CONSERVATION BUYS STABILITY AND FACTORING BUYS COUNTERFACTUALS IN PHYSICAL WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A learned simulator can reproduce its training conditions accurately yet fail in two distinct ways once those conditions change. Over long rollouts, small errors accumulate until the trajectory drifts away from physically plausible behavior; under an intervention on a physical parameter, the model may continue to follow the law seen during training rather than the intervened one. We show that these two failures require *different structural remedies*. Evolving a learned energy with a symplectic integrator preserves the geometry of the conservative dynamics and keeps rollouts bounded and physically meaningful for up to $100\times$ the training horizon, while equal-capacity predictors, an energy-regularized predictor, and a tuned neural ODE diverge. By contrast, encoding the physical coupling through an explicit linear factorization enables the model to follow a never-seen sign of that coupling, whereas an unrestricted parameterization remains locked to the training law. Crucially, the two mechanisms are separable: removing the structure responsible for long-horizon stability leaves counterfactual transfer intact, while removing the factorized coupling destroys counterfactual transfer without eliminating stability. This *double dissociation*, established with matched controls that remove or replace one structural component at a time, persists beyond the headline three-body system and remains visible when the physical state must be inferred from pixels rather than provided directly. The result is a concrete design principle for physical world models: long-horizon stability and changed-law generalization arise from distinct structural commitments, and each can be imposed deliberately without requiring the other.

1 INTRODUCTION

A learned simulator predicts the next state of a physical system and iterates that prediction into a trajectory. Modern models can perform well by conventional measures such as rollout accuracy or visual realism, yet they often fail in two distinct ways once the governing conditions move beyond training. First, prediction errors accumulate until long rollouts drift or become unstable. Second, when the physical law itself changes, for example through reversed gravity, a flipped coupling, or reversed time, the model often continues to follow the law seen during training. Both failures remain visible in current systems. The best reported video generator reaches only 22% joint accuracy on the hard split of VideoPhy-2, no model exceeds $3.3/5$ on PhyGround, and PhyWorldBench includes a dedicated "anti-physics" category for changed law failures (Yi et al., 2020; Bear et al., 2021; Gu et al., 2025; Bansal et al., 2025; Nam et al., 2026). Related limitations also appear in learned simulators across scales (Sanchez-Gonzalez et al., 2020; Zhong et al., 2021a). This paper asks: *which piece of physical structure fixes which failure?*

We address this question by encoding physical invariances directly rather than asking the model to infer them from finite data. The ingredients are individually established, including Hamiltonian and Lagrangian neural networks, generalized coordinate models, constrained dynamics, and symplectic structure preserving parameterizations (Greydanus et al., 2019; Cranmer et al., 2020; Lutter et al., 2019; Finzi et al., 2020; Zhong et al., 2020b). What remains unclear is which structural commitment produces which behavior and where each commitment stops helping. We show a *double dissociation*. Conserving a learned energy through a symplectic update keeps long rollouts physi-

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.

cally bounded but does not determine how the system should respond to a changed law. Conversely, making an interaction force enter linearly through its physical coupling enables transfer to a never seen sign of that coupling but does not prevent long horizon drift. Matched controls that remove or replace one ingredient at a time show that stability depends on geometric and energetic consistency, whereas counterfactual transfer depends on how the intervened parameter enters the dynamics.

This paper makes four contributions. First, we establish this double dissociation with matched controls, showing that conservation produces long horizon stability and a factored coupling produces counterfactual inversion, while neither produces the other and neither is explained by capacity, integrator choice, or formalism alone. Second, we formalize the mechanism behind each behavior and map the boundary at which each prior stops helping. Third, we show that the dissociation extends beyond smooth few body motion to generalized coordinates, contact, many body systems, mixed force families, and pixel based settings with grounded perception, including a learned object binder and frozen public video encoders. Fourth, we extend the same structural account beyond conservative dynamics, using a passive friction port that reaches the correct temperature and a factored drive that generalizes across unseen forcing strengths and signs.

We study a learned transition map from physical state, including positions and momenta, to the next state and iterate it into a trajectory. To separate dynamics from perception, we provide state in three ways: as oracle state; as state recovered from synthetic renders by an encoder whose position channel is anchored by either a fixed renderer or a label free centroid; and as state discovered from pixels with no anchor. The first two settings recover the target behaviors, while the third exposes a boundary of the current approach. Accordingly, this study concerns physical law fidelity rather than visual fidelity and does not claim improvements in real video generation, image quality, or leaderboard performance.

2 Related work

Structured dynamics and parameter conditioning. Hamiltonian and Lagrangian networks encode conservation laws or variational structure (Greydanus et al., 2019; Cranmer et al., 2020; Lutter et al., 2019). Generalised coordinate, constrained, and intrinsically symplectic variants extend these ideas to rigid and structured systems (Finzi et al., 2020; Zhong et al., 2020b; Jin et al., 2020), with continuous conservation statements for the learned energy and discrete behaviour determined in part by the numerical integrator (Zhong et al., 2021a). Conditioning dynamics on physical parameters, including coupling interactions through multiplicative coefficients, is also standard (Greydanus et al., 2019; Cranmer et al., 2020; Lutter et al., 2019; Zhong et al., 2020b; Jin et al., 2020; Brunton et al., 2016; Han et al., 2021; Raissi et al., 2019), while context adaptation extends prediction to new parameter draws within the training family (Kirchmeyer et al., 2022; Yin et al., 2021; Wang et al., 2022). Our setting differs in two respects. First, we intervene on a coupling whose sign lies outside the training support. Second, we compare matched factored and unrestricted versions of HNN, DeLaN, and SymODEN to determine which structural commitment supports stability and which supports counterfactual inversion.

Simulators, dissipation, contacts, and perception. Graph network simulators are strong relational predictors (Sanchez-Gonzalez et al., 2020; Battaglia et al., 2016; Satorras et al., 2021), and training noise is a common remedy for rollout instability; in the simulator studied here, it delays but does not prevent unbounded drift. Dissipative, port Hamiltonian, and GENERIC or metriplectic models separate reversible from irreversible dynamics (Sosanya & Greydanus, 2022; Zhong et al., 2020a; Desai et al., 2021; Lee et al., 2021). D-HNN evaluates unseen friction coefficients, port Hamiltonian networks recover energy, forcing, and dissipation, and compositional port Hamiltonian learning assembles systems from learned components (Neary & Topcu, 2023). Differentiable contact models further extend structured dynamics to collisions (Zhong et al., 2021b; Hochlehnert et al., 2021). On the perception side, Hamiltonian generative models and unsupervised Lagrangian models learn dynamics directly from images (Toth et al., 2020; Zhong & Leonard, 2020), while Botev et al. (2021) identify continuous time reversibility as the most broadly useful physical prior for pixel based learning. We show that structure alone does not induce a canonical latent representation, and that reversibility helps only when the underlying dynamics are themselves reversible. While recent video world models and physical reasoning benchmarks report persistent inconsistencies with physical laws (Yi et al., 2020; Bear et al., 2021; Riochet et al., 2018; Gu et al., 2025; Assran et al.,

2025; Bansal et al., 2025; Nam et al., 2026; Lin et al., 2026), our work isolates the specific structural commitments required to guarantee long-horizon stability and counterfactual generalisation.

3 METHODS

3.1 SETUP AND THE TWO BEHAVIOURS WE MEASURE

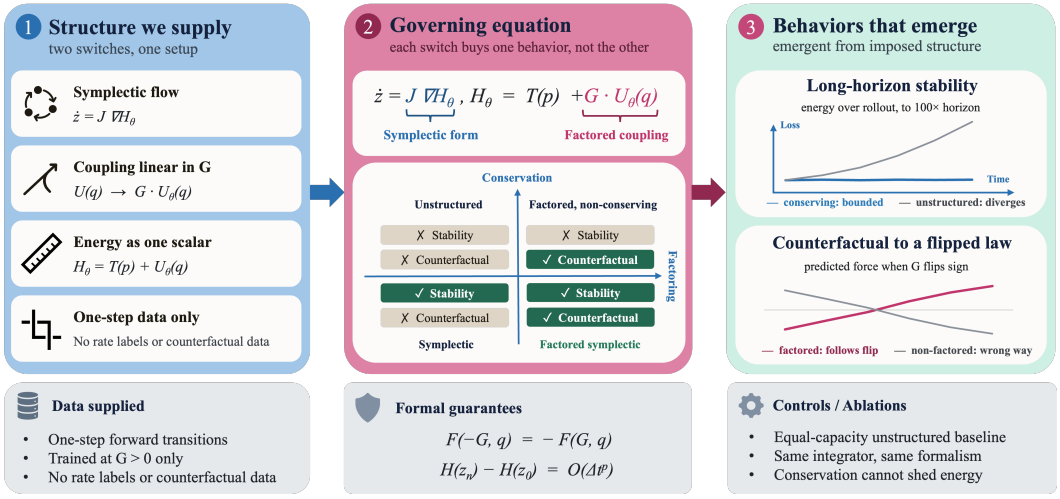


Figure 1: Overview of the two pieces of physical structure we impose (left), the double dissociation they produce (centre), and how each behaviour is measured and bounded (right and bottom).

Figure 1 summarizes the study. Left: we impose two structural commitments—energy conservation (evolving a learned energy H_θ with a symplectic step) and factored coupling (the intervened parameter G enters the force linearly, with all other content learned from forward-law data). Center: a double dissociation—conservation yields column-wise stability, factoring yields row-wise changed-law response, so only the model with both is stable *and* answers the counterfactual. Right/bottom: measurements, matched controls that rule out capacity, integrator choice, and formalism, and the limits where each structure stops helping. Below we formalize each component. For N bodies, the state is $x = (q, p)$ with positions q and momenta p, and parameters $c = (G, m_1, \dots)$ include the coupling G and masses. A dynamics model is a transition map iterated into a rollout,

$$x_{t+1} = f_\theta(x_t, c), \qquad x_{0:T} = \big(x_0, f_\theta(x_0, c), \dots\big), \tag{1}$$

and is trained only on one step prediction from forward law data with attractive gravity, $G > 0$. Our headline system is softened three body gravity,

$$V(q) = -\sum_{i<j} \frac{G m_i m_j}{\sqrt{\|q_i - q_j\|^2 + \epsilon^2}},$$

with evaluation rollouts of $T = 2000$ steps, one hundred times the training horizon. The remaining systems, including Coulomb charges, colliding disks, linear drag, a double pendulum in generalised coordinates, a confined thermal bath, and driven particles, together with the datasets, optimiser, and training settings, are specified in App. A.

We ask two questions of the trained transition map. The first is long horizon stability, measured by the relative drift of the true energy along a finite rollout,

$$D(t) = \frac{|E(x_t) - E(x_0)|}{|E(x_0)|}, \qquad t \le T, \tag{2}$$

where E is fixed by the reference Hamiltonian. We report both the per episode median and the worst episode, and count nonfinite trajectories explicitly rather than assigning them an arbitrary numerical value. The second behaviour is counterfactual inversion. We train only on attractive gravity, $G > 0$,

and evaluate the unseen repulsive regime, $G < 0$. Let $\hat{x}_{0:T}$ be the model's rollout, and let $x^-_{0:T}$ and $x^+_{0:T}$ be the ground-truth rollouts under the flipped repulsive law ($G < 0$) and the trained attractive law ($G > 0$), both started from the same true initial state so that initialisation error is counted rather than cancelled. The regime-match indicator asks which of these two references the model's rollout lies closer to,

$$\mathrm{RM} = \mathbb{1}\left[\|\hat{x}_{0:T} - x^-_{0:T}\| < \|\hat{x}_{0:T} - x^+_{0:T}\|\right], \tag{3}$$

so $\mathrm{RM} = 1$ when the rollout follows the changed law it never saw in training and $\mathrm{RM} = 0$ when it stays with the training law. Alongside it we report a normalized trajectory error, the squared error against the flipped-law reference $x^-_{0:T}$ divided by that reference's variance (Def. 1).

3.2 Two structural commitments, and what each guarantees

A symplectic model evolves a learned Hamiltonian H_θ through a symplectic update. In continuous time, the corresponding flow

$$\dot{q} = \frac{\partial H_\theta}{\partial p}, \qquad \dot{p} = -\frac{\partial H_\theta}{\partial q} \tag{4}$$

conserves H_θ. The discrete leapfrog update, however, conserves neither H_θ nor the reference energy exactly.

Proposition 1 (Near conservation of a modified Hamiltonian; formal statement in App. B)**.** *Under the regularity, bounded orbit, and small step assumptions of backward error analysis, leapfrog applied to Eq. equation 4 preserves a modified Hamiltonian $\widetilde{H}_\theta = H_\theta + O(dt^2)$ up to exponentially small error over exponentially long times (Hairer et al., 2006).*

This statement concerns $\widetilde{H}_\theta$, not the physical reference energy. Moreover, a generic MLP potential need not be coercive, so boundedness of the true energy remains an empirical finite horizon property. We therefore report learned energy conservation, physical energy drift, state escape, and trajectory fidelity as separate outcomes.

The second commitment is a coupling factored potential, which makes the interaction force exactly linear in G:

$$U_\theta(q) = \sum_{i<j} \varphi_\theta(q_i - q_j), \qquad V(q, c) = G\, U_\theta(q), \qquad F_i = -G\, \partial_{q_i} U_\theta(q), \tag{5}$$

where φ_θ is learned using attractive data only.

Proposition 2 (Sign inversion by construction)**.** *For any learned φ_θ and any G independent kinetic energy, the force in Eq. equation 5 is odd in G, $F_i(q, -G) = -F_i(q, G)$ for all q. The dynamics at $-G$ therefore follow the same learned interaction under the reversed coupling without requiring training examples from the opposite sign regime.*

The model family supplies the sign inversion, while the learned interaction φ_θ must still recover the correct magnitude and state dependence. Counterfactual accuracy therefore holds only where the learned content remains accurate on the states visited by the flipped rollout. By contrast, a nonfactored model conditions a black box potential $V_\theta(q, G)$ on G without imposing any relation between $G > 0$ and $G < 0$. Section 4.2 shows that such models can represent the flipped law but do not learn to select it from forward law data alone. Proofs are given in App. B.

3.3 Controls, and how claims are measured

We test each claim with ablations that remove or replace one structural ingredient at a time, attributing behaviour to the required component. For conservation, we compare capacity-matched baselines: an unstructured predictor, a fixed physical-energy penalty, and a tuned neural ODE, to separate the roles of model capacity, objective, and integration. For factoring, we compare a black box that receives the coupling without its linear form and a model that injects it via positive localized features, to test whether inversion arises from capacity or conditioning alone. We repeat these tests across Hamiltonian, Lagrangian, and symplectic-ODE formalisms and a graph-network simulator. In chaotic regimes, primary metrics are energy drift (Eq. equation 2) and regime match (Eq. equation 3); trajectory error is secondary, and divergence is distinguished from finite error. We

repeat experiments across independent initializations, reporting uncertainty via error bars, standard deviations, or p-values; near-binary outcomes are assessed per complete run. Results are assigned to one of four pre-registered evidence classes (well-supported, directional, supported, boundary). Full model lists, matching, statistics, and per-run results appear in App. A, App. C, and App. E.

4 RESULTS

In this section we first show that energy conservation stabilises long rollouts but does not answer a changed law. We then show that a factored coupling answers the changed law but does not provide stability. We combine both effects in a single dissociation table, rule out the main confounds, map where each structural prior stops helping, and finally show that the pattern extends beyond the headline system.

4.1 CONSERVATION MAKES LONG ROLLOUTS STABLE, BUT CANNOT ANSWER A CHANGED LAW

Table 1: Relative true-energy drift $D(t)$ over the rollout.

Model	t=100	t=500	t=2000	NF
Unstructured MLP (one-step map)	2.6	$>10^4$	$>10^4$	61
Energy-penalised MLP, λ=1	3.3	$>10^4$	$>10^4$	123
Energy-penalised MLP, λ=100	0.011	0.33	115–$>10^4$	0
Neural ODE (RK4, matched dt, tuned)	0.08	3×10^2–4×10^3	$>10^4$	0
Symplectic (ours)	0.40	2.7	7.1 [6.7, 7.7]	0

The first question is whether conserving a learned energy stabilises long rollouts where alternative approaches do not. We train the five models in Table 1 using one-step prediction on attractive-gravity trajectories and roll each model out for 2000 steps. Every reported value comes from a single raw record that retains the finite status of each episode (App. C.1). In the table, $>10^4$ denotes a per-episode median above the display cap, while NF counts trajectories that become non-finite by step 2000, out of 192 episodes.

