checkpoint each, with the joint-load target stored with the data, which differs from Equation (1) by the rotation between world and hand-base coordinates (Appendix A.4). Two further controls in the replication keep a derivative constraint and replace the measured loads, by permuting force–site pairs or by penalizing the joint derivative of the state-value head toward zero (Appendix E). Square brackets give 95% intervals from resampling whole held-out demonstrations.

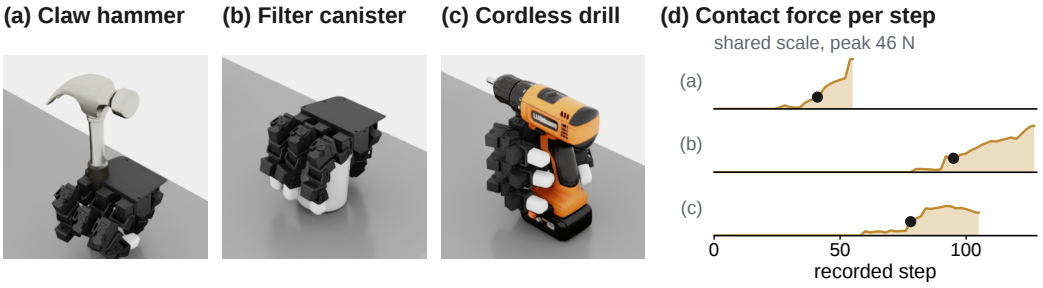

Figure 4: Recorded contact-rich demonstrations supply the contact-force supervision. (a)–(c) The recorded hand and object state of three demonstrations with objects of different geometry, each at three quarters of its demonstration. (d) The summed contact-site force of the same three demonstrations at every recorded step, on a shared scale; dots mark the displayed states.

Table 1: JFM improves on FP at every seed and measure, whereas FV's differences from FP change sign between seeds despite a four to five times lower auxiliary loss. Paired over seeds, JFM's return-error and return-rank gains are significant ($p = 0.027$ and $p = 0.008$). Entries are differences from FP trained from the same pretraining seed, on the same $8{,}000$ validation samples, and bold marks a difference that favors the method. The auxiliary-loss column gives each method's final auxiliary training loss relative to JFM's. Appendix C lists every interval.

		Auxiliary loss	Difference from FP		
Method	Seed	relative to JFM	Force error ↓	Return error ↓	Return rank ↑
FV	42	0.21	$+0.008$	$+0.693$	-0.021
	43	0.23	$+0.035$	$\mathbf{-0.562}$	$\mathbf{+0.208}$
	44	0.24	$\mathbf{-0.177}$	$+0.834$	$\mathbf{+0.017}$
JFM	42	1	$\mathbf{-0.086}$	$\mathbf{-0.277}$	$\mathbf{+0.061}$
	43	1	$\mathbf{-0.039}$	$\mathbf{-0.246}$	$\mathbf{+0.048}$
	44	1	$\mathbf{-0.117}$	$\mathbf{-0.153}$	$\mathbf{+0.064}$

Table 2: JFM's return error and force error stay below FP's on the 157 held-out demonstrations. Entries are differences on every contact state of these demonstrations at each seed, and bold marks a difference that favors the first method; Appendix C gives the means and intervals.

Comparison	Seed	Force error ↓	Return error ↓	Return rank ↑
FV − FP	42	$\mathbf{-0.037}$	$\mathbf{-1.380}$	-0.013
	43	$+0.051$	$+1.297$	$\mathbf{+0.049}$
	44	$\mathbf{-0.118}$	$+0.383$	-0.014
JFM − FP	42	$\mathbf{-0.001}$	$\mathbf{-2.207}$	-0.001
	43	$\mathbf{-0.019}$	$\mathbf{-0.063}$	$\mathbf{+0.066}$
	44	$\mathbf{-0.042}$	$\mathbf{-0.213}$	$\mathbf{+0.012}$

Closed-loop evaluation starts each policy from the 158 held-out initial states. Each attempt runs the policy for H steps, then removes the object's support while the hand holds its last command for 200 steps. The grasp succeeds when the object stays at least 0.2 m high, and a trajectory counts as covered when at least one of 32 attempts succeeds. The primary horizon $H = 100$ spans 96% of the demonstrated contact onsets. The main-study policies are evaluated as trained offline. A separately trained JFM and FP pair follows the earlier offline-to-online path of Appendix B and is fine-tuned online in simulation.

5.2 Value estimates and force prediction

JFM's critics give more accurate value estimates than FP's. They have lower return error and higher return rank at each seed on the validation samples, both significant when paired over the three pretraining seeds ($p = 0.027$ and $p = 0.008$, Table 1), and their return error is again lower at each seed on the 157 held-out demonstrations (Table 2). JFM also has the lowest centered return error of the three methods at each seed. JFM's return-rank advantage over FP is largest far from grasp completion, in the quarter of states with the most remaining steps, where FP ranks returns worst (Appendix C).

JFM predicts contact forces more accurately than FP after offline RL training. JFM's force error is below FP's at each seed on both evaluation sets, a consistency FV lacks (Tables 1 and 2), with a mean margin of 0.081 $[0.013, 0.188]$ on the validation samples. The gain holds under both joint-load targets: in the replication, which uses the stored target, JFM's mean force error over eight offline seeds is 1.418 against FP's 1.506 (paired $p = 0.014$, Table 6).

5.3 Source of the gain in value estimates

The energy-gradient requirement is the source of JFM's gain in value estimates. FV receives the same loads and reaches a four to five times lower auxiliary loss. Yet FV's critics have a higher mean return error than JFM's, by 0.55 $[0.22, 1.02]$ and 0.93 $[0.07, 2.38]$ on the two evaluation sets

Table 3: Of the three main-study policies, FV's covers the fewest held-out trajectories in closed loop despite a four to five times lower auxiliary loss than JFM's. The seed-42 policies, trained offline, cover most of the 158 trajectories, near the ceiling of the test. Covered: at least one of 32 attempts succeeds; H: policy steps before the support is removed; intervals from 10,000 paired trajectory resamples.