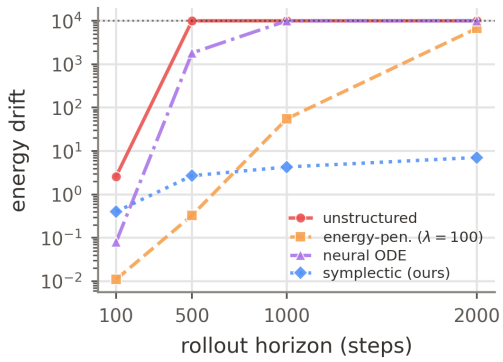


Figure 2: Long-horizon energy drift (log scale, capped at 10^4): only the symplectic model stays bounded, while the λ=100 penalty and the neural ODE reach lower drift at short horizons before diverging.

The symplectic model is the only one that stays finite and below the 10^4 cap in all episodes (Fig. 2), but finiteness does not imply accuracy. Its true-energy drift increases from 0.40 at 100 steps to 7.1 at 2000 (710% error) because Prop. 1 guarantees near-conservation for a modified learned Hamiltonian, not the physical energy. The one-step MLP exceeds the cap in the median episode by step 500, and 61/192 trajectories become non-finite between steps 1000 and 2000. Adding an energy penalty ($\lambda = 1$) does not help: 123/192 episodes become non-finite. Because non-finite trajectories have undefined (not infinite) error, we report non-finite counts, energy drift, and rollout error separately rather than as a single ratio. Thus, the comparison is finite rollouts with large drift versus divergence, and the cap should not be treated as a measured lower bound.

The neural ODE is the stronger control: at matched step size, objective, capacity, and learning-rate sweep, every trajectory stays finite and its 100-step drift (0.08) is five times smaller than the symplectic model's, yet by 500 steps it reaches 3×10^2–4×10^3 and exceeds the cap by 1000. The short horizon favours the neural ODE, the long horizon favours Hamiltonian structure, and since equal step size does not imply equal integration

error for RK4 and leapfrog, the comparison is between complete updates rather than integrators in isolation.

The most informative control is the energy-penalised model, whose target is the analytic physical energy: its short-horizon drift is far smaller than the symplectic model's, yet grows without bound by 2000 steps (with trajectory error following), so penalising the true energy delays drift without bounding it, and the two are always reported together. Two further probes on the factored model sharpen this picture: a short-rollout loss more than halves the 2000-step drift while preserving inversion, and over 10,000 steps the factored model's drift saturates from 1.9 toward 0.89 while the non-factored symplectic model keeps drifting.

Stability is nevertheless all that conservation provides. Trained only on attractive gravity, the same symplectic model does not follow the unseen repulsive law. Its rollout remains closer to the attractive world, with regime match 0. We turn to that second behavior next.

4.2 A FACTORED COUPLING ANSWERS A CHANGED LAW, BUT CANNOT MAKE ROLLOUTS STABLE

Table 2: Following a never-trained sign flip of the coupling to $G \in \{-0.5, -1.0, -1.5\}$.

Model	regime match	rollout error (nMSE)
factored symplectic	1 at every G	0.03–0.18
non-factored symplectic	0 at every G	$\approx$2.7

We next ask whether a model trained only on attractive gravity can follow the unseen repulsive law (Table 2). Regime match measures the fraction of rollouts that follow the repulsive rather than the attractive reference, while rollout error is measured against the true repulsive trajectory. The factored model follows the flipped-law regime at every tested coupling, with regime match 1 and nMSE 0.03–0.18 for $G \in \{-0.5, -1.0, -1.5\}$. In contrast, the non-factored symplectic model remains bounded but never inverts, with regime match 0 at every coupling and nMSE near 2.7. Its rollout remains closer to the attractive reference. At $G = -1$, only the factored model falls below nMSE 1, by four orders of magnitude (Fig. 3). The mechanism is Prop. 2: because the force is linear in G, the sign reversal is supplied by construction, while the learned interaction shape φ_θ determines the magnitude of the repulsive trajectory. The non-factored model fails not because it lacks capacity, but because its black-box potential imposes no relation between the unseen $G < 0$ regime and the observed $G > 0$ regime.

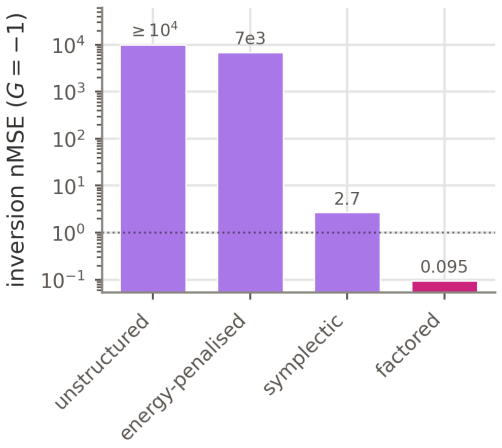


Figure 3: Inversion error at $G = -1$; only the factored model falls below nMSE 1.

The converse also holds: factoring provides the counterfactual but not stability. Placed on a non-conserving base, the factored coupling still inverts the regime but diverges over long rollouts (§4.3). On smooth parameter shifts the distinction becomes quantitative rather than binary: every model's one-step error rises smoothly, and the factored model achieves the lowest extrapolation error (10 to 30$\times$ below the training band) for the coupling it explicitly factors, but gains no advantage for the mass parameter it does not.

4.3 THE DISSOCIATION, AND THE CONTROLS THAT RULE OUT CONFOUNDS

The two effects meet in Table 3, which crosses energy conservation with coupling factorization in one matched comparison. The stability column reports true-energy drift at 2000 steps, where bounded models remain below 2 and divergent models reach the 10^4 display cap. The inversion

Table 3: The double dissociation, from one matched comparison on oracle three-body gravity.

Model	conserves	factored	stability: drift	inversion: match (nMSE)
unstructured	no	no	$>10^4$	no (0)
symplectic	yes	no	3.9	no (0)
factored graph-net	no	yes	$>10^4$	**yes** (1; 0.35)
factored symplectic	yes	yes	1.7	**yes** (1; 0.08)

column reports regime match together with rollout nMSE under the sign flip. The pattern is a clean double dissociation. Stability appears when the update is symplectic, while inversion appears when the coupling is factored, and the two effects vary independently. The factored symplectic model is the only model that both remains stable and answers the changed law; the unstructured model does neither. We next remove the most plausible confounds, namely capacity, integrator choice, and the choice of formalism.

The first concern is that the counterfactual benefit may come from some feature that accompanies the factored potential rather than from factorization itself. We test this by varying the state representation, parameter map, and integrator one at a time using an exact linear diagnostic. Each candidate fits a stiffness matrix by ridge regression to force labels containing a fixed small non-gradient curl component and observation noise. Hamiltonian candidates are projected onto the symmetric cone, while the conditional parameter map is constructed from positive-localised features. The design crosses leapfrog, RK4, and implicit midpoint at three step sizes in two coercive coupled-oscillator systems. Every candidate fits the training regime comparably, so the comparison is not driven by state escape (paired runs, Holm $p \approx 0.003$). At matched in-distribution fit, the factored map reaches error 0.0035 at the unseen coupling, while an unrestricted map trained on the same data reaches 0.985. The Hamiltonian representation keeps true-energy error at 5 to 8×10^{-4}, compared with 0.31 to 0.33 for an unrestricted vector field.

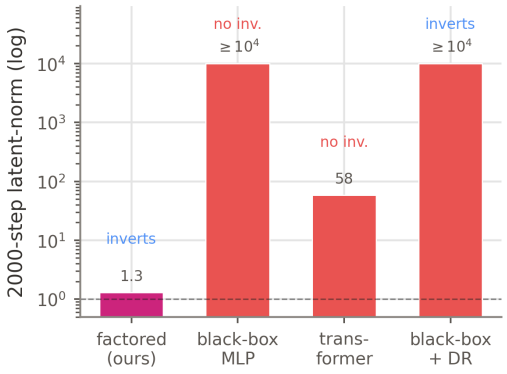


Figure 4: Stronger baselines on the sign flip: 2000-step latent-norm (stability, log), with each model's inversion outcome marked.

Two controls clarify this result. First, the energy advantage tracks the injected curl: at zero curl both representations drift little (0.0007 and 0.002), but as contamination rises to curl fractions 0.03, 0.06, 0.10 the unrestricted model climbs to 0.35, 0.94, 4.30 while the Hamiltonian stays near 0.0008, and since no configuration escapes, the factorial isolates fidelity and inversion rather than mere prevention of divergence. Second, the inversion gap reflects the extrapolation prior, not capacity: under the same objective a multiplicative map transfers whether it has eight scalars or 4.5k parameters, a flexible network receiving the coupling directly only partially extrapolates (0.071–0.075), and one receiving it through positive-localised features stays near 1.0 despite the most parameters. The transferable ingredient is the multiplicative structure, supplied rather than identified from positive-support data.

The separation persists across formalisms: on the engine gravity flip both factored models invert, while the three non-factored structured models fail at comparable stability. Capacity is ruled out directly: across a width sweep the unstructured model diverges at every width and the factored model inverts at every width, while adding capacity to the non-factored conserving model only worsens its drift. Graph-network simulators give the complementary control: factoring their messages restores the short-horizon response, but both variants fail the 2000-step test, so parameter transport alone does not provide long-horizon prediction (App. C.2).

A stronger architecture and a direct data remedy leave the dissociation unchanged (Fig. 4). A permutation-equivariant set-transformer, used as a stronger relational black box, still diverges over 2000 steps and fails to invert the unseen sign flip. Domain randomisation trains a black box on both

Table 4: Sign-flip inversion error (nMSE at $G < 0$) across state groundings.

State grounding	factored	unstructured twin
oracle state	0.067	—
pixel latent + canonical kinetic	0.065–0.092	0.47 / div.
grounded pixel model	0.12	—
decode-free (JEPA) + centroid anchor	0.40	0.68
learned object binder	0.15	never
frozen VideoMAE	0.13	0.65
frozen V-JEPA2	0.59	0.82
DINOv2 + frame differencing	0.096	0.62

signs of the coupling and therefore recovers inversion because the counterfactual is now part of its training distribution, but it provides no stability and diverges like the ordinary black box. Only the factored symplectic model both remains bounded and inverts, showing that neither attention nor data coverage substitutes for the relevant structural commitment.

4.4 Where conservation and factoring each stop helping

Both structural priors have clear limits. Under linear drag, a port-Hamiltonian model dissipates energy with the correct sign and order of magnitude, at roughly half the true rate ($-0.014/-0.063$ compared with $-0.033/-0.091$). A conserving model cannot dissipate at all. Conservation is therefore the wrong prior once the physical system genuinely loses energy.

A factored law with the wrong parameter dependence gives a subtler failure. A linear-G prior trained on a G^2 law fits the observed regime well ($0.014 \approx 0.010$) but predicts a spurious sign flip with error 0.343. Only the correctly specified G^2 prior restores the counterfactual (0.022). A wrong interaction structure, such as a pairwise prior fitted to a three-body interaction, instead produces poor in-distribution fit (0.43), so coupling-power misspecification is the more dangerous case, failing without an in-distribution warning. The ambiguity nonetheless has a constructive resolution: when linear (G) and even ($|G|$) couplings fit the positive-support data equally well, positive-only queries reach unique selection only half the time (0.50 across four candidate maps), whereas a single fixed query at $G = -1$, using no ground truth, reaches 1.00, and a non-adaptive query over the full domain is intermediate (0.775; App. C.7).

Perception introduces a third boundary: a single frame recovers position and fields but not velocity, so the inferred state falls off the correct energy shell, and the limiting variable is momentum rather than perception as a whole (App. C.4). One further diagnostic sharpens the map. Time-reversal consistency helps exactly when the underlying dynamics are reversible, giving retrace match 1 under conservative contact and only partial recovery under dissipation, which refines the conclusion of Botev et al. (2021). A truth-assisted transport score from the bound in Prop. 3 ranks which representation and map roll out worst under the flip (pooled Spearman 0.87 versus 0.09 for the initial-state term alone), a model-selection diagnostic rather than a certificate.

4.5 The dissociation is not an artefact of the headline system

The same two mechanisms extend well beyond smooth three-body gravity with oracle state. In generalized coordinates, a Hamiltonian model with a learned mass metric on a MuJoCo double pendulum keeps its energy drift near 0.1 while unstructured baselines diverge, and the gravity flip still inverts only when the coupling is factored, with regime match 1 versus 0. Under contact, the symplectic model remains bounded as collisions become stiffer while the unstructured model diverges (App. C.3). The factored coupling also transfers across object count and force family. A model trained on three bodies still inverts with thirty-two; it transfers to Coulomb charges under a held-out sign combination, with error 0.52 compared with 1.42 for a black box; and a single factored model trained on a mixed gravity-plus-Coulomb generator inverts both laws about $120\times$ more accurately than a black box receiving the same couplings.

The inversion comes from the factored channel itself rather than from additional data or an unrestricted shortcut. Increasing the training set does not improve the non-factored model and slightly worsens it: its inversion error rises from nMSE 0.76 with 32 trajectories to 2.67 with 512, because additional attractive trajectories reinforce a continuation that becomes incorrect after the sign flip. The structural channel is also necessary rather than incidental. When a free black-box shortcut is added alongside the factored pathway, it absorbs about 42% of the force and degrades inversion. Penalising the shortcut reduces its contribution to 10^{-4} and restores fidelity, improving nMSE from 0.43 to 0.09 (App. C.5).

Because neither behavior should rely on a hand-built state, we also ground the counterfactual in pixels (Table 4). A latent learned from images recovers inversion once a fixed canonical kinetic form is imposed, reaching near oracle-state error. A decode-free objective inverts once a label-free image centroid fixes the latent gauge. A learned object binder over a fixed detector restores inversion after removing object identity had broken it, succeeding on $10/10$ seeds versus $0/10$ for the unstructured twin (exact binomial $p \approx 10^{-3}$). The same structure transfers to frozen public video encoders we did not train (VideoMAE, V-JEPA2), where the factored readout inverts more reliably than an unstructured readout.

Encoder scale alone does not recover the physical readout: in V-JEPA2 from $\sim$ 300M to 1B parameters, inversion error is essentially flat ($1.02\times$) and the dissociation persists, so perception-side scale does not supply the counterfactual. The split appears within a single model: a readout recovers position and momentum from frozen V-JEPA2 tokens (nMSE 0.28–0.43), yet its action-conditioned predictor under no-op actions is no better than repeating the final frame. The missing ingredient is momentum: per-frame encoders capture position but not velocity, and inversion returns only when momentum is estimated by differencing features across frames, reducing error from 0.62 to 0.096 and matching video encoders. Thus the momentum boundary of §4.4 reappears at the representation level (App. C.8).

The structural account also extends beyond conservative dynamics. In a confined thermal bath, a positive-semidefinite friction port with fluctuation–dissipation noise reaches the correct stationary temperature, while a conserving model does not thermalize. The same passive port cannot represent anti-damping (Prop. 4), so energy injection must enter via a separate factored drive. That drive recovers the driven steady state across trained and held-out strengths, including an unseen cooling sign (Prop. 5). With video observations, a grounded energy-balance estimator recovers the damping rate; on time-reversed footage it yields a negative rate that the sign-constrained port cannot represent, giving an arrow-of-time measure paired with an architectural refusal (App. C.6, C.8).

5 Conclusion and Limitations

In this work we investigated which structural commitments make a learned world model obey physical laws, evaluating each through long-horizon stability and counterfactual generalisation. First, energy conservation through a symplectic update provides long-horizon stability where equal-capacity predictors, tuned neural ordinary differential equations, and fixed energy penalties fail, yet it does not improve counterfactual generalisation. Second, making the interaction force linear in the physical coupling enables extrapolation to never-seen signs of that coupling, which an unrestricted parameter map fails to recover, yet it does not improve stability. Third, we map the boundary of each mechanism and identify where its structural prior ceases to help. Fourth, we extend the same account beyond conservative dynamics, using a passive friction port that thermalises correctly and a factored drive that generalises across unseen forcing strengths and signs. The broader lesson is that physical structure does not make a model uniformly more physical. Specific commitments produce specific behaviours, allowing a modeller to combine structural components according to the capabilities a task requires.