	Covered				
Horizon	JFM	FP	FV	JFM − FV	FP − FV
$H = 100$	149	148	144	$+5\ [0, 10]$	$+4\ [-1, 10]$
$H = 50$	142	141	133	$+9\ [3, 16]$	$+8\ [2, 14]$

(Appendix C). FV's differences from FP change sign between seeds, whereas every JFM difference favors JFM (Table 1). Proposition 1 explains how FV can reach a lower loss, since FV can also fit the non-conservative remainder.

Two controls in the replication show that the measured loads matter as well (Appendix E). Permuting force–site pairs keeps the energy-gradient requirement and reaches JFM's loss, yet it raises the critics' temporal-difference (TD) residual above JFM's in both runs. The TD residual measures how consistent a critic is with its own bootstrap target. Adding to FP a zero-target penalty on the joint derivative of the state-value head constrains a derivative without the measured loads, and FP's force error stays above JFM's. FV keeps the loads and removes the energy-gradient requirement, whereas these two controls keep a derivative constraint and remove the measured loads.

In closed loop, FV covers the fewest held-out trajectories of the three main-study policies at both horizons despite its lower auxiliary loss, significantly fewer than JFM at $H = 50$ ($p = 0.012$, Table 3). Matching the same loads through one energy thus gives the policy more grasps than fitting them as outputs.

5.4 Gain without the energy-gradient input

JFM's gain in value estimates survives zeroing the energy-gradient input of its critics at evaluation. Zeroing this input w moves JFM's action values, so the critic reads it, yet JFM keeps its return error and return rank (Table 4, Appendix C). The zeroed JFM critic also keeps the direction of each difference from FP, and averaged over seeds its return error on the validation samples stays below that of the zeroed FV critic, by $0.39\ [0.08, 0.89]$ (Table 11). The gain therefore lies in the trained critic, whose representation the JFM loss trains through the backbone Jacobian of Equation (8).

Table 4: JFM keeps its value estimates when the auxiliary input w is zeroed. Each row is one pretraining seed, and each entry gives the result with w intact / zeroed; Table 11 adds the comparisons with FP and FV.

	Validation samples		157 held-out demonstrations	
Seed	Return error	Return rank	Return error	Return rank
42	6.305 / 6.301	0.566 / 0.565	4.630 / 4.631	0.705 / 0.704
43	7.514 / 7.550	0.457 / 0.454	5.266 / 5.277	0.730 / 0.730
44	7.330 / 7.372	0.398 / 0.393	5.333 / 5.351	0.661 / 0.659

5.5 Closed-loop grasping

JFM's offline-to-online policy grasps the object on more held-out trajectories than FP's. At the primary horizon $H = 100$, JFM covers 33 of the 158 held-out trajectories against FP's 18, a significant paired difference of $15\ [5, 25]$ (exact McNemar $p = 0.006$). JFM solves 21 trajectories that FP misses, against 6 the other way, and both solve 12 (Table 5, Figure 5). At $H = 50$, JFM again covers more trajectories, 14 against 9. Unlike the main-study policies near the ceiling of the test (Table 3), this pair covers few enough trajectories for the two methods to separate. JFM reaches this coverage despite three pretraining epochs against FP's twenty and without the transition residual of Equation (3) (Appendix B).

Table 5: JFM's offline-to-online policy covers significantly more held-out trajectories than FP's (exact McNemar $p = 0.006$ at $H = 100$) and keeps most of those FP covers. Covered and H as in Table 3; the solved-by columns partition the 158 trajectories, and the interval comes from 10,000 paired resamples.

	Covered		Trajectories solved by				
Horizon	JFM	FP	both	JFM only	FP only	neither	Difference
$H = 100$ (primary)	33	18	12	21	6	119	$+15\ [5, 25]$
$H = 50$	14	9	6	8	3	141	$+5\ [-1, 12]$

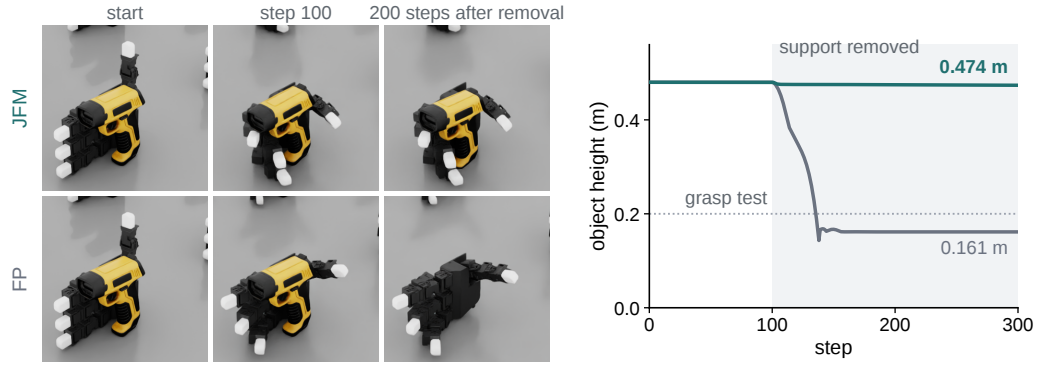

Figure 5: From the same start, JFM's grasp survives support removal and FP's releases the object. Offline-to-online policies on a held-out barcode-scanner trajectory that JFM covers and FP does not, at identical initial state and seed. The left frames show the start, step 100 and 200 steps after support removal, and the right panel shows object height over the same attempt, with shading after support removal and a dotted line at the 0.2 m grasp test.

6 Discussion and Conclusion

Contact data give a critic more than force values. Force prediction supervises what the representation outputs, and JFM also supervises its Jacobian with respect to the joint angles, through which the critic's values change as the hand moves, and ties this supervision across joints and hand poses to one scalar. FV can also fit the frictional remainder, which lowers its auxiliary loss without improving its critic, so auxiliary objectives for RL critics are best compared by the critics' value estimates.

JFM's target extends to other contacts, sensors and learners. It applies wherever kinematics maps measured forces into generalized coordinates, such as foot contacts, tool contacts and bimanual grasps. JFM needs no force input at evaluation, so demonstrations recorded with tactile gloves or fingertip sensors could train critics for hands without force sensing. The Hessian of the learned energy gives the conservative stiffness of a grasp and could score grasp stability or shape rewards. A structured head for the non-conservative remainder could let friction inform the critic as well, and the same target could train world-model dynamics and online critics.

Conclusion. This work asks which structure in contact forces improves an RL critic's value estimates in grasping, where existing methods ignore how the forces load the hand's joints. JFM supplies this mechanics: it turns measured forces into joint loads with the hand's Jacobians and asks one energy on the critic's representation to produce them as its negative joint-angle gradient, keeping force prediction and needing no force input at evaluation. This requirement makes the critic's value estimates more accurate than force prediction alone, lowering the return error at every pretraining seed; it also improves force prediction, and JFM's offline-to-online policy grasps significantly more held-out trajectories despite less pretraining. The gain comes from the requirement itself: a free-vector head fits the same loads four to five times more closely yet leaves its critic with a higher mean return error, and JFM's advantage survives when the critic stops reading the energy gradient at evaluation. Under frictional contact, the energy targets the conservative component of the joint loads, which brings force matching from molecular modeling to robot hands.

AI USE STATEMENT

We used large language model (LLM) assistants to aid and polish the writing, to retrieve and discover related work, and in research ideation and execution, including code, experiments and analyses. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Section 3 defines the model inputs, joint-load supervision, shared objectives, and gradient boundaries. Appendix A states the projection assumption, proofs, and coordinate conventions. Appendix B describes the data, reward, training budgets and bootstrap conditioning. Appendix C gives the main study with demonstration-level intervals, including the zeroing of the auxiliary input; Appendix D gives paired closed-loop trajectory coverage and Appendix E the supervision controls of the replication.

REFERENCES

Trevor Ablett, Oliver Limoyo, Adam Sigal, Affan Jilani, Jonathan Kelly, Kaleem Siddiqi, Francois Hogan, and Gregory Dudek. Multimodal and force-matched imitation learning with a see-through visuotactile sensor. *IEEE Transactions on Robotics*, 41:946–959, 2025. doi: 10.1109/TRO.2024.3521864.

Harsh Bhatia, Gregory Norgard, Valerio Pascucci, and Peer-Timo Bremer. The Helmholtz-Hodge decomposition—a survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(8):1386–1404, 2013. doi: 10.1109/TVCG.2012.316.

Roberto Calandra, Andrew Owens, Manu Upadhyaya, Wenzhen Yuan, Justin Lin, Edward H. Adelson, and Sergey Levine. The feeling of success: Does touch sensing help predict grasp outcomes? In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 314–323, 2017. URL https://proceedings.mlr.press/v78/calandra17a.html.

Claire Chen, Zhongchun Yu, Hojung Choi, Mark Cutkosky, and Jeannette Bohg. DexForce: Extracting force-informed actions from kinesthetic demonstrations for dexterous manipulation. *IEEE Robotics and Automation Letters*, 2025.

Jie Cheng, Ruixi Qiao, Yingwei Ma, Binhua Li, Gang Xiong, Qinghai Miao, Yongbin Li, and Yisheng Lv. Scaling offline model-based RL via jointly-optimized world-action model pretraining. In *International Conference on Learning Representations*, 2025.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017. doi: 10.1126/sciadv.1603015.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Furio Ercolessi and James B. Adams. Interatomic potentials from first-principles calculations: The force-matching method. *Europhysics Letters*, 26(8):583–588, 1994. doi: 10.1209/0295-5075/26/8/005.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A. Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pp. 158–168. PMLR, 2022.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Haotian He, Zeyu Yan, Qipeng Liu, Ning Guo, and Wenzhao Lian. FAWAM: Force-aware world action models for closed-loop contact-rich manipulation. *arXiv preprint arXiv:2606.08555*, 2026.

Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. FoAR: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

Carolina Higuera, Sergio Arnaud, Byron Boots, Mustafa Mukadam, Francois Robert Hogan, and Franziska Meier. Visuo-tactile world models. *arXiv preprint arXiv:2602.06001*, 2026.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL `http://jmlr.org/papers/v6/hyvarinen05a.html`.

Sergei Izvekov and Gregory A. Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005. doi: 10.1021/jp044629q.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Jaehwi Jang, Zhaoyuan Gu, Alfred Cueva, Zimeng Chai, Junjie Sheng, Thong Nguyen, Himank Galundia, Yifan Wu, Huishu Xue, Isaac Legene, Ojas Mediratta, Davin Doan, Andrew Collins, Sarah Sadegh, KyoungMok Kim, Rishita Dhalbisoi, Zun Chen, and Ye Zhao. WT-UMI: Tactile-based whole-body manipulation via force-supervised contact-aware planning. *arXiv preprint arXiv:2606.13232*, 2026.

Lei Jin, Yiding Ma, Xin Zhang, Chen Gao, Wei Wu, and Yong Li. TacWAM: Anchor-guided world action model with mechanics-aware tactile prediction. *arXiv preprint arXiv:2607.28391*, 2026.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *IEEE International Conference on Robotics and Automation*, 2019. doi: 10.1109/ICRA.2019.8793485.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Yang Li, Zhaxizhuoma, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, Qiaojun Yu, Cewu Lu, Yu Qiao, and Jiangmiao Pang. ForceVLA2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jason Jingzhou Liu, Yulong Li, Kenneth Shaw, Tony Tao, Ruslan Salakhutdinov, and Deepak Pathak. FACTR: Force-attending curriculum training for contact-rich policy learning. *arXiv preprint arXiv:2502.17432*, 2025.

Michael Lutter, Christian Ritter, and Jan Peters. Deep Lagrangian networks: Using physics as model prior for deep learning. In *International Conference on Learning Representations*, 2019.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1–9, 2021. URL `https://proceedings.mlr.press/v130/lyle21a.html`.

Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, Vinod Krishna, Sergei Izvekov, Gregory A. Voth, Avisek Das, and Hans C. Andersen. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics*, 128(24):244114, 2008. doi: 10.1063/1.2938860.

Domenico Prattichizzo and Jeffrey C. Trinkle. Grasping. In Bruno Siciliano and Oussama Khatib (eds.), *Springer Handbook of Robotics*, pp. 955–988. Springer, 2016.