However some limitations remains. Conservation becomes the wrong prior when the true system genuinely dissipates energy, while factoring helps only when the intervened parameter follows the supplied functional form. A misspecified coupling power can therefore fit the training regime well while failing under intervention. A fixed kinetic term and a label-free anchor reduce the ambiguity, but joint learned detection and binding remain open. Finally, all evidence in this study comes from synthetic simulators and rendered observations rather than real video, the general deployment guarantee of our work is the next step we aim to ensure.

AI USE STATEMENT

A large language model assisted the authors in polishing the language of the manuscript. The authors verified that all claims, proofs, mathematical formulations, and reported values are valid, checked them against the implementation and results, and take full responsibility for the manuscript.

ETHICS STATEMENT

This work uses only synthetic physics simulations, so no human subjects, personal data, or scraped datasets are involved. We do not foresee direct harmful applications of the findings, which concern the architectural conditions for physical generalisation in learned simulators. The main integrity-relevant practices, namely prospective committed protocols with metrics fixed in advance, equal-capacity baselines, and the reporting of negative and boundary results, are described in the paper and its record appendices.

REPRODUCIBILITY STATEMENT

App. A specifies all systems, models, integrators, metrics, and the perception pipeline. App. B gives the derivations and proofs. App. C gives the per-experiment protocols, seeds, and per-seed results. App. F gives the pre-registration protocol (metrics fixed in advance, the equal-capacity rule, and non-finite capping), the statistics conventions, and the compute and hyperparameter defaults (Table 23). Our implementation and the pre-registration documents are provided as supplementary material.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Yann LeCun, Michael Rabbat, Nicolas Ballas, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint*, 2025.

Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Daniel M. Bear et al. Physion: Evaluating physical prediction from vision in humans and machines. In *NeurIPS Datasets and Benchmarks*, 2021.

Aleksandar Botev, Andrew Jaegle, Peter Wirnsberger, Daniel Hennes, and Irina Higgins. Which priors matter? benchmarking models for learning latent dynamics. In *NeurIPS Datasets and Benchmarks*, 2021.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences (PNAS)*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *ICLR 2020 Deep Differential Equations Workshop*, 2020.

Shaan A. Desai, Marios Mattheakis, David Sondak, Pavlos Protopapas, and Stephen J. Roberts. Port-Hamiltonian neural networks for learning explicit time-dependent dynamical systems. In *Physical Review E 104, 034312*, 2021. doi: 10.1103/PhysRevE.104.034312.

Marc Finzi, Ke Alexander Wang, and Andrew Gordon Wilson. Simplifying hamiltonian and lagrangian neural networks via explicit constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. Phyworldbench: A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics. Springer, 2nd edition, 2006. doi: 10.1007/978-3-662-05018-7.

Chen-Di Han, Bryan Glaz, Mulugeta Haile, and Ying-Cheng Lai. Adaptable hamiltonian neural networks. *Physical Review Research*, 3(2):023156, 2021. doi: 10.1103/PhysRevResearch.3.023156.

Andreas Hochlehnert, Alexander Terenin, Steindor Sæmundsson, and Marc Peter Deisenroth. Learning contact dynamics using physically structured neural networks. In *Artificial Intelligence and Statistics (AISTATS)*, 2021.

Pengzhan Jin, Zhen Zhang, Aiqing Zhu, Yifa Tang, and George Em Karniadakis. Sympnets: Intrinsic structure-preserving symplectic networks for identifying hamiltonian systems. *Neural Networks*, 2020.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics adaptation. In *International Conference on Machine Learning (ICML)*, 2022.

Kookjin Lee, Nathaniel Trask, and Panos Stinis. Machine learning structure preserving brackets for forecasting irreversible processes (GFINNs). In *NeurIPS*, 2021.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Pu Zhao, Yanzhi Wang, et al. Phyground: Benchmarking physical reasoning in generative world models. *arXiv preprint*, 2026.

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

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-jepa: Learning world models through object-level latent masking. *arXiv preprint*, 2026.

Cyrus Neary and Ufuk Topcu. Compositional learning of dynamical system models using port-Hamiltonian neural networks. In *Learning for Dynamics and Control Conference (L4DC)*, volume 211 of *Proceedings of Machine Learning Research*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Lyndsey C. Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Schölkopf, and William T. Freeman. Seeing the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. doi: 10.1109/CVPR.2014.262.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint*, 2018. doi: 10.1109/TPAMI.2021.3083839.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning (ICML)*, 2020.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

Andrew Sosanya and Sam Greydanus. Dissipative hamiltonian neural networks: Learning dissipative and conservative dynamics separately. *arXiv preprint*, 2022.

Molei Tao. Explicit symplectic approximation of nonseparable hamiltonians: Algorithm and long time performance. *Physical Review E*, 94(4):043303, 2016. doi: 10.1103/PhysRevE.94.043303.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian generative networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Rui Wang, Robin Walters, and Rose Yu. Meta-learning dynamics forecasting using task inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.52202/068431-1573.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00840.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *International Conference on Learning Representations (ICLR)*, 2020.

Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, and Patrick Gallinari. Leads: Learning dynamical systems that generalize across environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yaofeng Desmond Zhong and Naomi Leonard. Unsupervised learning of lagrangian dynamics from images for prediction and control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Dissipative symoden: Encoding hamiltonian dynamics with dissipation and control into deep learning. *ICLR 2020 Workshop DeepDiffEq*, 2020a.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. In *International Conference on Learning Representations (ICLR)*, 2020b.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Benchmarking energy-conserving neural networks for learning dynamics from data. In *Learning for Dynamics and Control (L4DC)*, 2021a.

Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Extending lagrangian and hamiltonian neural networks with differentiable contact models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

A Formal setup and definitions

A system of N bodies has state $x = (q, p)$, with positions $q \in \mathbb{R}^{Nd}$ and conjugate momenta $p \in \mathbb{R}^{Nd}$ (d is the spatial dimension, equal to 2 except for the pendulum, where q are joint angles). A parameter vector $c = (G, \{m_i\})$ collects the coupling constant and masses. A dynamics model is a map $f_\theta : (x_t, c) \mapsto x_{t+1}$, iterated to a rollout $x_{0:T}$. Ground truth comes from a reference integrator Φ. Training minimises one-step squared error on forward-law data only. The k-step variant instead sums this error over a k-step rollout.

The evaluated systems are as follows. (i) **Conservative gravity** (oracle state): $H = \sum_i \|p_i\|^2/2m_i + V(q)$ with softened-core potential

$$V(q) = -\sum_{i<j} \frac{G\, m_i m_j}{\sqrt{\|q_i - q_j\|^2 + \epsilon^2}}, \qquad \epsilon = 0.1,\ dt = 0.002. \tag{6}$$

The softening keeps the ground-truth integrator near-conservative (drift $\sim$0.01; $\epsilon = 10^{-2}$ increases the reference drift by $110\times$). This softening and the associated reference-integrator drift are reported with all gravity results. (ii) **Coulomb**: the same form with Gm_im_j replaced by a charge product $k\, z_i z_j$, $z_i \in \{\pm 1\}$. (iii) **Springs-with-cutoff** (necessity screen). (iv) **Soft-sphere disks** on a periodic torus (minimum-image), contact stiffness series set by the stiffness $k \in \{30, 300, \infty\}$. (v) **Dissipative gravity**: Eq. equation 6 plus linear drag, $\dot{p}_i = -\partial V/\partial q_i - \gamma\, p_i/m_i$. (vi) **Double pendulum** (MuJoCo): joint angles q, configuration-dependent $M(q)$, generalised momentum $p = M(q)\dot{q}$, RK4 engine ground truth; conservative and joint-damped variants.

Parameter counts are approximately matched in the five-model gravity comparison. Other comparisons use the experiment-specific budgets reported below. Equal parameter count does not imply equal functional capacity. The unstructured MLP is a residual MLP, $x_{t+1} = x_t + g_\theta(x_t, c)$. The energy-penalised model is the unstructured MLP plus an energy-drift penalty $\lambda\, |E_\theta(x_{t+1}) - E_\theta(x_t)|$. The interpretation of this baseline depends on how E_θ is obtained: a jointly learned unconstrained energy can become constant and make the penalty vacuous. Its failure alone therefore cannot establish the inadequacy of a penalty evaluated using a fixed physical energy. The penalty energy is *fixed and analytic*, not learned or pretrained: the true physical energy $E = \sum_i \|p_i\|^2/2m_i - G\sum_{i<j} m_i m_j/\sqrt{r_{ij}^2 + \epsilon^2}$ with the true masses and G and the same softening $\epsilon = 0.1$ as the generator, applied as $\lambda\,(E(\hat{x}_{t+1}) - E(x_t))^2$ on the de-standardised prediction. The penalty does not by itself exclude the constant-energy solution (a wrong trajectory that holds E near its initial value), which is why rollout error is always co-reported with energy drift. In the raw record the $\lambda = 100$ arm does not take that route: its energy error and its rollout error grow together (drift 115–$>10^4$ with normalised rollout error 82–$>10^4$ at 2000 steps), after an early phase in which its energy error is below the symplectic model's (Table 6). The symplectic model is a symplectic HNN, $H_\theta = T_\theta(p) + V_\theta(q, c)$, one leapfrog step. The factored symplectic model is the factored HNN, potential Eq. equation 5; the canonical *fixed-kinetic* variant sets $T = \sum_i \|p_i\|^2/2m_i$. DeLaN (Lutter et al., 2019): Lagrangian $L_\theta = \frac{1}{2}\dot{q}^\top M_\theta(q)\dot{q} - V(q, c)$, Euler–Lagrange accelerations, factored/non-factored V. SymODEN (Zhong et al., 2020b): control-Hamiltonian, learned $M_\theta(q)^{-1}$ and V_θ, RK4. Generalised-coordinate HNN: $H = \frac{1}{2}p^\top M_\theta(q)^{-1}p + g\, U_\theta(q)$, Cholesky-SPD metric, factored potential. Port-Hamiltonian: conservative flow plus a learned dissipative port $-\gamma\, R_\theta(p)$. The graph-network simulator and its factored variant (Sanchez-Gonzalez et al., 2020; Battaglia et al., 2016): an interaction network; the factored form multiplies the aggregated message by G. Hybrid (necessity): factored pathway plus a free black-box shortcut with an L_2 penalty λ on the shortcut force.

Separable Hamiltonians use symplectic leapfrog. The non-separable generalised-coordinate Hamiltonian uses Tao's explicit symplectic scheme (Tao, 2016), admitted only after a calibration gate (a known Hamiltonian must stay within 10^{-4} relative energy drift over 2000 steps).

The relative energy drift is

$$D(t) = \frac{|E(x_t) - E(x_0)|}{|E(x_0)|}, \tag{7}$$

with E the true energy of the predicted state; regime-match and the affine identifiability residual are Def. 1. Energy-shed rate (dissipation) and time-reversal match (retrace a reversed-momentum end state) are defined analogously.

For the stochastic dissipative results (§4.5, App. C.6), we separate mechanical friction from a closed-system GENERIC formulation. For canonical kinetic energy and momentum-space friction, the continuous model is

$$dq = \nabla_p H\, dt, \qquad dp = \big[-\nabla_q H - M\nabla_p H + f_{\text{ext}}\big]dt + \Sigma\, dW, \qquad \Sigma\Sigma^\top = 2k_B T\, M, \quad M \succeq 0. \tag{8}$$

The deterministic friction contribution satisfies $\dot{H}_{\text{fric}} = -(\nabla_p H)^\top M \nabla_p H \leq 0$. This is a passive mechanical channel, not a claim that subsystem energy cannot rise through work or thermal fluctuations. State-dependent mobilities require a separately specified stochastic drift and convention;

the isolated OU result below concerns constant coefficients. A closed-system GENERIC model additionally requires degeneracy conditions $L\nabla S = 0$ and $M\nabla E = 0$, beyond antisymmetry and positive semidefiniteness (Lee et al., 2021). Those closed-system laws are not established by the passive-channel tests here. The consequences of the imposed relation are reported. (vii) **Confined Langevin N-body**: softened gravity plus a harmonic trap and an exact Ornstein–Uhlenbeck thermostat at bath temperature T; the diagnostic observable is the kinetic temperature $\hat{T} = \langle p^2/m \rangle$ (equal to T at equilibrium by equipartition, k_B=1). (viii) **Damped double pendulum** with linear and nonlinear (v^2) joint friction (the engine of §4.5). (ix) **Driven-dissipative particles**: system (vii) plus a linear active drive of strength A opposing friction γ, reaching a *non-equilibrium steady state* at $T_{\text{eff}} = T\gamma/(\gamma - A)$ (App. B). Models add a dissipation channel to the symplectic backbone: a diagonal positive-semidefinite friction with fluctuation–dissipation noise (the *guarantee* side, "$M \succeq 0$"); the full PSD friction operator $M = LL^\top$ (Cholesky) with correlated noise; a *factored* friction $M = \gamma M_0$ whose sign flips with γ and permits inversion with respect to γ; and, for system (ix), a factored drive $A\, w_\theta$ with net damping $\gamma_f - A\, w$ and fluctuation–dissipation noise tied to γ_f. Primary metric is the temperature relative error $|\hat{T} - T^*|/T^*$ (against the correct equilibrium or steady-state temperature T^*, the true simulator's measured value which targets agreement with the discrete reference rather than cancellation of model-dependent integration bias); secondary is the energy-shed rate and the injection rate at $\gamma < 0$.

A scene is rendered to synthetic frames by a fixed differentiable renderer R (Gaussian or coloured-disk splats at body positions); an encoder E_ϕ maps a frame stack to a latent $(\hat{q}, \hat{p})$. In the supervised setting, E_ϕ is trained against the oracle state. In the discovered setting, it is trained only through future-frame prediction, $\|R(\text{rollout of } E_\phi(\cdot)) - \text{frames}\|^2$, without oracle-state supervision. The symplectic-consistency prior adds $\|\hat{p}/m - v_{\text{obs}}\|^2$, v_{obs} a finite difference of soft-argmax centroids across frames (an observed velocity, not oracle state). In the end-to-end pixel comparison every model shares E_ϕ and R; a separate branch removes R to test unsupervised discovery.

B Derivations

Proof of Proposition 2. Write $f_i(q, G) = -\partial V/\partial q_i = -G\, \partial_{q_i} U_\theta(q)$. V is linear in G with G-independent coefficient $\sum_{i<j} \varphi_\theta$, so $\partial V/\partial q_i$ is linear in G and $f_i(q, -G) = -f_i(q, G)$. The Hamiltonian vector field's momentum component $\dot{p} = f$ therefore changes sign with G, whereas $\dot{q} = \partial T/\partial p$ is unchanged. The resulting trajectory is the flow generated by the same learned potential under coupling $-G$. □

Comparison with the non-factored model. A black-box conditional potential $V_\theta(q, G)$ has $\partial V_\theta/\partial q$ with no constraint linking $G > 0$ to $G < 0$, so $f_\theta(q, -G)$ is unconstrained by $G > 0$ training and inversion is not guaranteed; empirically it fails (App. C.2). Thus, the dissociation isolates *factoring*, rather than conservation, as the inversion ingredient.

Proposition 1, formal statement. Let H_θ be analytic on a neighbourhood of a compact set that contains the numerical orbit, and let the step dt be below the threshold of backward-error analysis for that set. Then the leapfrog map applied to equation 4 is the exact time-dt flow of a modified Hamiltonian $\widetilde{H}_\theta = H_\theta + O(dt^2)$ up to a remainder of size $O(e^{-c/dt})$ per step, so $\widetilde{H}_\theta$ is conserved up to $O(e^{-c/dt})$ over times of order $e^{c/dt}$ and the numerical H_θ stays within $O(dt^2)$ of its initial value over that interval (Hairer et al., 2006). The statement is about $\widetilde{H}_\theta$ and the learned H_θ; it says nothing about the physical energy unless H_θ equals it. *Empirical regularity (not a theorem).* A generic explicit map $x_{t+1} = x_t + g_\theta(x_t)$ has no corresponding conservation guarantee. Its energy behaviour therefore depends on the learned map; a contraction, for example, may remain bounded. In these experiments, the unstructured and graph-network models exhibit secular drift and reach the reporting cap. The proposition supplies a sufficient condition for long-time near-conservation, whereas divergence of the unstructured models is an empirical observation. *Remark.* Training-noise injection reduces the per-step prediction error but does not itself provide an energy-boundedness guarantee. In the graph-network simulator it delays, but does not prevent, the observed drift (App. C.2). The learned $\widetilde{H}_\theta \neq H_{\text{true}}$, which is why the symplectic model is *bounded but not flat*; multi-step training and longer horizons shrink and saturate the gap.