Fabian Schramm, Nicolas Perrin-Gilbert, and Justin Carpentier. First-order Sobolev reinforcement learning. *arXiv preprint arXiv:2511.19165*, 2025.

Max Schwarzer, Ankesh Anand, Rishab Goel, R. Devon Hjelm, Aaron Courville, and Philip Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. LEAP hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In *Robotics: Science and Systems*, 2023.

Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science*, 5(5):755–767, 2019. doi: 10.1021/acscentsci.8b00913.

Yujie Zang, Yuhang Zheng, Xian Nie, Yupeng Zheng, Shuai Tian, Songen Gu, Chen Gao, Zining Wang, Shuicheng Yan, and Wenchao Ding. TacForeSight: Force-guided tactile world model for contact-rich manipulation. *arXiv preprint arXiv:2606.11184*, 2026.

A Structural analysis and implementation conventions

A.1 Fixed-measure gradient projection

We formulate the structural result as an orthogonal projection in a specified measure. Write $x = (q, u)$, where u fixes the velocity, action conditioning, contact flags, descriptors, and any derived features supplied independently to the network. The measure ρ is the distribution from which the rows are drawn: all rows during pretraining, where contact-free rows carry a zero target, and contact-active rows during offline learning. Let $\mathcal{V}$ be the linear space of admissible potentials of Section 4.1, and let $\mathcal{G}$ be the closure in $L^2(\rho; \mathbb{R}^{23})$ of the gradients $-\nabla_q v$ with $v \in \mathcal{V}$. $\mathcal{G}$ is a closed subspace by construction, and the orthogonal projection onto it exists and is unique. The measured target F^* is square-integrable. Here $\mathcal{V}$ and the measure ρ specify the projection.

The field entering the projection is the conditional mean of the force target when force records are stochastic at fixed x. The measured-target squared loss then adds the conditional force variance to the mean-field loss in (5). That variance is independent of the energy parameters, so the population-gradient statement is unchanged.

Projection assumption. We assume that the orthogonal projection of F^* onto $\mathcal{G}$ is attained by a scalar potential $\Phi \in \mathcal{V}$ with square-integrable joint gradient. Then

$$F^* = -\nabla_q \Phi + F^*_\perp, \qquad \langle F^*_\perp, \nabla_q v \rangle_{L^2(\rho)} = 0 \tag{9}$$

for every $v \in \mathcal{V}$. The measure weights both the fit and the orthogonality condition, so they describe the same learning objective. Each admissible energy, including every energy reachable by the parameterization, lies in $\mathcal{V}$. Potentials are identified through their joint gradients, with additive functions of the fixed conditioning variables left free.

A.2 Proof details for the projection identity

The proof is a squared-error decomposition. Orthogonality removes the cross-term, leaving an energy-dependent fit error and a parameter-independent residual. The same separation yields the population-gradient consequence below.

Proof of Proposition 1. Substituting the projection in Equation (9) into the residual gives

$$-\nabla_q E - F^* = -\nabla_q (E - \Phi) - F^*_\perp,$$
$$\mathcal{L}_{\mathrm{JFM}}(E) = \|\nabla_q (E - \Phi)\|^2_{L^2(\rho)} + \|F^*_\perp\|^2_{L^2(\rho)} + 2\langle \nabla_q (E - \Phi), F^*_\perp \rangle_{L^2(\rho)}.$$

Since $\mathcal{V}$ is linear and both E and Φ lie in it, $E - \Phi \in \mathcal{V}$; the final inner product is therefore zero by Equation (9), yielding Equation (6). □

Population-gradient consequence. Equation (7) follows by differentiating the same identity. Assume the energy is differentiable in its parameters. The residual norm contains no energy parameters for fixed F^* and ρ. Only the squared Sobolev semi-norm then contributes to the population gradient, provided differentiation may pass through the expectation.

The population identity preserves the distinction between optimization and sampling. The residual can have a nonzero inner product with a parameter derivative of the predicted gradient on a minibatch, so stochastic gradients can retain variation from the residual.

A.3 Derivation of the shared-Jacobian factorization

Equation (8) follows from the shared representation $z = \phi(x)$. Write $E = h(z, c)$ and $Q = g(z, a, w)$, with g selecting the summary token internally. The chain rule gives $\nabla_q E = (\partial\phi/\partial q)^\top \nabla_z h$ and $\partial_q Q = (\partial\phi/\partial q)^\top \nabla_z g$ at fixed u, a, w for differentiable heads and backbone.

The factorization identifies shared parameters with different output roles. The energy head combines summary and contact tokens, while the value head uses the summary token. Both derivatives contain the backbone Jacobian, with the auxiliary feature held fixed. The identity describes the partial derivative with the detached feature w held fixed; the stop-gradient keeps critic and policy updates from passing through w.

A.4 Coordinate and gradient conventions

The implementation differentiates in the full 23-dimensional coordinate input. Its first 16 entries are hand joint angles and its final seven entries represent the object's position and quaternion. Site Jacobians are computed from the hand joints and padded with zeros in the object block. Consequently, the implemented auxiliary term, before averaging over JFM rows, is

$$\frac{1}{23}\left(\left\|-\partial_{q_{\text{hand}}}E - F^{*}_{\text{hand}}\right\|^2 + \left\|\partial_{q_{\text{object}}}E\right\|^2\right). \tag{10}$$

The object term is identically zero in the evaluated input graph, because the encoder reads the hand angles from q and obtains object pose through independently supplied body features, and the zero object block supplies no additional flattening penalty.

Target conventions. Site Jacobians are computed in the hand-base frame, and the collector stores world-frame force vectors. The replication and its controls use the stored target $F^{*}_{\text{stored}} = \sum_i c_i J_{i,B}^{\top} f_{i,W}$. The main study, including FV, uses the target of Equation (1), $\sum_i c_i J_{i,B}^{\top} R_{WB}^{\top} f_{i,W}$, where R_{WB} maps hand-base vectors into world coordinates, so it evaluates the virtual-work target of Equation (1). Section 5.2 reports the force results of the two studies separately.