Proposition 3 (Finite-time transport bound)**.** *Let $F^\star$ be the reference field, F_θ^{true} the learned content evaluated with the true parameter map, and F_θ^{decl} the deployed field using the declared map, all at the tested coupling. Suppose the reference and deployed trajectories remain in a compact region Ω on $[0,T]$, and F_θ^{decl} is Lipschitz there with constant $L \geq 0$. Define*

$$\epsilon_c = \sup_{\Omega} \|F_\theta^{\mathrm{true}} - F^\star\|, \qquad \epsilon_m = \sup_{\Omega} \|F_\theta^{\mathrm{decl}} - F_\theta^{\mathrm{true}}\|.$$

Then

$$\|\hat{x}_t - x_t\| \leq e^{Lt}\big(\|\hat{x}_0 - x_0\| + t(\epsilon_c + \epsilon_m)\big), \qquad 0 \leq t \leq T.$$

Proof. Adding and subtracting the deployed field at the reference state bounds the trajectory-error growth by $L\|\hat{x} - x\| + \epsilon_c + \epsilon_m$. Integration and Gronwall's inequality give the displayed bound. □

The auxiliary true-map field and both suprema are truth-assisted quantities. They are not generally available at deployment. The bound assumes trajectories remain in Ω and therefore does not establish global stability. The reported pooled Spearman association (0.87 versus 0.09 using only initial-state error) describes a ranking on the evaluated models (§4.4). It is not coverage calibration or a certified prediction of error on new model families.

Proposition 4 (Passive friction and sign-reversed anti-damping)**.** *Consider the deterministic momentum-friction channel $\dot{p}|_{\mathrm{fric}} = -M\nabla_p H$ with $M \succeq 0$. Its contribution to $\dot{H}$ is nonpositive for every represented state. If $M = \gamma M_0$, with $M_0 \succ 0$, changing γ to a negative value makes this contribution positive whenever $\nabla_p H \neq 0$. Thus a globally PSD parameterisation cannot represent that sign-reversed channel. For a bath with $T > 0$, negative γ also prevents $\Sigma\Sigma^\top = 2k_B T\gamma M_0$ from being a real noise covariance.*

Proof. The chain rule gives $\dot{H}_{\mathrm{fric}} = -(\nabla_p H)^\top M \nabla_p H$. Its sign follows from positive semidefiniteness; for negative γ and nonzero velocity the sign reverses. A covariance is positive semidefinite, whereas $2k_B T\gamma M_0$ is negative definite. □

This restriction concerns the same passive channel. It neither forbids external work or thermal energy injection nor makes every anti-damped stochastic model ill-posed. It also does not establish the invariant law of a coupled numerical simulator (App. C.6).

Proposition 5 (Effective stationary temperature for linear drive)**.** *Let $H(q,p) = U(q) + \|p\|^2/(2m)$, $m > 0$, $\gamma > 0$, $T > 0$, and $A < \gamma$. Consider the continuous underdamped dynamics*

$$dq = (p/m)\,dt, \qquad dp = -\nabla U(q)\,dt - (\gamma - A)(p/m)\,dt + \sqrt{2\gamma k_B T}\,dW.$$

If the density is normalizable and boundary terms vanish, then $\rho_\infty(q,p) \propto \exp[-H(q,p)/(k_B T_{\mathrm{eff}})]$ is stationary, where $T_{\mathrm{eff}} = T\gamma/(\gamma - A)$. Each momentum component has variance $mk_B T_{\mathrm{eff}}$.

Proof. Hamiltonian transport annihilates any differentiable density depending only on H. The momentum probability current from friction and noise is

$$(\gamma - A)(p/m)\rho_\infty + \gamma k_B T \nabla_p \rho_\infty = \big((\gamma - A) - \gamma T/T_{\mathrm{eff}}\big)(p/m)\rho_\infty = 0.$$

Thus the Fokker–Planck operator annihilates the stated density. Its momentum marginal is Gaussian with the given variance. □

The isolated OU momentum process has no stationary probability law when $A \geq \gamma$ and $T > 0$. The proposition makes no corresponding unqualified assertion for arbitrary coupled nonlinear models, and proves neither uniqueness nor exact preservation by a numerical integrator. The externally driven bath interpretation is distinct from the mathematical effective Gibbs form. A factored drive supplies the coefficient dependence; predictive accuracy still depends on the remaining learned dynamics.

Definition 1 (Regime-match and affine recovery)**.** Regime-match. *Let $\hat{x}$ be the model's rollout under the flipped law, and x^-, x^+ the ground-truth flipped-law and training-law (default) rollouts, both from the* true *initial state (so initialisation error is counted, not cancelled). The regime is matched iff $\|\hat{x} - x^-\| < \|\hat{x} - x^+\|$; we report the matched fraction and the normalised MSE*

$\|\hat{x} - x^-\|^2/\mathrm{Var}(x^-)$*, for nonzero reference variance. Ties do not count as matches. A non-finite rollout is reported as a failure rather than assigned a nearest-reference label. Two conventions are made explicit here. First, the distance is the batch-pooled normalised MSE over the whole rollout, which gives one regime-match per run rather than per episode. Second, the sign convention is fixed: the correct reference is rolled at the tested coupling* G_{test} *and the default reference at* $|G_{\mathrm{test}}|$*, and the regime is scored as* $\mathbb{1}[\mathrm{nMSE}(\hat{x}, x^-) < \mathrm{nMSE}(\hat{x}, x^+)]$. Affine recovery residual *of a discovered latent* $\hat{z}$ *with respect to the true* (q, p)*:* $\min_{A,b}\|A\hat{z} + b - (q, p)\|^2/\mathrm{Var}(q, p)$*. This measures linear recoverability under a fitted affine map, not uniqueness of the latent coordinates. Values below* 0.5 *are a descriptive threshold used in this study, not an identifiability theorem.*

C Extended results and findings

This appendix reports the findings summarised in the main text, together with their figures, multi-seed estimates, experiment-specific methods, and stated limitations. The unified notation of App. A is used: state $x = (q, p)$, coupling G, relative energy drift $D(t)$ (Eq. 2), regime-match and affine identifiability residual (Def. 1). Each finding is tagged *[well-supported]*, *[directional]*, *[supported]*, or *[boundary]* (criteria in App. E).

Aggregates are the mean over at least three seeds, and at least five seeds at scale, with the available aggregate summaries in App. C. Where a bracket is labelled a normal-approximation 95% interval, it retains that meaning even for small seed counts and is not an observed range. Such small-sample intervals are descriptive and do not justify significance from non-overlap. Individual seed values are not tabulated for every comparison.

Display-clipped finite errors and non-finite numerical failures are distinct. A sentinel substituted for a failure is not a measured error and cannot establish a quantitative error ratio. These figures retain their plotting thresholds, and the text distinguishes what the two mean. A pre-registered absolute-error *backstop* marks a perception result invalid when even its in-distribution fidelity is worse than the threshold (such runs are reported as invalid, not compared). Regime-match is meaningful only for a *bounded* rollout. For a diverged rollout, such as that of the graph network, comparison with the nearest reference degenerates and is not used as evidence. Such cases are identified explicitly. Table 5 summarises all the findings; in it, "match" is the regime match (1 = follows the unseen flipped law), "nMSE" the normalised inversion error, and "drift" the relative true-energy drift of Eq. 2.

Table 5: Summary of findings and evidence classes.

Finding	System	Summary statistic	Class
Stability	gravity	symplectic model drift $0.40 \to 7.1$ vs cap 10^4	well-supported
Tuning-robust	gravity	every unstructured arm passes the cap; λ=100 delays but does not bound drift	well-supported
Multi-step	gravity	factored symplectic model 2000-step drift $4.4 \to 1.9$	directional
10k-step	gravity	factored symplectic model drift saturates $1.9 \to 0.89$	well-supported
Non-factored	gravity	symplectic model match 0 (no inversion)	boundary
Factored	gravity	factored symplectic model match 1, nMSE 0.03–0.18	well-supported
Param. extrap.	gravity	smooth: $2\times$ edge, no abrupt degradation	directional
DeLaN/SymODEN	engine	factored match 1 vs non-factored 0	well-supported
Graph-net	gravity	reported response labels 1 vs 0; both fail long-horizon test	boundary
Contacts	disks	bounded stiffness series; sym-inversion 1	directional
Engine	pendulum	factored symplectic model match 1, nMSE $8.7e{-4}$	well-supported
Contact aug.	disks	no better than smooth model	boundary
Dissipation	drag	conservation sheds 0; port sheds	boundary
Supervised perc.	pixels	inversion retained; error increases with speed	directional

continued on the next page

Table 5 (continued)

Finding	System	Summary statistic	Class
Discovered perc.	pixels	fixed-kinetic nMSE $0.075 \approx 0.067$	well-supported
Momentum recov.	pixels	p-resid $0.72 \to 0.35$; pos. open	directional
Pixel comparison	pixels	factored fixed-kinetic model nMSE 0.12; others fail	well-supported
Necessity	gravity	shortcut $42\% \to 10^{-4}$; nMSE $0.43 \to 0.09$	directional
Comp. & scale	gravity	match 1 to $N{=}32$; drift $0.2 \to 102$	directional
Coulomb, held-out sign	Coulomb	factored 0.52 vs black-box 1.42; $2.7\times$	well-supported
Mixed two-family inversion	Coulomb	factored 0.0035 vs black-box 0.425; $120\times$	well-supported
Data-efficiency	gravity	inverts with 32 traj.; symplectic model error rises with data	well-supported
Capacity	gravity	unstructured MLP diverges at every width	well-supported
Composition	drag	three ingredients compose	directional
Control	gravity	structured plans transfer; unstructured MLP misestimates	directional
Rendering	pixels	structured video consistent; unstructured jumps	supported
Noise screen	gravity	robust to 30% state noise	supported
Temperature	Langevin	structured $\hat{T}$ has low error; unstruct. diverges	well-supported
Tension	Langevin	passivity precludes channel anti-damping; factored form permits it	well-supported
Full operator	Langevin	full M has higher error than diagonal M	supported
Engine dissipation	pendulum	port sheds energy, correct sign, $\sim$half rate	directional
Noise robustness	Langevin	correct $\hat{T}$ to $\sim 10\%$ meas. noise	directional
Driven drive	driven	factored drive inverts strength+sign, err ≤ 0.09	well-supported
Misspecification	gravity	low in-dist. error, incorrect counterfactual	well-supported
Transformer/dom-rand	gravity	transformer diverges and does not invert	well-supported
Grounded JEPA	pixels	anchored factored hybrid match 1, nMSE 0.40 vs 0.68	well-supported
Anonymous object	pixels	identity removed: both collapse to match 0.33	boundary
Learned binding	pixels	learned tracker + factored inverts ($10/10$ seeds, nMSE 0.15; control $0/10$)	well-supported
Detection+binding	pixels	slot-attention does not reach it (match 0)	boundary
External encoder	pixels	frozen VideoMAE/V-JEPA2 invert (0.13/0.59, both 10-seed, factored $10/10$ $p{=}0.001$; readout-matched $2.6\times$ over our CNN; scale-flat 300M$\to$1B, gain $1.02\times$)	well-supported
Momentum axis	pixels	differencing DINOv2 tokens closes the image-SSL gap ($0.62 \to 0.096$)	well-supported
Encoder composition	pixels	VideoMAE + learned detect/bind + prior does not compose (unstable, no clean inversion)	boundary
Second law from video	pixels	reported damping diagnostic orders damped $1.25 >$ cons 0.84 ($5/5$ seeds); passive friction by construction	supported (ordering only)
Floor located + refusal	pixels	oracle control rules out grounding noise as the sole explanation; training-level refusal directionally consistent $5/5$ but below the pre-registered bar	boundary
Energy-aux unification	pixels	the auxiliary-loss lever fails all three pre-registered predictions; with a learned meter it backfires (σ_{damped} $1.25 \to 0.62$)	pre-registered negative
Measure, don't fit	pixels	the energy-balance estimator removes the residual bias ($\hat{\gamma}_{\text{cons}}$ $0.84 \to -0.006$); arrow-of-time rate read with $\sigma \geq 0$ refusal; composition instability cured, grounding boundary stands	well-supported (one gate marginal)

C.1 A. STABILITY FROM CONSERVATION (EXTENDS §4.1)

Conservation → stability. Five approximately capacity-matched models (unstructured MLP, energy-penalised MLPs, and neural ODE at ∼139k parameters; symplectic model at ∼138k) predict one step of conservative gravity and roll to $T = 2000$ ($100\times$ the 20-step training horizon). The unstructured MLP and the $\lambda = 1$ penalised MLP pass the 10^4 cap in the median episode by step 500 and lose 61 and 123 of 192 episodes to non-finite states between steps 1000 and 2000. The $\lambda = 100$ penalised MLP and the neural ODE stay finite in every episode but pass the cap. The symplectic model stays finite and below the cap in every episode (Table 6). It remains bounded but has non-zero drift because near-conservation concerns a modified learned Hamiltonian under the assumptions of Prop. 1, not the true energy. Multi-step and longer-horizon training reduce this discrepancy.

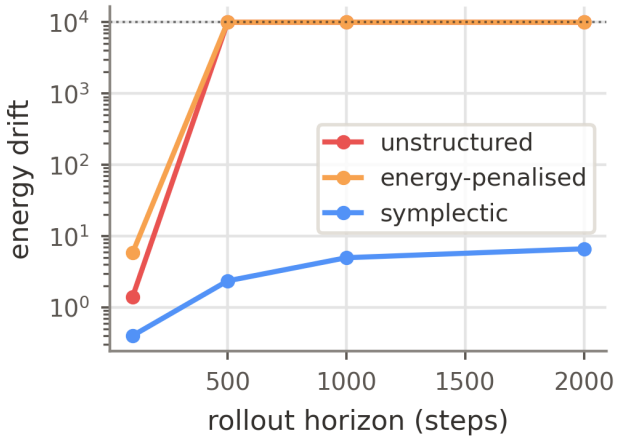


Figure 5: Relative energy drift $D(t)$ vs rollout step (log y, 3 seeds; the shaded band is the plotting normal-approximation interval, descriptive at three seeds) in the original three-model display. The unstructured MLP and energy-penalised model reach the 10^4 display cap by ∼step 300; the symplectic model plateaus. Numerical summaries are taken from the raw record (Table 6), not from this capped display.

Table 6: Raw stability record: per-episode median relative energy drift over 64 held-out episodes, mean [min, max] over three runs; $>10^4$ marks a median above the cap. NF is the number of episodes non-finite by step 2000 (of 192; none is non-finite by step 1000); err_{2000} is the median normalised rollout error at 2000 steps, range over seeds. Learning rates were selected per seed on held-out one-step error from $\{10^{-3}, 3\times10^{-4}, 10^{-4}\}$.

model	t=100	t=500	t=1000	t=2000	NF	err_{2000}
unstructured MLP	2.6 [0.6, 5.1]	$>10^4$	$>10^4$	$>10^4$	61	$>10^4$
energy-penalised, λ=1	3.3 [2.1, 4.0]	$>10^4$	$>10^4$	$>10^4$	123	$>10^4$
energy-penalised, λ=100	0.011 [0.009, 0.016]	0.33 [0.16, 0.49]	56 [1.4, 141]	115–$>10^4$	0	82–$>10^4$
neural ODE (RK4)	0.08 [0.07, 0.11]	1.8×10^3 [3.0×10^2, 4.0×10^3]	$>10^4$	$>10^4$	0	$>10^4$
symplectic model	0.40 [0.33, 0.46]	2.7 [2.0, 3.7]	4.3 [3.4, 6.2]	7.1 [6.7, 7.7]	0	10–14

The neural-ODE and energy-penalty controls. The neural vector field with RK4 uses matched capacity, the same step size and one-step objective, and the same learning-rate sweep. All three seeds remain finite in every episode. Its median energy drift is 0.08 at 100 steps, 3×10^2–4×10^3 at 500, and above the cap by 1000. Equal step size does not imply equal integration error across RK4 and leapfrog. The $\lambda = 100$ energy-penalised MLP has the lowest energy error of all the models at 100 and 500 steps, after which the drift grows without bound, with the rollout error growing alongside it.

All models share the evaluation episodes within a seed, so the record supports paired comparisons. The summaries here support a long-horizon energy-error contrast, not a claim that Hamiltonian

structure is necessary for finiteness. The comparison is generated from one raw record rather than from capped plots. For each model and run, that record stores the learning-rate grid and the selected rate, the held-out one-step error used for selection, the per-episode finite and non-finite status and counts at each checkpoint, the uncapped relative energy drift on finite episodes (median, minimum, maximum, and fraction above 1), and the rollout error, together with the selected checkpoint for each model and seed. Seed-level summaries are the mean, minimum, and maximum over seeds of per-episode medians.

C.2 B. Law-inversion and the double dissociation (extends §4.1–4.2)

Conservation alone does not provide inversion. Under never-trained repulsive $G < 0$, the symplectic model is stable but its rollout stays closer to the training-law trajectory (regime-match 0 at every G and seed). It learned a black-box $V_\theta(q, G)$ on $G > 0$ only (Prop. 2 contrast).

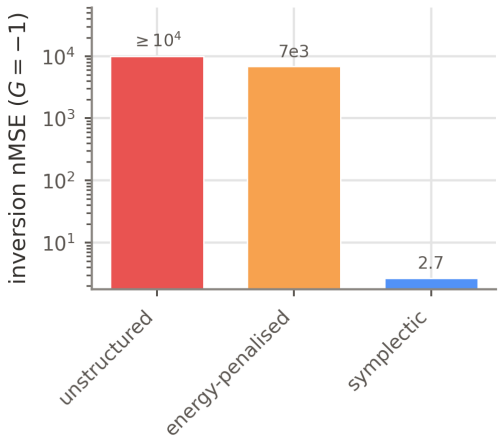
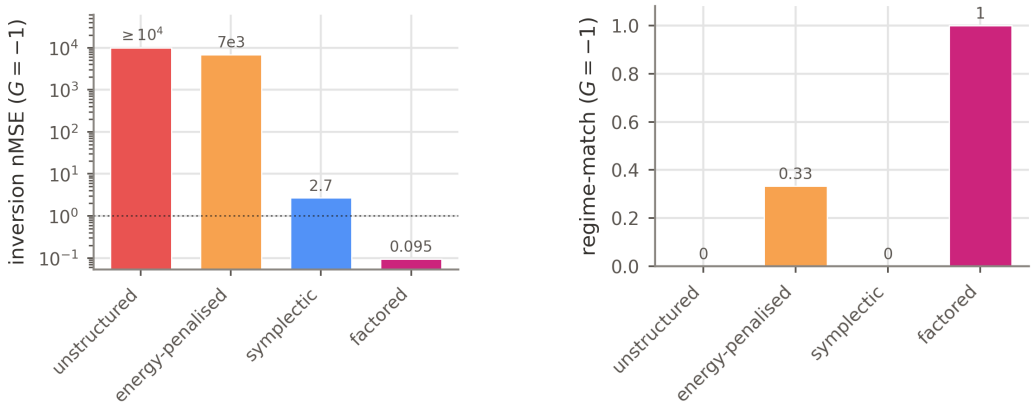


Figure 6: Non-factored HNN: stable but does not invert; repulsive rollout matches the attractive default.

The factored coupling yields inversion. Trained on attraction only, the factored symplectic model (Eq. 5) tracks unseen repulsion at nMSE 0.03–0.18 across $G = -0.5$ to -1.5 (regime-match 1 everywhere), more than an order of magnitude below the non-factored symplectic model and far below the (capped) unstructured baselines, while attaining the lowest in-distribution error with fewer parameters. Together with the non-factored result, this is the double dissociation.



(a) inversion error

(b) regime-match

Figure 7: Factored HNN tracks the true reversed ($G < 0$) reference trajectory; the other evaluated models diverge or remain closer to the training-law trajectory.

DeLaN/SymODEN: comparison across formalisms and integrators. On the engine gravity-flip, three non-factored structured models fail and both factored models invert (Table 7), isolating the axis to the factoring.

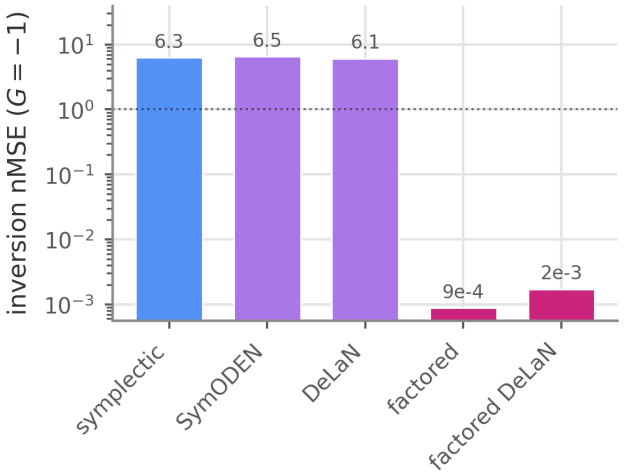


Figure 8: **DeLaN/SymODEN.** Inversion across five structured models: non-factored fail (match 0), factored invert (match 1), all equally stable.

Table 7: **DeLaN/SymODEN** engine gravity-flip inversion.

model	regime-match	inversion error
symplectic model (Tao, non-fac.)	0	$\sim$6
SymODEN (RK4, non-fac.)	0	$\sim$6
DeLaN (non-fac.)	0	$\sim$6
factored symplectic model (factored)	1	8.7×10^{-4}
factored DeLaN (factored)	1	1.7×10^{-3}

Matched causal factorial. A natural worry is that our stability and inversion contrasts change several things at once, namely the state representation, the update rule, and the numerical integrator. We therefore run a controlled factorial that varies these three axes independently. We cross a Hamiltonian representation against an unrestricted vector field, a coupling-factored parameter map against an unrestricted conditional map, and three integrators (leapfrog, RK4, and implicit midpoint) at three step sizes, over twelve paired seeds in two coercive coupled-oscillator systems, with every model fitted to the same observed regime. The unrestricted vector field has more coefficients than the Hamiltonian model. Parameter counts alone do not establish equal approximation capacity. Two results follow. First, at matched fit the Hamiltonian representation reaches a maximum true-energy error of about 5×10^{-4} to 8×10^{-4}, against 0.31 to 0.33 for the unrestricted field, and a trajectory error of 5×10^{-4} to 1.7×10^{-3} against 0.044 to 0.059 (Table 8). Second, at the unseen negative coupling the factored map reaches an error of about 0.0035, while the unrestricted conditional map stays near 0.985, a gap that holds in both systems (Holm-adjusted $p \approx 0.003$, twelve seeds; Table 9). The differences persist across the evaluated representations, parameter maps, and solvers. This does not establish equal functional capacity or exclude all optimisation and discretisation effects. Every factorial cell stays bounded here (escape fraction 0), so the factorial isolates a fidelity and inversion effect rather than a boundedness effect. The boundedness contrast comes instead from the softened-gravity system of §4.1, where the unrestricted and $\lambda = 1$ energy-penalised models pass the 10^4 cap by step 500 and lose episodes to non-finite states by step 2000, while the symplectic model stays finite and below the cap.

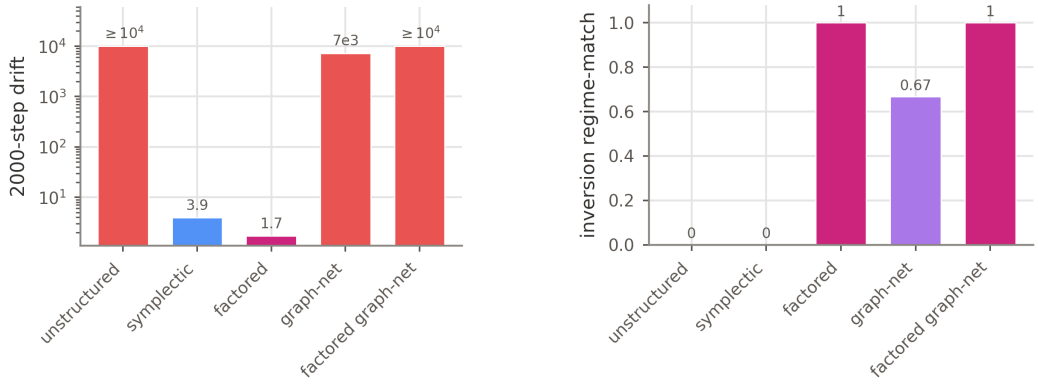
Table 8: Controlled factorial at step size 0.05 over twelve training seeds. Ranges span the two systems and the relevant model cells.

Comparison	Hamiltonian representation	Unrestricted vector field
Maximum true-energy error	0.00050 to 0.00084	0.314 to 0.332
Trajectory normalised error	0.00054 to 0.00169	0.0439 to 0.0586
Escape fraction	0	0

Table 9: Matched-state response at the unseen negative coupling.

System	Coupling-factored map	Unrestricted conditional map
Coupled oscillator	0.00348	0.9853
Modified oscillator	0.00382	0.9853

Graph-network response and long-horizon failure. Both graph variants fail the 2000-step test, so message factoring alone does not prevent their observed long-horizon failure. The source reports nearest-reference labels of 1 for the factored variant and 0 for the other variant. Without verified short-horizon finiteness and fidelity, those labels are descriptive and are not counted here as validated counterfactual success. Training-noise injection does not remove the reported long-horizon failure (Table 10).



(a) 2000-step drift

(b) inversion regime-match

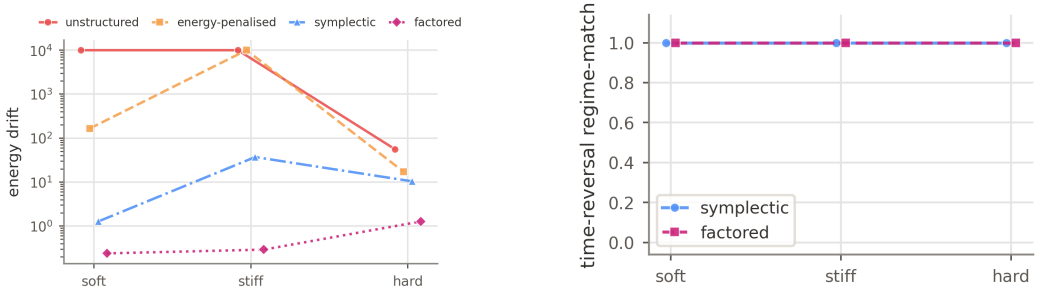
Figure 9: Graph-network long-horizon failure and reported response labels. The labels require separate short-horizon validity checks before being interpreted as counterfactual success.

Table 10: Graph-network summaries. Failure is separated from finite energy error; reported response labels are descriptive pending horizon-validity checks.

model	2000-step drift	regime-match
unstructured MLP (MLP)	numerical failure	0
graph-network simulator	long-horizon failure	0
factored graph-network simulator	long-horizon failure	1
symplectic model/factored symplectic model	bounded	–/1

C.3 C. Contacts and the engine (extends §4.5)

Contacts. On colliding disks (contact stiffness series: soft k=30, stiff k=300, impulsive), the factored symplectic model's energy stays bounded across the whole series and time-reversal (symmetry) inversion remains present (Table 11); parameter-inversion regime-matches (match 1) at soft/stiff, though its fidelity degrades (nMSE 0.76 → 2.9), and has no referent at the impulsive level (§C.4).



(a) energy drift across the stiffness series

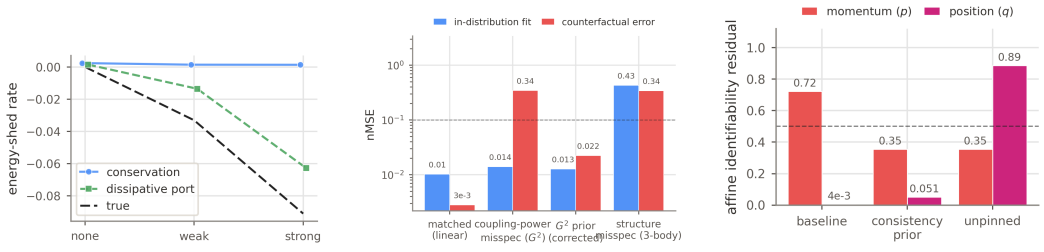
(b) time-reversal regime-match

Figure 10: Contact stiffness series: bounded energy and surviving symmetry-inversion where unconstrained models diverge.

Table 11: Across the stiffness series (factored symplectic model).

level	drift	sym-inversion match	param-inversion nMSE
soft	0.24	1	0.76
stiff	0.29	1	2.9
impulsive	1.27	1	undefined

C.4 D. THE BOUNDARY MAP (EXTENDS §4.4)



(a) Dissipation: conservation is incompatible

(b) A misspecified factored law

(c) Perception: momentum is the open axis

Figure 11: The boundaries. Conservation cannot shed energy under drag; a factored law of the wrong power fits in-distribution yet gives a wrong counterfactual; and a single frame recovers position but not momentum.

The impulsive-contact boundary. The stiffness-sign intervention used here has no specified counterpart in the impulsive hard-contact model. This does not rule out interventions on restitution, friction, or forcing in a separately defined hybrid law. The tested differentiable-restitution variant does not improve on the smooth model and loses time-reversal performance at the impulsive level; the free-learned impulse model diverges (Table 12). These results delimit the tested augmentation rather than all contact-capable structured models.

Table 12: Drift at the soft, stiff, and impulsive levels.

model	soft	stiff	impulsive
factored symplectic model (smooth)	0.11	0.29	1.06
diff.-restitution	0.92	16.9 $[0.25, 33.6]$	0.91 (loses retrace)
free-learned impulse	not evaluated	not evaluated	3360 (diverges)

Conservation is incompatible with dissipative dynamics. With linear drag the truth sheds energy (Fig. 11); symplectic model/factored symplectic model shed $\sim$0, and only a port-Hamiltonian variant more closely reproduces the reference shedding rate and inverts the drag sign (Table 13). Energy-shed rate is primary here (trajectory-MSE does not separate the regimes short-horizon).

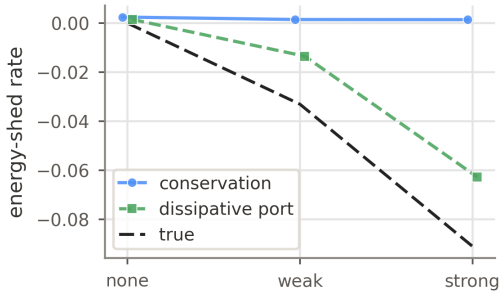


Figure 12: Conservative models have approximately zero energy-shed rate; the port variant more closely reproduces the reference rate.

Table 13: Energy-shed rate (negative = losing energy).

	weak drag	strong drag
truth	-0.033	-0.091
symplectic model/factored symplectic model	~ 0	~ 0
port-Hamiltonian	-0.014	-0.063

Supervised perception. With a soft-argmax encoder supervised to the oracle state, position is near-perfect and flat across a visual-difficulty series; the limit is object *speed* (per-body momentum error 0.01 slow $\rightarrow$ 0.87 fast), and downstream inversion is retained (factored symplectic model nMSE $0.062/0.064 \approx$ oracle 0.09). Reference trajectories are rolled from the true initial state so that perception error is counted rather than cancelled.

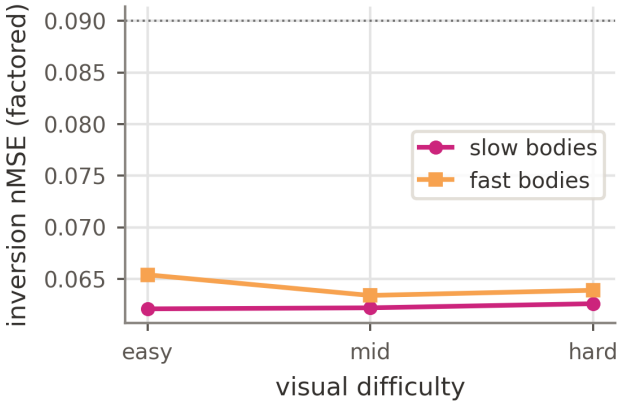


Figure 13: Supervised perception: position error remains stable; inversion is retained; momentum error increases with object speed.