B Experimental setup details

Demonstrations and dataset. Human operators collect demonstrations in Isaac Lab using VR hand tracking retargeted to a four-finger LEAP hand with 16 actuated joints and a 6-DoF wrist. Contact flags and forces come from simulator sensors at 17 sites. The demonstrations cover 71 rigid objects. Each (object, pose, variant) tuple has five human demonstration trajectories; the median length is 79 steps and the 90th percentile is 181 steps. Four trajectories per tuple are used for training and one for validation, with disjoint trajectory identifiers. Eligible states belong to successful demonstrations and have at least two active contact sites. The eligible rows of a trajectory with T transitions run from row 30 through $T - \lfloor 0.15T \rfloor - 21$, inclusive (zero-based indexing), reserving a successor window after the contact segment. The main study draws its 8,000 validation samples from the eligible states with a fixed random seed of zero.

Architecture. The shared transformer has six layers, eight attention heads, and width 256. Its tokens encode kinematics, actions, object and environment descriptors, and contact sites. The energy combines a summary-token term with contact-gated site terms. The input-coordinate and derivative conventions are specified in Appendix A.4.

Reward. Offline reward labels come from a single physical reward. The contact term is the fraction of the 17 sites in contact at the next state, with weight 0.1. The lift term is the object's height in meters above the initial scene height recorded in the metadata, with weight 1.0. The squared change in per-site forces is penalized with weight 0.01. A stability term penalizes squared object displacement with weight 0.5, gated by a smooth ramp in lift height. Each step incurs a cost of 0.01. A bonus of 10 is assigned to the last transition of a trajectory that passes the grasp test: after the supporting platform is removed and the hand holds its last commanded joint target for 200 simulator steps, the object's height is at least 0.2 m. We use this reward throughout and add no energy-based shaping term.

World-model pretraining (Stage 0). Pretraining uses 20 epochs with $\lambda_s = 1.0$, $\lambda_c = 0.1$, $\lambda_f = 0.05$, and $\lambda_{\text{jfm}} = 0.2$; FP sets $\lambda_{\text{jfm}} = 0$. Adam uses learning rate 3×10^{-4}, batch size 1024, and gradient-norm clip 1.0. The JFM loss averages all batch rows. The force head regresses current recorded forces during pretraining in the replication and successor forces in both stages for the main study. Predicted forces retain the recorded world frame. The main study repeats pretraining at seeds 42, 43 and 44, each setting the initialization and the data order, so every seed yields one JFM, one FV and one FP checkpoint. FV replaces the scalar energy head with a vector head that regresses the same joint-load target, matched to JFM's energy head in capacity to within ten parameters and with identical routing; FP receives no joint-load target. The offline-to-online policies of Section 5.5 follow an earlier path that uses the replication's joint-load target and transition convention, with 3 pretraining epochs for JFM and 20 for FP.

Offline RL training (Stage 1). The final pretrained checkpoint initializes the shared backbone. IQL heads are then attached and trained for 15 epochs at batch size 1024. Predictive losses remain active, with $\lambda_\pi = \lambda_q = 0.3$ and $\lambda_{\text{jfm}} = 1.0$. This stage uses successor force labels and averages JFM over contact-active rows. Adam uses learning rates 10^{-4} for world-model parameters and 3×10^{-4} for IQL heads. Gradient norms are clipped at 0.5, and the target-network update coefficient is 0.005. The return-learning objectives and hyperparameters are specified below.

The JFM coefficient increases from 0.2 to 1.0 when offline RL training begins. This gives derivative supervision five times its pretraining weight relative to each predictive loss. The schedule is fixed across the eight replication seeds, the force–site permutation controls, and the three main-study pretraining seeds, whose offline stage uses seed 0.

Algorithm 1 One JFM training step in the offline RL stage. Pretraining runs the same step over all rows, without the value and policy losses.

Input: transitions with model inputs x_t, site forces f_i^* and contact flags c_i

1: $F^* \leftarrow \sum_i c_i J_i(q_t)^\top f_i^*$ — joint loads, Equation (1)
2: $z \leftarrow \phi(x_t)$ and $E \leftarrow h_0(z) + \sum_i c_i h_i(z)$ — shared representation and energy
3: $g \leftarrow -\partial_q E$, kept in the computation graph — second-order backpropagation
4: $\widehat{\mathcal{L}}_{\text{JFM}} \leftarrow$ mean of $\|g - F^*\|_2^2$ over contact-active rows — Equation (2)
5: $w \leftarrow \text{sg}[g]$ and $\hat{q}_{t+1} \leftarrow q_t + \Delta_{\text{dyn}}(x_t) + \eta w$ — Equation (3)
6: state, contact and force losses, and IQL losses $\mathcal{L}_Q, \mathcal{L}_V, \mathcal{L}_\pi$ with input w
7: update ϕ and every head on $\mathcal{L}$ of Equation (4), and the target critic by a moving average

Predictive objectives. The predictive objectives used during offline RL training are

$$\begin{aligned} \mathcal{L}_{\text{state}} &= \text{MSE}(\hat{q}_{t+1}, \tilde{q}_{t+1}) + \text{MSE}(\hat{p}_{t+1}, \tilde{p}_{t+1}), \\ \mathcal{L}_{\text{contact}} &= \text{BCE}(\hat{c}_{t+1}, \tilde{c}_{t+1}), \qquad \mathcal{L}_{\text{force}} = \text{MSE}(\hat{f}_{t+1}, \tilde{f}_{t+1}). \end{aligned} \tag{11}$$

Mean squared error (MSE) averages the regression errors, and binary cross-entropy (BCE) supervises contact flags. Tildes identify the recorded successor labels supplied by the data loader. Here p collects hand-joint and object velocities.

Return-learning objectives. Writing the actual input conditioning as x, the losses are

$$\begin{aligned} \mathcal{L}_V &= \mathbb{E}[|\tau - \mathbf{1}\{v_t < 0\}| v_t^2], & v_t &= \text{sg}[Q_{\bar{\psi}}(x_t, a_t; w_t)] - V_\psi(x_t; w_t), \\ \mathcal{L}_Q &= \mathbb{E}[(Q_\psi(x_t, a_t; w_t) - y_t)^2], & y_t &= \text{sg}[r_t + \gamma(1 - d_t) V_\psi(\hat{x}_{t+1}; 0)], \\ \mathcal{L}_\pi &= -\mathbb{E}[\min(e^{\beta \hat{A}_t}, c_{\max}) \log \pi_\psi(a_t \mid x_t, w_t)], & \hat{A}_t &= \text{sg}[v_t]. \end{aligned} \tag{12}$$

The target critic $Q_{\bar{\psi}}$ is an exponential moving average of Q_ψ. We use expectile $\tau = 0.7$, temperature coefficient $\beta = 3$, weight cap $c_{\max} = 100$, and discount $\gamma = 0.99$.