Discovered perception leaves a large linear momentum-recovery error; a canonical kinetic recovers it. A latent learned end-to-end by prediction leaves a large linear momentum-recovery error (affine p-residual 0.72) with renderer-pinned position (q-residual 0.004); a canonical *fixed* kinetic recovers inversion with error comparable to the oracle result (3-seed, 500-epoch nMSE 0.065–0.092 $\approx$ oracle 0.067; backstop 0.045). A learned decoder is invalid against the backstop. The value is the 500-epoch, three-seed result; a 60-epoch run is undertrained and disagrees.

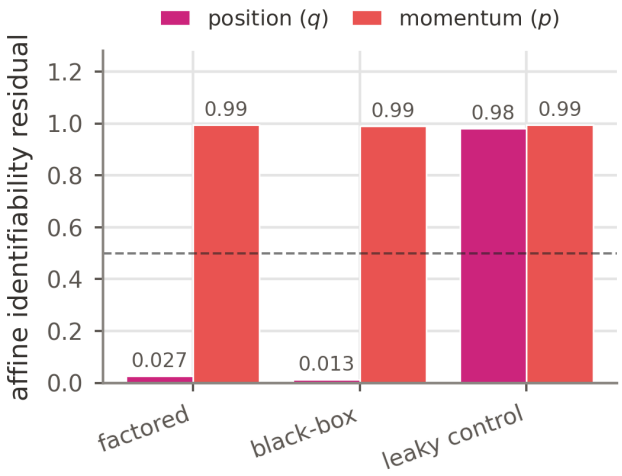


Figure 14: Discovered latent: momentum poorly recovered by an affine readout (high p-residual); the fixed canonical kinetic reduces inversion nMSE to 0.065–0.092, compared with the oracle value 0.067.

Momentum partially recoverable; position grounding open [directional/boundary]. A symplectic-consistency prior drops the momentum residual $0.72 \rightarrow 0.35$ (< 0.5, improved linear recovery) while keeping position (q-residual 0.05) and low inversion error (nMSE 0.10), at valid fidelity (backstop 0.149). Here v_{obs} is frame-derived, not oracle. Removing the renderer to *discover* position increases the residual (q-residual 0.89; backstop 1.84, invalid): the residual boundary is decoder-free position grounding.

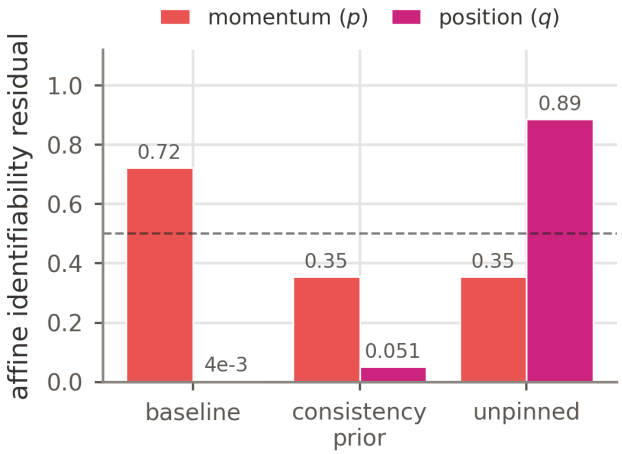
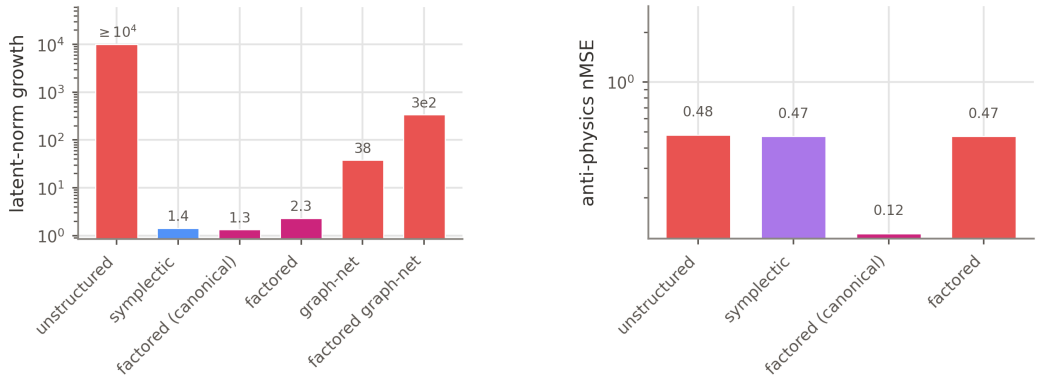


Figure 15: Symplectic-consistency improves linear momentum recovery (p-residual $0.72 \rightarrow 0.35$) while retaining position; removing renderer-based grounding produces a high position residual.

Pixel-based dynamics comparison under grounded perception [well-supported, scoped]. Trained end-to-end on synthetic renders with a shared fixed disk-renderer (only the dynamics differ), symplectic models stay bounded while unstructured MLP/graph-network simulator/factored graph-network simulator/JEPA-style predictor diverge, and only the factored fixed-kinetic model approaches the oracle inversion error (Table 14). The factored symplectic model (learned kinetic) fails like the non-factored models, reproducing the empirical recovery condition on pixels. The comparison evaluates dynamics *on* pixels with renderer-grounded position, not unsupervised inference

from pixels. The baselines are restricted to the evaluated architecture families and do not constitute a published leaderboard. Oracle-latent reconstruction is 0.22–0.53. The fully unsupervised branch is invalid (backstops 2.8–3.3) even though the decoder renders from oracle latents, a position-discovery failure that confirms, rather than resolves, the identifiability and position-grounding boundaries above. The decode-free JEPA-style predictor column here is the bare objective (no grounding). It is developed into a grounded hybrid, and into its own boundary, in App. C.8.



(a) stability (latent-norm growth)

(b) anti-physics inversion error

Figure 16: End-to-end on pixels: (a) latent-norm growth (log) diverges for all but symplectic models; (b) anti-physics inversion error is comparable to the oracle result only for the factored-plus-canonical-kinetic model. In-distribution backstops are within the valid range throughout (Table 14).

Table 14: End-to-end on pixels (grounded perception).

model	latent-norm growth	inversion nMSE
unstructured MLP	10^4	0.478
GNS / factored GNS	38 / 338	div.-contaminated
JEPA-style predictor (decode-free)	22	1.46 (diverges)
symplectic model	1.43	0.468
factored symplectic model (learned T)	2.3	0.468
factored fixed-kinetic model (canonical T)	1.33	**0.121**

C.5 E. Necessity, compositionality, generality, and scale (extends §4.5)

Necessity. A first screen yields a null result because always-active interactions provide no available shortcut. A second screen makes a shortcut available: a hybrid (factored pathway + free black-box term) assigns $\sim$42% of the force, with inversion degraded 5–10×. An L_2 penalty drives the share to 10^{-4} and restores fidelity (nMSE 0.43 $\rightarrow$ 0.09). These results support the shortcut-use interpretation within this physical system.

Cross-family. On Coulomb with mixed signs, a charge-product coupling has the lowest error on a held-out all-repulsive configuration (error 0.52 vs black-box 1.42, a modest $\sim$2.7× margin). A single factored model spanning a mixed gravity+Coulomb generator inverts both $\sim$120× better than a black-box that reads the couplings (nMSE 0.0035 vs 0.425).

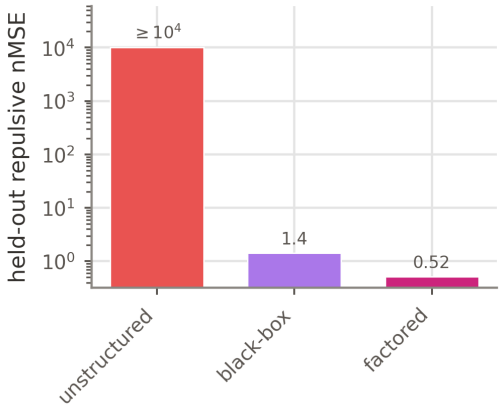


Figure 17: Held-out all-repulsive Coulomb: the charge-product coupling generalises across sign.

Data-efficiency. The factored symplectic model inverts at every training-set size down to 32 trajectories. Giving the non-factored symplectic model *more* data makes its inversion *worse* ($0.76 \rightarrow 2.67$ across $32 \rightarrow 512$). This is not undertraining at small data: the symplectic model's *in-distribution* ($G{>}0$) rollout stays bounded with low drift at every size (drift $\sim$1.4 at 32 trajectories), so it fits the training law well. Within this experiment, the inversion failure is therefore not explained by insufficient training data.

Capacity and composition [well-supported / directional]. Across a width sweep $32 \rightarrow 512$, the unstructured MLP diverges at every width and the factored symplectic model inverts at every width. Non-factored conservation worsens with capacity. Factored potential, dissipative port, and necessity penalty compose without interference under dissipation, with the necessity-fidelity effect muted in-regime (shortcut 13% vs the earlier 42%).

Calibration of in-regime planning. Planning an impulse through each model's rollout and executing it in the true physics, structured models' plans transfer (executed $\approx$ predicted) while unstructured MLP misestimates its own plans. The unstructured model reports the lowest predicted error but the highest realised error. Because confidence intervals for realised error overlap, the supported result concerns prediction–execution calibration rather than a control-performance advantage.

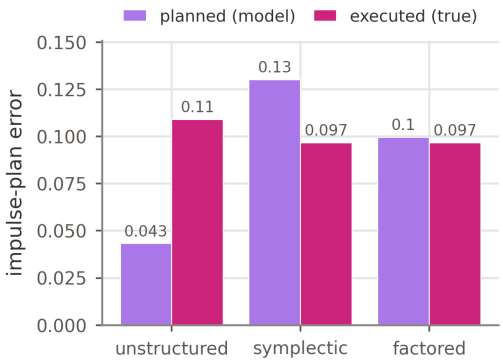


Figure 18: Planned vs executed error: structured plans transfer; the unstructured model misestimates its own plans.

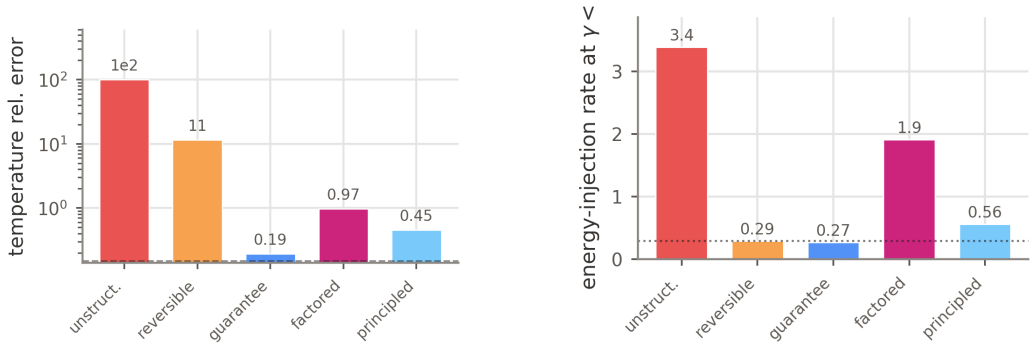
Rendering and noise screen. A structured model's rendered rollout is physically consistent (re-perceived energy conserved) whereas the unstructured model's bodies jump discontinuously, a re-visualisation of the stability gap in pixels, not a benchmark. Under i.i.d. initial-state noise to 30%

of the state s.d., the factored symplectic model keeps regime-match 1 and inversion nMSE $\sim$0.1, a screen (not perception error) that motivated the move to pixels.

C.6 F. Beyond conservation: stochastic dissipation and driven injection (extends §4.5)

This subsection extends §4.5 using the passive-friction formulation of Eq. equation 8 and the temperature metric of App. A. All aggregates are mean [95% CI] over $\geq$ 3 seeds; temperature error is capped at 100 and compared against the true simulator's *measured* equilibrium or steady-state temperature which targets agreement with the discrete reference rather than cancellation of model-dependent integration bias.

Structured fluctuation–dissipation models approach the target temperature; guarantee–invertibility tension. On the confined Langevin system (three bodies, three seeds, four (γ, T) settings), a structured model with a positive-semidefinite friction and fluctuation–dissipation noise reaches the target temperature with relative error 0.19 ($\leq$ 0.10 at strong drag), while an unstructured neural SDE *diverges* ($\hat{T}$ error $\sim 10^3$–10^4) and a reversible-only model *never thermalises* (11.4): the conservation$\rightarrow$stability analogue for the thermal case (Table 15). The same table shows the tension: the guarantee-side model has the lowest temperature error among the evaluated models but injects energy only at the reversible baseline (injection 0.27 $\approx$ 0.29, i.e. it *cannot* invert), whereas the factored-friction model injects $\sim 7\times$ the baseline (1.91) at the cost of being a worse thermostat without the passive-channel sign guarantee (Prop. 4).



(a) temperature error by model structure

(b) guarantee–invertibility trade-off

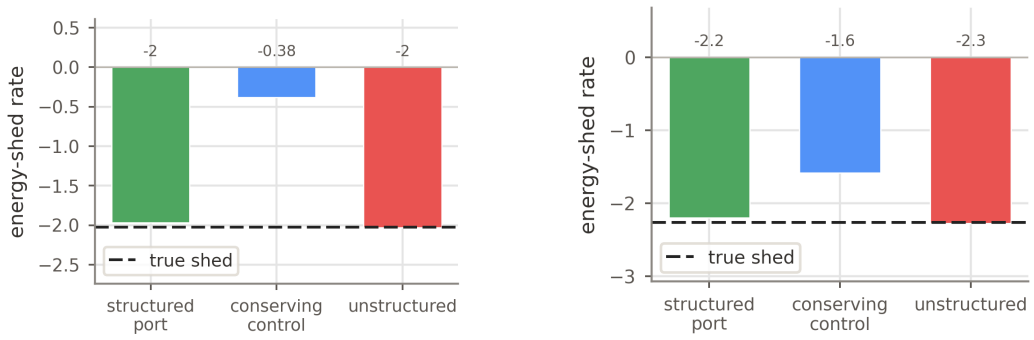
Figure 19: The structured FDT model has the lowest temperature error; the unstructured model diverges and the reversible-only model does not thermalise. The positive-semidefinite model cannot reverse its passive friction channel, whereas the signed factored model can.

Table 15: Temperature error and energy-injection rate at $\gamma < 0$ (confined Langevin; injection near the reversible baseline 0.29 means "cannot invert").

model	temperature rel-err	injection rate ($\gamma < 0$)
unstructured neural SDE	100 (reporting cap; diverges)	not evaluated
reversible-only	11.4 (does not thermalise)	0.29 (baseline)
structured, guarantee ($M \succeq 0$)	**0.19**	0.27 (cannot invert)
factored friction (invertible)	0.97	**1.91** (injects, no guarantee)
principled full $M = LL^\top$	0.46	0.56

Full and diagonal friction parameterisations. The full positive-semidefinite operator $M = LL^\top$ (Cholesky) with correlated fluctuation–dissipation noise retains the qualitative thermalisation result but has higher temperature error (0.46 versus 0.19) and is more difficult to train. In these mechanical experiments, the diagonal parameterisation therefore performs better; evaluation in non-mechanical dissipative domains remains future work.

Dissipation in MuJoCo, with linear and nonlinear friction. On the MuJoCo double pendulum in generalised coordinates, the structured dissipative port reproduces the true energy-shed rate for both linear and nonlinear (v^2) friction, while an energy-conserving control under-sheds (Table 16). The unstructured model also fits smooth deterministic friction. The comparison is therefore between the structured model and the conserving control. The distinction from the unstructured model occurs in the stochastic temperature results above. Under anti-friction ($\gamma < 0$), the port injects energy at rate $+10.1$ for linear friction but diverges under nonlinear friction in all seeds. This instability is consistent with Prop. 4.



(a) linear friction

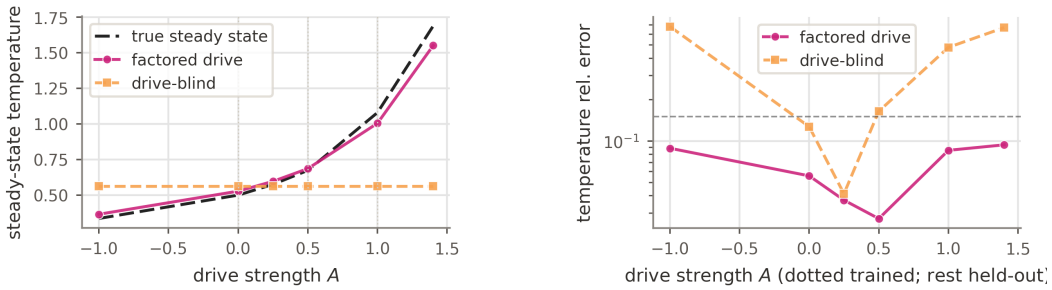
(b) nonlinear (v^2) friction

Figure 20: The structured port sheds the true energy-loss rate on the engine for linear and v^2 friction; the conservation control under-sheds.

Table 16: Energy-shed rate (negative = losing energy), model vs true.

friction	true shed	structured port	conserving control
linear	-2.03	$\mathbf{-1.97}$	-0.38 ($\sim$19%)
nonlinear (v^2)	-2.26	$\mathbf{-2.20}$	-1.59

Energy injection as a factored drive generalises over strength and sign. Modelling energy injection as an explicit factored *drive* A (Prop. 5), a structured model reaches the correct driven non-equilibrium steady state $T_{\text{eff}} = T\gamma/(\gamma - A)$ across trained and held-out drive strengths and the never-seen cooling sign (rel-err ≤ 0.094, Table 17), while a control without drive input remains at the bath temperature (rel-err climbing to 0.68 at the extremes) and the unstructured model diverges. Beyond the friction ($A \geq \gamma$) there is no steady state and the driven models exhibit unbounded growth (diverged fraction: factored drive 67%, unstructured 100%, control without drive input 0% because it does not represent the drive), consistent with the expected unbounded energy growth when net damping is non-positive. This parallels the gravity-coupling result: a factored external drive generalises across sign by construction, whereas a friction-sign flip is thermodynamically ill-posed and should not generalise at all.



(a) correct driven steady state

(b) generalising over drive strength/sign

Figure 21: The factored-drive model follows the true driven-steady-state curve across drive strengths and both signs; the control without drive input cannot track it.

Table 17: Driven-steady-state temperature rel-err ($*$ = trained drive; rest held-out).

model	A=−1	A=0*	A=0.25	A=0.5*	A=1*	A=1.4
factored drive	**0.088**	**0.056**	**0.037**	**0.027**	**0.085**	**0.094**
control (A omitted)	0.68	0.13	0.041	0.16	0.48	0.67
unstructured	diverges (rel-err 100 at all A)					

C.7 G. Misspecification and stronger baselines (extends §4.2 and §4.4)

Inaccurate counterfactuals under coupling-power misspecification. Every earlier inversion used a prior whose form matched the law. Here we deliberately misspecify it (softened 3-body gravity, factored prior $V = G\sum\phi_\theta$, force linear in G, 3 seeds). When the true coupling is instead *even* in G (force $\propto G^2$, so flipping G leaves it unchanged), the model fits in-distribution as well as the matched case yet predicts a sign flip the truth never performs, with no in-distribution warning. The correct G^2 structure restores the expected counterfactual, and a mismatched *interaction* structure (a 3-body term under a pairwise prior) instead produces poor in-distribution fit (Table 18). Parameter factoring therefore depends on the assumed functional form: coupling-power misspecification is not detected by in-distribution error here, whereas interaction-structure misspecification produces elevated in-distribution error.

Table 18: Misspecification series (nMSE; 3 seeds). "true-cf differs" ≈0 means the true counterfactual does not change under the flip.

law	prior	in-dist nMSE	counterfactual nMSE	true-cf differs
linear (matched)	factored	0.010	**0.003**	0.33
quadratic (power misspec)	factored	0.014	**0.343**	0.0
quadratic	factored-G^2	0.013	**0.022**	0.0
threebody (struct. misspec)	factored	**0.43**	0.34	1.23

Transformer and domain-randomisation comparisons. A permutation-equivariant set-transformer (bodies as tokens, self-attention), a relational unstructured model, *diverges* over 2000 steps (∼2 orders more slowly than the MLP, but still unbounded) and, evaluated at the same 150-step inversion horizon, does not invert (regime-match 0, counterfactual nMSE ∼0.3; Table 19). Thus, stability is associated with symplectic structure and inversion with factoring. The same distinction is observed with this additional baseline. Regime-match is primary here. At 50 steps, the attractive and repulsive trajectories remain close, producing a small transformer nMSE despite regime-match 0. This value reflects the short horizon rather than inversion. Domain randomisation, in which the black-box model is trained on both signs of G, inverts (regime-match 1) because the counterfactual

is then in-distribution, but it still diverges, so it gains no stability. The factored model is the only evaluated model that both inverts (zero-shot) and stays bounded.

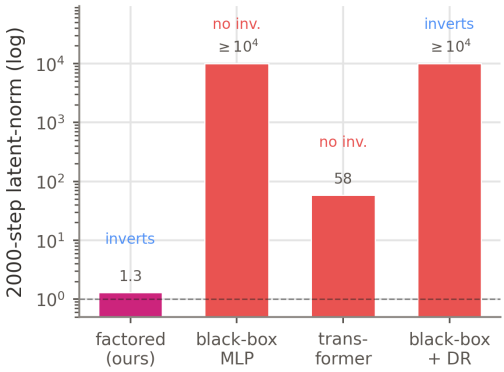


Figure 22: The factored model is bounded and inverts; the transformer diverges and does not invert; domain randomisation inverts after training on both coupling signs but diverges.

Table 19: Transformer + domain-randomisation (3 seeds; inversion read at the 150-step horizon). Regime-match is primary; nMSE secondary.

model	2000-step stability	inversion regime-match	counterfactual nMSE (150-step)	inverts short-horizon (50)?
factored HNN	**1.3**	**1**	0.002	yes
black-box MLP	10^4 (div.)	0	3.6×10^3	yes
transformer	58 (div.)	**0**	0.29	no
black-box + dom.-rand.	10^4 (div.)	1	1.5	yes

C.8 H. Grounding a decode-free objective, and the object-binding boundary (extends §4.5)

Encoder readout, scaling, and image-SSL. Is the advantage in the encoder or in the readout? VideoMAE's 0.13 improves on our own CNN's 0.45, but that comparison changes two things at once, namely the encoder *and* the readout (our CNN used the earlier pooled head, whereas the public encoders use the spatial cross-attention). We deconfound by running our *own* trained CNN, frozen, through the *same* spatial readout: it improves to 0.34, so the localisation-aware readout explains part of the gap, but it stays 2.6× worse than VideoMAE under the identical readout. The "better than our own encoder" claim therefore softens from ∼5× to a readout-matched ∼2.6×, and the residual earns a mechanism: the public video-SSL representation is richer than our task-trained CNN's, not merely better-read. The readout-limited reading was then tested at scale: the same frozen-encoder harness on V-JEPA2 at ViT-L (300M), ViT-H (600M), and ViT-g (1B), identical spatial readout and head budget. The best-rung gain over ViT-L is 1.02× (factored inversion nMSE 0.585 → 0.573, both rungs 10 seeds; ViT-H 0.60 at 3) against a pre-registered 1.5× weak-gain bar, and the escalation itself is instructive: a 3-seed ViT-g interim read 3× better (0.195) until the pre-registered 10-seed escalation surfaced the rung's own heavy-tail seed (3.6), collapsing the gain. The full dissociation reproduces at 1B (factored drift ∼3 bounded on 10/10 seeds; unstructured 0.77, diverges 5/10): perception-side scale does not buy the physics readout here.

Image-SSL and the momentum axis. A frozen *image*-SSL encoder, DINOv2 (Oquab et al., 2024), transfers only partially under the spatial readout: the stability dissociation holds (factored drift 2.8 vs unstructured 6.8) but the inversion-nMSE gap is small (0.62 vs 0.65), consistent with a per-frame encoder exposing position but not momentum. We tested that mechanism directly: a readout that reads a per-frame position from DINOv2's tokens and *differences* it across strided frames (momentum lives in differences, at the representation level) closes the gap to nMSE 0.096, matching

VideoMAE, while its unstructured twin stays at 0.52. So image-SSL encoders *do* transfer fully once momentum is differenced rather than read from a single frame. The apparent "video-encoders-only" advantage resolves into one momentum-visibility law. (A global pooled readout remains a confound throughout: VideoMAE + pool is null, nMSE 0.76.) The law was then *quantified* on the cited model itself (a supervised (q, p)-readout probe at matched head budget, 3 seeds $\times$ 4 temporal configurations): position readout is configuration-independent (nMSE 0.04–0.09 everywhere), while momentum readout splits exactly along the temporal-difference channel: V-JEPA2's native multi-frame clip reads momentum at 0.54 against 1.35 for the same clip with its centre frame repeated (temporal information ablated at identical compute; a $2.5\times$ advantage, pre-registered bar $2\times$), and DINOv2-single-frame sits at $1.29 \approx$ the frame-repeated video encoder (ratio 0.95). A video encoder therefore buys exactly the temporal-difference channel the momentum boundary demands, that is, the channel rather than the architecture (Table 4). This is encoder transfer on synthetic renders, not real-video object physics (joint learned detection-and-binding remains the open rung).

This block develops the decode-free JEPA-style predictor column of the pixel comparison into a grounded family (3 seeds, matched encoder/data/objective/anchor/budget, so the only variable within a pair is the dynamics module). The unstructured dynamics head has $36{,}748$ parameters and the factored fixed-kinetic head $4{,}417$ ($\sim 8\times$ smaller), so every hybrid win is against a strictly larger unconstrained competitor. CI lower bounds are clipped at the physical floor of each non-negative metric for display.

A label-free image anchor lets the physics prior separate on a decode-free objective. A decode-free JEPA-style objective (predict future *latents* against an exponential-moving-average target; no decoder) with the physics prior alone does not invert: the fixed-kinetic factored JEPA-style predictor fails (nMSE 1.46) and its latent grows (~ 22), reproducing the empirical recovery condition that the prior needs a fixed latent gauge to act on. Supplying that gauge *without a renderer or state labels* (a soft-argmax image centroid for position and a central finite difference for momentum) separates dynamics structure from grounding cleanly (Table 20, Fig. 23): the anchored unstructured model never inverts (match 0, nMSE 0.68), while the anchored factored hybrid inverts on every seed (match 1, nMSE 0.40) and stays bounded (1.39), $\sim 41\%$ lower error under matched everything. This partially closes the residual position-grounding boundary (position grounding without a decoder). Two limits: the anchor assumes calibrated per-body channels, and its centroid/finite-difference grounding is noisier than the renderer, so the hybrid (0.40) trails the renderer-grounded factored fixed-kinetic model (0.12): a reliable *regime* inversion decoder-free, not near-oracle fidelity. A grounding-quality control isolates which of the two differences drives that gap: replacing the noisy centroid with an exact-position anchor (same world coordinates and finite-difference momentum, minus centroid and occlusion noise), holding the JEPA objective fixed, does *not* move the number (nMSE 0.405 vs 0.401, match 1 every seed), so the anchor is exonerated and the residual gap to 0.12 is a property of the decode-free objective, not of grounding quality. A composited-RGB ladder confirms the operative ingredient is persistent temporal grounding, not per-body channels: naive RGB differencing breaks both variants, but a seven-frame robust tracker (palette demixing, visibility-weighted trajectory fitting; a controlled probe, not a general perception method) restores the hybrid advantage (nMSE 0.45 vs 0.82, every seed).

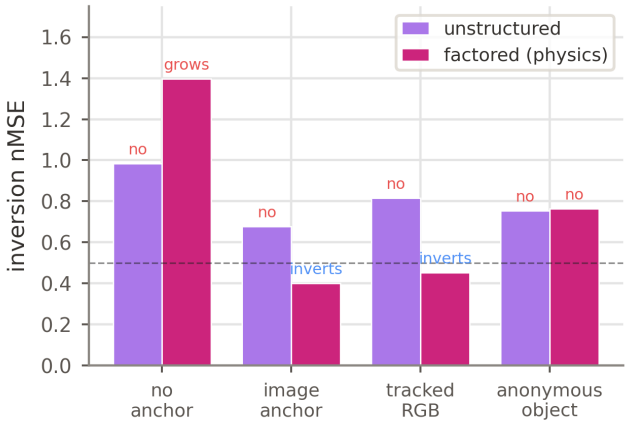


Figure 23: Decode-free inversion nMSE across the grounding ladder: the image-anchored factored hybrid inverts (match 1) where its unstructured twin does not, on channels and on tracked RGB; with object identity removed the advantage disappears.

Table 20: Decode-free JEPA family (3 seeds). "match" is anti-physics regime-match; nMSE the inversion error; growth the 60-step latent-norm ratio.

variant (grounding)	dynamics	match	inversion nMSE
none (no anchor)	unstructured	0	0.99
none (no anchor)	factored	1	1.46 (grows)
image anchor	unstructured	0	0.68
image anchor	factored	**1**	**0.40**
tracked RGB	unstructured	0	0.82
tracked RGB	factored	**1**	**0.45**
anonymous object	unstructured	0.33	0.76
anonymous object	factored	0.33	0.76

Removing object identity relocates the blocker to learned object binding. When per-body channels *and* colour identity are removed (all bodies share one monochrome frame, bound across time by anonymous peak detection and matching), the hybrid advantage vanishes: both the structured and unstructured variants match the flipped law on only $1/3$ seeds with indistinguishable error (nMSE $\sim$0.76), because close encounters make the momentum/identity assignment ambiguous. The physics dynamics are not the blocker; learned persistent *object binding* (long-context slots or tracks) is. The label-free anchor moves the boundary from decoder-free *position* grounding down to *object* binding, and this anonymous-object result marks where the hand-engineered binder fails; the next two results attack that blocker directly.

A learned binder resolves the anonymous-object boundary, and the physics prior disambiguates the learned binding. The anonymous-object collapse traces to the hand-engineered binder: it detects N blobs and assigns identity across frames by a hard minimum-acceleration match, which is ambiguous at a close encounter, so the prior is grounded on mis-bound (q, p). Replacing that match with a learned, differentiable, confidence-aware soft-assignment (a "learned tracker"; the detector is unchanged, and the binder is the encoder, its grounded centroids rolled and matched under the decode-free objective) resolves it (Table 21, Fig. 24): the learned-tracker factored hybrid inverts on every seed (match 1, nMSE 0.15, bounded growth $\sim$1.4), while its unstructured twin does not (match 0, nMSE 0.64) and the hand-engineered floor does not (match 0.33, nMSE 0.76). Extended to the ten-seed standard, the dissociation is exact: factored inverts on $\mathbf{10/10}$ seeds and the unstructured control on $\mathbf{0/10}$ (one-sided exact binomial $p \approx 10^{-3}$ each). This is the same double dissociation once more, now on *learned* binding: with the same binder, factored inverts and unstructured fails. It holds whether the binder is trained through the physics loss (0.152) or self-supervised

on temporal smoothness and then frozen (0.155), so physics need not train the binder end-to-end to obtain the separation; and 0.15 is below the channelised hybrid (0.40), approaching the renderer-grounded 0.12, because the compact anonymous disks detect cleanly and only the matching needed fixing.

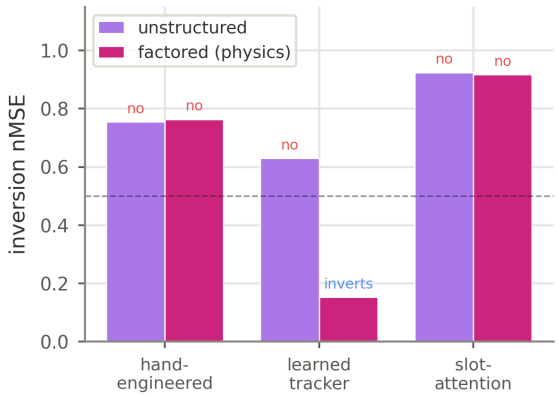


Figure 24: Inversion nMSE on the anonymous world across the binding ladder: only the learned-tracker factored hybrid inverts (below the 0.5 guide), where the hand-engineered floor and slot-attention do not.