Offline-to-online training (Stage 2). Fine-tuning uses 15,800 training steps, batch size 1024, and learning rate 10^{-4}. The policy collects a 20-step Isaac Sim rollout in 32 parallel environments every 50 training steps. The online fraction of each batch is zero during the first 5% of training. It rises linearly to 0.5 at the halfway point, stays there through 95% of training, and reaches 0.75 at the end. This stage applies to the two offline-to-online policies, and the main-study policies end at Stage 1.

Residual and gradient settings. The residual scale is fixed at $\eta = 1/2$ in every stage. The offline-to-online JFM policy is the one exception: it trains and acts with the transition residual of Equation (3) disabled. The auxiliary gradient uses second-order backpropagation in full precision. The transition residual and IQL conditioning receive detached copies, and target-critic readouts are also detached.

Policy observation interfaces. No trained policy receives measured forces at evaluation: JFM reads the simulator contact flags through its energy term, and FP receives a zero energy feature.

Encoder and bootstrap conditioning. The one-step bootstrap encodes predicted coordinates and velocities together with the recorded next action and recorded next body features. The bootstrap value head receives a zero energy-gradient feature. The force-input control also supplies the recorded next-step forces to the bootstrap value head.

Recorded transitions and force labels. The collectors apply a joint target, advance simulation, and then record the state and target, so the command recorded at the next row is the one applied between two records. The main study uses this command as the transition's action, keeps the current row's command as state context, and stores each episode's final successor explicitly. The replication and its controls use the stored-record convention, which takes the command at the current row and forms successors by shifting stored rows; the final transition of each episode is terminal, so no repeated row enters a bootstrap target.

Offline randomization. The eight replication runs of each method use offline seeds 0 to 7 for the IQL-head initialization and the model random stream, and run k in the tables denotes seed $k-1$. Each starts from one pretrained checkpoint per method with the replay sampler at a fixed initial seed, and runs are paired by seed.

Replication force results. JFM ends offline RL training with a lower held-out force error than FP, 1.418 against 1.506 (Table 6). JFM also has the lower contact training loss, and its state-prediction loss stays at FP's level, so the derivative target improves force prediction at no measurable cost in state prediction.

Table 6: Mean $\pm$ standard deviation across the eight replication seeds. Force error uses the eligible held-out states; contact and state losses are recorded training losses. Lower is better. Paired t tests use the training run as the unit.

Quantity	FP	JFM	Paired p
Force error	1.5058 ± 0.0750	1.4175 ± 0.0665	0.0144
Contact loss	0.0658 ± 0.0015	0.0642 ± 0.0005	0.0444
State-prediction loss	0.2543 ± 0.0070	0.2584 ± 0.0039	0.2207

C MAIN STUDY DETAILS

The main study trains JFM, FV and FP at each of the pretraining seeds 42, 43 and 44 with the same training budget and evaluates the nine models at fixed checkpoints on the same $8{,}000$ validation samples. JFM and FV match the joint-load target of Equation (1); FP has no joint-load target, so no auxiliary loss is defined for it. Intervals resample whole held-out demonstrations $4{,}000$ times at bootstrap seed 0 while keeping the pairing between models, and the mean over seeds resamples the same demonstrations for all of them. Paired t tests over the three pretraining seeds treat each training run as the unit, as in Table 6. On the validation samples they give $p = 0.027$ for JFM's lower return error and $p = 0.008$ for its higher return rank against FP. Return error and return rank compare each state's action value for the applied command with the discounted recorded return $G_t = \sum_{k \geq 0} \gamma^k r_{t+k}$ up to the episode boundary, with $\gamma = 0.99$. Table 7 reports the paired differences at each seed and their mean over seeds in every pairwise direction. Averaged over seeds, JFM's force error is below FP's and its return error below FV's, both with intervals excluding zero. FV's return error exceeds FP's at two of the three seeds, and FV's mean return-rank gain over FP comes from seed 43 ($+0.208$), against -0.021 and $+0.017$ at the other seeds, whereas JFM's lies between $+0.048$ and $+0.064$ at each seed. Remaining trajectory length correlates with recorded return at -0.609, so Table 10 repeats the return-rank comparison within its quartiles at each seed. JFM's advantage is largest in the quartile farthest from completion at each seed, where FP orders returns worst.

Table 7: Paired differences at each pretraining seed and their mean over seeds, with 95% intervals from resampling whole held-out demonstrations. Negative force-error and return-error differences and positive return-rank differences favor the first method. JFM keeps the favorable sign against FP on force error, return error and return rank at each seed, while FV's differences change sign between seeds.

Quantity	Seed	JFM − FP	JFM − FV	FV − FP
Force error	42	$-0.086\,[-0.331, +0.072]$	$-0.094\,[-0.272, +0.029]$	$+0.008\,[-0.278, +0.265]$
	43	$-0.039\,[-0.122, +0.034]$	$-0.075\,[-0.213, +0.008]$	$+0.035\,[-0.083, +0.221]$
	44	$-0.117\,[-0.265, -0.003]$	$+0.060\,[-0.017, +0.153]$	$-0.177\,[-0.358, -0.030]$
	Mean	$-0.081\,[-0.188, -0.013]$	$-0.036\,[-0.078, +0.001]$	$-0.044\,[-0.132, +0.022]$
Return error	42	$-0.277\,[-0.659, +0.437]$	$-0.971\,[-1.496, -0.362]$	$+0.693\,[+0.047, +1.541]$
	43	$-0.246\,[-0.603, +0.221]$	$+0.316\,[-2.072, +1.935]$	$-0.562\,[-2.365, +1.943]$
	44	$-0.153\,[-1.261, +1.670]$	$-0.986\,[-2.056, +0.487]$	$+0.834\,[+0.548, +1.549]$
	Mean	$-0.225\,[-0.773, +0.583]$	$-0.547\,[-1.015, -0.225]$	$+0.322\,[-0.303, +1.386]$
Return rank	42	$+0.061\,[-0.044, +0.148]$	$+0.082\,[-0.049, +0.187]$	$-0.021\,[-0.119, +0.072]$
	43	$+0.048\,[-0.065, +0.176]$	$-0.161\,[-0.389, +0.120]$	$+0.208\,[+0.000, +0.439]$
	44	$+0.064\,[-0.072, +0.203]$	$+0.047\,[-0.076, +0.199]$	$+0.017\,[-0.075, +0.094]$
	Mean	$+0.058\,[-0.020, +0.128]$	$-0.011\,[-0.060, +0.060]$	$+0.068\,[-0.023, +0.148]$

All held-out demonstrations. The same nine models are also read out on every transition of each successful held-out demonstration whose current state has at least two active contact sites, the final transition excluded, which gives $7{,}151$ states from 157 demonstrations. Table 8 lists every difference from FP and the mean over seeds. JFM leads in eight of the nine comparisons and FV in four, and averaged over seeds FV's return error exceeds JFM's with an interval excluding zero. On these states JFM's centered return error, defined below, again beats both other methods at each seed: 4.45, 4.86 and 4.99 against 6.71, 4.87 and 5.33 for FP and 5.27, 6.50 and 5.44 for FV. Zeroing the auxiliary input w moves JFM's return error on these states by at most 0.018 at any seed (Table 11). On force error, every JFM difference from FP is negative on both evaluation sets, while FV's differences take both signs.

Table 8: Paired differences on every contact state of all 157 held-out demonstrations, at each seed and as the mean over seeds with 95% intervals from resampling whole demonstrations. Negative force-error and return-error differences and positive return-rank differences favor the first method.

Comparison	Seed	Force error	Return error	Return rank
JFM − FP	42	-0.001	-2.207	-0.001
	43	-0.019	-0.063	$+0.066$
	44	-0.042	-0.213	$+0.012$
	Mean	$-0.020\,[-0.081, +0.026]$	$-0.828\,[-2.091, +0.157]$	$+0.026\,[-0.005, +0.059]$
FV − FP	42	-0.037	-1.380	-0.013
	43	$+0.051$	$+1.297$	$+0.049$
	44	-0.118	$+0.383$	-0.014
	Mean	$-0.035\,[-0.137, +0.038]$	$+0.100\,[-0.202, +0.513]$	$+0.007\,[-0.027, +0.041]$
JFM − FV	Mean	$+0.014\,[-0.035, +0.068]$	$-0.928\,[-2.385, -0.072]$	$+0.018\,[-0.005, +0.049]$

Return scale and offset decomposition. Once each critic's mean offset is removed, JFM's centered return error is the lowest of the three methods at each seed (Table 9). The expectile bootstrap of Appendix B places the values of every critic above the recorded returns, by 3.5 to 5.3 on these samples. For each critic, let b be the mean of $Q - G$ over these samples; its squared return error equals b^2 plus the mean of $(Q - G - b)^2$, and Table 9 reports the return error, mean offset and centered return error at each seed.

Table 9: Return error decomposed into offset and centered error, on the same $8{,}000$ validation samples, whose returns have mean 0.181. The offset is the mean of $Q - G$; the centered return error removes it. Lower error and higher return rank indicate more accurate value estimates.

Seed	Method	Mean Q	SD Q	Offset	Centered error	Return error	Return rank
42	FP	4.021	4.697	3.840	5.347	6.583	0.505
	FV	3.802	6.500	3.621	6.311	7.276	0.484
	JFM	3.866	4.202	3.685	5.117	6.305	0.566
43	FP	5.056	2.948	4.875	6.037	7.759	0.410
	FV	3.645	7.842	3.464	6.309	7.197	0.618
	JFM	5.055	3.482	4.874	5.719	7.514	0.457
44	FP	4.346	3.595	4.165	6.216	7.482	0.333
	FV	5.466	2.695	5.285	6.421	8.316	0.350
	JFM	4.136	4.386	3.955	6.171	7.330	0.398

Table 10: JFM minus FP return rank within quartiles of remaining trajectory length, at each main-study seed. Intervals resample whole held-out demonstrations 4,000 times while keeping the pairing between methods.

Remaining steps	n	Seed 42	Seed 43	Seed 44
29–56	1,924	$+0.061$ $[-0.141, +0.283]$	$+0.093$ $[-0.113, +0.297]$	-0.067 $[-0.198, +0.054]$
57–84	2,048	$+0.146$ $[-0.016, +0.319]$	$+0.250$ $[+0.009, +0.501]$	$+0.004$ $[-0.152, +0.187]$
85–135	2,018	$+0.109$ $[-0.031, +0.263]$	$+0.068$ $[-0.126, +0.304]$	-0.070 $[-0.179, +0.027]$
136–353	2,010	$+0.266$ $[-0.059, +0.396]$	$+0.380$ $[-0.136, +0.542]$	$+0.478$ $[-0.394, +0.675]$

Zeroing the auxiliary input. All nine main-study checkpoints are read out on both evaluation sets with the auxiliary input w of their critics unchanged or set to zero. Each pair of readouts uses the same encoded features and applied action; no parameters are updated, and FP's auxiliary input is already zero. Table 11 reports JFM's return error and return rank before and after zeroing and the paired differences of the zeroed critic at each seed and as the mean over seeds, with intervals for the mean from 4,000 demonstration resamples shared by the three seeds. The zeroed JFM critic keeps the direction of each difference the intact critic has from FP at each seed and on both sets. Zeroing does move the outputs, with a root-mean-square Q change at seed 42 of 0.186 for JFM and 0.317 for FV. The input w is therefore active, and the advantage that survives its removal belongs to the trained critics.

Table 11: Zeroing the auxiliary input w at each main-study seed. Error is return error and rank is return rank; lower error and higher rank indicate more accurate value estimates, and each JFM entry gives the intact and the zeroed result. Paired differences use the zeroed JFM critic (JFM_0), FP and the zeroed FV critic (FV_0); the mean rows give 95% intervals from resampling whole held-out demonstrations.

	JFM, intact / zeroed		$\text{JFM}_0 - \text{FP}$		$\text{JFM}_0 - \text{FV}_0$
Seed	Error	Rank	Error	Rank	Error
Validation samples					
42	6.305/6.301	0.566/0.565	-0.281	$+0.060$	-0.948
43	7.514/7.550	0.457/0.454	-0.209	$+0.045$	$+0.716$
44	7.330/7.372	0.398/0.393	-0.110	$+0.060$	-0.940
Mean			-0.200	$+0.055$	-0.391
			$[-0.716, +0.578]$	$[-0.020, +0.123]$	$[-0.889, -0.080]$
All held-out demonstrations					
42	4.630/4.631	0.705/0.704	-2.206	-0.002	-0.805
43	5.266/5.277	0.730/0.730	-0.052	$+0.066$	-1.081
44	5.333/5.351	0.661/0.659	-0.195	$+0.009$	-0.568
Mean			-0.818	$+0.024$	-0.818
			$[-2.101, +0.159]$	$[-0.006, +0.058]$	$[-2.196, +0.010]$

D CLOSED-LOOP TRAJECTORY COVERAGE

Coverage counts trajectories with at least one successful grasp among 32 attempts on the same 158 held-out trajectories. Demonstrations first make contact after a median of 38 steps, and 96% of them do so within the primary horizon $H = 100$. Results are grouped by the policies' training histories.

Offline-to-online policies. These policies are pretrained, trained offline, and then fine-tuned online through the Stage 2 procedure of Appendix B, with 3 pretraining epochs for JFM and 20 for FP. They cover 33 and 18 trajectories at $H = 100$: 12 are solved by both, 21 by JFM alone, and 6 by FP alone. The paired coverage difference is 15 trajectories, with 95% interval $[5, 25]$ and two-sided exact McNemar $p = 0.0059$. Coverage at $H = 50$ is 14 and 9, with 6 solved by both, 8 by JFM alone, and 3 by FP alone; the difference is 5 trajectories, with interval $[-1, 12]$ and exact McNemar $p = 0.227$.

Main-study offline policies. These are the seed-42 Stage 1 policies of Appendix C, evaluated under the same grasp test without the online stage. All three methods are evaluated at both horizons on the same 158 trajectories, each with 32 attempts. Coverage is 149, 148 and 144 for JFM, FP and FV at $H = 100$, and 142, 141 and 133 at $H = 50$. Table 12 lists every paired comparison.

Table 12: Paired closed-loop coverage for the seed-42 main-study policies, evaluated without the online stage. Counts are trajectories, of 158. A positive difference favors the first method. $H = 100$ is the primary horizon.

Horizon	Comparison	Both	Only 1st	Only 2nd	Neither	Diff. [95%]	McNemar p
$H = 100$	JFM − FP	144	5	4	5	$+1$ $[-5, +7]$	1
	JFM − FV	143	6	1	8	$+5$ $[0, +10]$	0.125
	FP − FV	142	6	2	8	$+4$ $[-1, +10]$	0.289
$H = 50$	JFM − FP	137	5	4	12	$+1$ $[-5, +7]$	1
	JFM − FV	132	10	1	15	$+9$ $[+3, +16]$	0.012
	FP − FV	132	9	1	16	$+8$ $[+2, +14]$	0.021

FV trails both other methods at both horizons, significantly so at $H = 50$ ($+9$ $[+3, +16]$ against JFM and $+8$ $[+2, +14]$ against FP), even though its auxiliary loss is four to five times lower than JFM's (Table 1). Its lower auxiliary loss thus comes with lower closed-loop coverage, matching its higher return error in Section 5.3.

Intervals use $10{,}000$ paired bootstrap resamples of the 158 trajectory indicators at bootstrap seed 0 and are expressed in numbers of covered trajectories.

E SUPERVISION-CONTROL DETAILS

These controls vary the supervision itself within the replication, and each is read against the reference in its own row. A temporal-difference (TD) residual compares a critic with its own bootstrap target; since that target differs between methods, each TD residual is read against the same method's reference.

One of them appends the measured forces to the IQL heads and so supplies force to the value function as an input. The measure is the within-run interaction $(M_{\mathrm{JFM+F}} - M_{\mathrm{JFM}}) - (M_{\mathrm{FP+F}} - M_{\mathrm{FP}})$, where M is the mean squared TD residual. It is -10.49 and -2.62 in the two runs, and -10.33 and -2.58 with the energy input zeroed in the JFM cells. Adding a raw-force channel thus costs FP more consistency with its own bootstrap target than it costs JFM, the direction expected once the derivative target has placed that information in the shared representation.

The auxiliary controls have two training runs per condition, each compared with its own reference. The permuted runs reach final auxiliary losses of 0.001605 and 0.001575 against 0.001608 and 0.001614 for their JFM references, so the permutation changes the target field and keeps the fit as close as JFM's. Zeroing the auxiliary input w at inference changes JFM's mean squared TD residual by -4.3, $+0.6$, -0.3, 0.0, $+2.0$, -6.0, -1.6 and -0.7 percent in the eight replication runs. Each value-gradient penalty coefficient was set at training entry, before any result of the penalty runs was observed, so that the penalty's loss contribution equals 1.6×10^{-3}, the contribution of the JFM term; the resulting penalty coefficients are 16.84 and 11.31. Table 13 lists every intervened and reference number behind these contrasts, including the held-out force errors of the value-gradient penalty.

Table 13: Supervision controls in the replication, two runs each, written as run 1 / run 2. Each control is read against the reference entries in its own row. Entries are mean squared temporal-difference residuals except where the measure is force error; the interaction is JFM's change minus FP's, so it needs no separate reference.

			Reference	
Control	Measure	Result	JFM	FP
Force input to IQL heads	Interaction	-10.49 / -2.62		
	Interaction, w zeroed	-10.33 / -2.58		
Permuted force–site pairing	TD residual MSE	8.49 / 91.75	2.82 / 1.34	1.34 / 0.29
Value-gradient penalty	TD residual MSE	1.37 / 4.69		1.34 / 0.29
	Force error	1.4285 / 1.4601	1.4108 / 1.4166	1.4689 / 1.4897