Table 21: Learned binding on the anonymous world (3 seeds). "match" is anti-physics regime-match; nMSE the inversion error.

binder	dynamics	coupling	match	inversion nMSE
hand-engineered	factored	—	0.33	0.76
learned tracker	unstructured	e2e	0	0.63
learned tracker	factored	e2e	**1**	**0.15**
learned tracker	factored	twostage	**1**	**0.16**
slot-attention	factored	e2e	0	0.92
slot-attention	unstructured	e2e	0	0.93

Joint learned detection-and-binding is the next rung. A slot-attention binder, which must learn *detection* as well as assignment, does not reach the inversion (match 0, nMSE 0.9–1.0 across factored/unstructured and both couplings). Learning the binding on top of a fixed detector is therefore solved; learning detection and binding jointly and decode-free is not. The perception frontier accordingly moves one rung further: from persistent object binding, now resolved by a learned soft-assignment, to *joint* learned detection-and-binding. The learned quantity in the tracker is only the assignment (4,609 params, detector fixed); the slot-attention binder is 105,728 params.

The grounded approach transfers to a frozen public video encoder we did not build. To test whether the pixel result depends on our specific encoder, we replace the trained CNN with a frozen public encoder (weights not trained; only a small head and the dynamics train) in the tracked setting, reading out (q, p) with N object queries that cross-attend over its patch tokens. On two frozen *video*-SSL encoders, VideoMAE (Tong et al., 2022) and the V-JEPA2 model (Assran et al., 2025) this paper's introduction cites (both arms now at the full 10-seed standard: factored inverts $10/10$, exact binomial $p = 0.001$), the factored approach inverts on every seed with bounded drift while its unstructured twin is far worse and unstable. Over 10 seeds VideoMAE inverts at nMSE 0.13 (oracle 0.067) against 0.65 unstructured (which diverges on a seed); V-JEPA2 inverts at 0.59 against 0.82 unstructured (the 10-seed expansion widens the 3-seed 0.25 with two heavy-tail seeds; the unstructured twin diverges $5/10$, drift up to $\sim$3100). This is the same dissociation on representations we did not build (Table 4).

The world-model half of V-JEPA2, characterised, with the positive control as the finding. Our external-encoder result tested the public model's *encoder*; here we test its *predictor*: the official action-conditioned V-JEPA2-AC (ViT-g encoder + 300M frame-causal predictor, frozen, official release weights), rolled autoregressively on our rendered physics (native 256px, 16-frame context at the dynamics cadence) with action tokens held at no-op, in both the zero pose and the release example's real end-effector pose (disclosed; a no-op stream is itself part of the domain gap for a model trained on robot manipulation). A pre-registered positive control gates the physics probes: the predictor's one-step latent prediction on our renders must beat a copy-last-frame baseline by a stated margin before any probe is interpretable. It does not: pred/copy $= 0.978$–0.986 across 3 seeds and both no-op variants, against a certified-fair baseline (consecutive-frame representations sit at 82% of the *unrelated*-frame distance, since bodies traverse 8% of the scene per frame, so there is real predictable signal that copying does not capture). The same gate shows the *encoder* half does encode the state: our spatial readout recovers (q, p) from its tokens at 0.28–0.43 nMSE. Per the pre-registered rule, the imagined-rollout probes (energy drift, counterfactual tracking, second-law flags) are VOID and reported as such, not as model failures. The finding is that the representation transfers across the domain gap (consistent with the encoder-transfer result) while the *predictive* half, under no-op conditioning far from its training distribution, adds nothing over frame persistence. This is a characterisation rather than a strawman: the domain gap (Droid tabletop $\rightarrow$ rendered N-body) is the headline variable, the fairness artifact and both no-op variants are in the record, and a same-domain test would require fine-tuning the predictor, which is out of scope for a frozen-model claim.

Composing a public encoder, learned binding, and the physics prior, the boundary that localises the residual. The learned-binder result and the frozen-encoder result each remove one obstacle separately: the frozen-encoder result keeps the hand-engineered 7-frame tracker anchor, and the learned-binder result's detections are *real* image centroids. We composed them so that no part of the perception stack is author-built: a frozen VideoMAE reading the *anonymous* world, a *learned* soft-argmax detector replacing the hand-built centroids, and the learned tracker, under a factored prior. It does *not* compose. The factored cell is unstable across seeds (inversion nMSE 0.8–7.5, latent growth 1.2–$42\times$; the same seed even flips between bounded and divergent under GPU nondeterminism, i.e. it sits on the edge of stability) and never reaches the clean inversion of the learned-binder result (0.15) or the frozen-encoder result (0.13). The unstructured control is worse (nMSE up to 315). The factored prior keeps only a weak stability edge. The boundary is informative: this composition removes *both* physical-grounding sources that made the two earlier results work (real centroids; a built anchor), and the physics prior, which can disambiguate binding and transfer across frozen encoders *given* grounding, does not bootstrap grounding in this frozen-token configuration. This is the joint learned detection-and-grounding residual at the representation level: the tested joint detection-and-grounding configuration remains unresolved.

Damping-rate ordering from video [partial]. A grounded seven-frame tracker supplies latent (q, p) for a disk world with conservative or damped dynamics. The model uses the positive port $p \leftarrow p \exp[-\mathrm{softplus}(a)\Delta t]$. The effective rate is $r = \mathrm{softplus}(a) > 0$; the raw parameter a is not constrained to be nonnegative. This update contracts canonical quadratic kinetic energy during the isolated port, not arbitrary total energy.

The source reports rate diagnostics 1.25 for damped video and 0.84 for conservative video, against reference rates 1.5 and 0, across five seeds. Their ordering is recovered, and a label-free window head tracks a medium switch. The parameter-to-rate reporting convention must be checked before interpreting these diagnostics as calibrated physical rates. The reported σ values are transformed effective rates, $\sigma = \mathrm{softplus}(\log \sigma)$ read from the trained port (not the raw parameter and not a fitted diagnostic), in units of inverse physical time at the frame cadence; the estimator validation below gives the grounding comparison and the untested step-size sensitivity. The conservative diagnostic remains nonzero; per-clip arrow-of-time detection is not demonstrated (prediction nMSE approximately 0.9).

Oracle-grounding and reversed-video controls. Replacing grounded pixels with true simulator states leaves the conservative diagnostic near 0.885, versus 0.84 from pixels (five seeds). Oracle grounding therefore does not eliminate the floor. This directs attention to model, objective, optimisation, and temporal-resolution limitations rather than grounding alone; it does not uniquely identify the cause. On reversed damped clips, the constrained-port momentum-loss floor rises from 1.03 to 1.24, while the sign-free control changes from 0.84 to 0.87. The pooled $1.42\times$ contrast misses the

declared $2\times$ threshold. The retained findings are ordering and a channel constraint, not accurate conservative-rate recovery or demonstrated video-level refusal.

A grounded energy-balance estimator [pre-registered; estimator amended before the run, disclosed]. The preceding experiments show that oracle grounding does not eliminate the learned-model floor; they do not uniquely identify its cause. A separate estimator uses grounded positions and the known interaction law. For unit masses and linear velocity drag, it regresses the windowed energy balance $\Delta E_j \approx -2\gamma \bar{K}_j \Delta t$, with $E = K + V(q; G)$. For general masses under $\dot{p}_i|_{\text{drag}} = -\gamma p_i/m_i$, the dissipated power is instead $-\gamma \sum_i \|p_i\|^2/m_i^2$. A time-constant additive measurement bias cancels in the energy difference; time-varying tracking noise and discretisation error need not cancel. *Estimator validation.* The learned-rate readout stores the transformed rate $\sigma = \text{softplus}(\log \sigma)$, applied as $p \leftarrow p\, e^{-\sigma \Delta t}$ with Δt the frame cadence (M simulator steps), so σ is a rate per unit physical time; masses are 1 in this world, so the p/m and p/m^2 conventions coincide. The ~ 0.85 positive bias of the learned conservative rate is *not* tracking noise: the oracle-grounded arm (true positions) gives $\sigma_{\text{cons}} = 0.85$–$0.93$ and the pixel-grounded arm 0.80–0.89 (five seeds each), so the bias is the co-training transient described above, which the grounded energy-balance estimator removes ($\hat{\gamma}_{\text{cons}} = -0.006$). The damped rate is recovered at 1.22–1.30 against a truth of 1.5 (about 15% low) under both groundings. Sensitivity to the frame cadence was not tested, so the video result is an ordering and an approximate rate, not a calibrated measurement. One physics amendment was made to the pre-registration *before* the full run, on smoke evidence: the originally specified kinetic-envelope form fails on self-gravitating worlds, where K *grows* even without drag (virial infall, measured $0.066 \to 1.52$); the energy-balance form replaced it with all gates unchanged. Result: pixel-grounded $\hat{\gamma}_{\text{cons}} = -0.006$ and $\hat{\gamma}_{\text{damped}} = 1.63$ (truth 1.5), both gates pass at 5 seeds; the port carries $\sigma = \max(\hat{\gamma}, 0)$, so the estimator supplies the rate and the architecture the guarantee. The estimator is label-free by construction (per-episode, no medium index), and a sliding version reads a mid-clip medium switch cleanly ($\hat{\gamma}$ median $-0.05 \to 1.62$ across a conservative→damped concatenation), a label-free discovery pattern completed at the estimator level. Scope: the meter here uses the known interaction law $V(q; G)$; on real video, selecting the form that *becomes* the meter is an open form-selection problem. On time-reversed footage the estimator reads $\hat{\gamma} = -1.66$ while the $\sigma \geq 0$ port clamps its carried rate to zero, giving arrow-of-time *rate* measurement with architectural refusal, extending the classification-only contrast of Pickup et al. (2014) and Wei et al. (2018). The training-floor refusal channel was pursued through two pre-registered designs and is now *retired as an instrument*: the original sign-free control reached $1.93\times$ against the $2\times$ bar (not demonstrated under the pre-registered threshold), and the exponential-twin control (the port's exact unrestricted twin, the cleanest nested pair) inverted the prediction for an identified reason: a fixed growth rate amplifies the core's error by $e^{|\hat{\gamma}|\Delta t}$ per step (compounding $\approx 6.9\times$ in squared error over the horizon; predicted floor 6.6, measured 5.5), so under exponential growth a training floor measures *error amplification*, not rate-fit, so both models degrade on reversed footage on $10/10$ seeds and the channel cannot discriminate. This is a metric retired by its own control; the refusal claim rests on the estimator channel and the architectural guarantee. Finally, a distill-initialisation probe on the composition boundary (detector distilled from the centroid detector, then released with EMA, gradient clipping, and $0.1\times$ detector learning rate) *removes the instability* (latent growth 1.2–$1.3\times$ on every seed, versus the bounded↔ divergent flipping) while inversion plateaus at nMSE 0.73, and the residual is now *localised*: the cured detector's median position error against the centroid teacher is 0.71–0.75 world units, larger than the disk radius (0.6), so frozen tokens fail at *localisation* on this anonymous world before binding even starts (measured on the trained stack; the tokens' information ceiling is not probed). That number is the opening problem for future work.

D CLAIM–EVIDENCE MAP

This appendix maps each claim of the paper to the evidence that supports it and the class we assign that evidence.

Claim	Evidence	Class
In the evaluated conservative systems, conservation by construction is associated with lower long-horizon energy error than the evaluated unstructured baselines; neural-ODE finiteness shows that structure is not necessary for bounded finite-horizon states.	Bounded vs diverged energy and robustness to tuning; 10k-step saturation.	Well-supported.
Within the evaluated model families, law inversion occurs with factored coupling and not with conservation alone.	Non-factored fails while factored inverts in the reported fidelity-valid comparisons; the double-pendulum engine and DeLaN/SymODEN comparison. Graph-network response labels are descriptive pending short-horizon validity checks.	Well-supported.
Under grounded pixel training, a canonical fixed kinetic is additionally required for inversion in the evaluated factored models.	The discovered-perception and pixel-comparison results (learned-kinetic factored fails like non-factored).	Supported.
The results extend to contact dynamics and a generalised-coordinate double-pendulum engine.	Contacts; the engine with learned $M(q)$.	Supported.
The evaluated boundaries include hard contacts, dissipation, reversibility, and ungrounded visual-state discovery.	Contacts; dissipation; the reversibility flip; perception; the unsupervised pixel branch (invalid).	Supported as boundaries.
When an explicit shortcut is available, penalising its use increases reliance on the factored pathway and improves inversion fidelity.	The shortcut dose–response; the null result without an available shortcut.	Supported.
The factored model exhibits object-level transfer and cross-family generalisation within the tested ranges.	N-transfer with the regime persisting to 32; Coulomb and the mixed generator.	Supported (degradation reported).
The dissociation is retained under end-to-end pixel training with grounded perception.	The end-to-end pixel comparison.	Supported, scoped.
Discovered-from-pixels perception does not provide position grounding through the objective alone; a label-free image anchor supplies it without a decoder, and a learned soft-assignment binder then resolves object binding, leaving joint learned detection-and-binding as the open rung.	The discovered-perception results show the objective alone fails; a decoder-free image anchor grounds position and the physics prior then separates; a hand-engineered binder collapses on anonymous objects; a learned tracker + the physics prior inverts (match 1, nMSE 0.15); slot-attention, learning detection too, does not yet reach it.	Supported (boundary moved two levels down).

continued on the next page

Claim	Evidence	Class
The structured FDT model has the lowest stationary-temperature error among the evaluated models.	The structured model is correct while the unstructured model diverges and the reversible-only model never thermalises; this holds under noise.	Well-supported.
Under the passive-friction formulation of Prop. 4, a global positive-semidefinite friction guarantee is incompatible with friction-sign inversion.	Prop. 4; the passive channel cannot anti-damp, whereas sign reversal relinquishes that channel guarantee; the double-pendulum engine.	Well-supported (with proof).
Energy injection, as a factored drive, generalises over drive strength and sign.	The driven steady state at trained and held-out drives, including the unseen sign ($A \geq \gamma$ produces the expected divergence).	Well-supported.
Accurate counterfactual inference from factoring depends on correct specification of the coupling's functional form.	Wrong coupling power gives good fit but a wrong counterfactual with no warning, and the correct structure restores the result.	Well-supported (a boundary).
The dissociation is reproduced with the evaluated graph-network and transformer baselines; domain randomisation inverts only when both coupling signs are included in training and does not provide bounded rollouts.	The transformer diverges and does not invert at the inversion horizon; domain randomisation includes both signs in training and does not provide stability.	Well-supported.
On a decode-free future-latent objective, the physics prior separates once a label-free image anchor fixes the latent gauge; grounding *quality* is not the limiter.	The anchored factored hybrid inverts every seed at nMSE 0.40 vs anchored unstructured 0.68; an exact-position anchor does not improve on the noisy centroid, so the residual gap to the renderer-grounded 0.12 is a property of the objective.	Well-supported.
The grounded factored approach does not depend on our own encoder: it transfers to frozen public video encoders, one within $2\times$ of oracle fidelity (VideoMAE 0.13 vs oracle 0.067) and one at $\sim 8.8\times$ (V-JEPA2 0.59), with a localisation-aware readout.	Frozen VideoMAE and V-JEPA2 + spatial readout + factored invert at nMSE 0.13 and 0.59 vs unstructured 0.65/0.82 and diverging [10 seeds throughout]; a global pooled readout misses it; readout-matched, VideoMAE stays $2.6\times$ better than our own frozen CNN; scale-flat across V-JEPA2 ViT-L/H/g, gain $1.02\times$; a frozen image encoder transfers fully once momentum is differenced across frames.	Well-supported (encoder transfer; synthetic renders).

E EVIDENCE CLASSIFICATION

We classify evidence as **well-supported** when effects are large, rest on an exact test or a clear per-seed separation rather than on the non-overlap of marginal intervals, and replicate across the stated architectures, seeds, or epochs. Findings in this class are conservation-to-stability; factored-coupling-to-inversion (across MLP, Hamiltonian, Lagrangian, and graph-net; pixel-trained under grounded perception); the canonical-fixed-kinetic condition; the capacity and data-efficiency results; the misspecification boundary; the transformer / domain-randomisation dissociation; the grounded decode-free hybrid, where an image anchor lets the physics prior separate on a JEPA-style objective; the learned-binding crack, where a learned soft-assignment binder + the physics prior inverts the anonymous-object world while its unstructured twin fails; the external-encoder transfer, where the factored approach inverts on two frozen public video encoders we did not build, one within $2\times$ of oracle fidelity; and, beyond conservation, structure-to-correct-temperature, the guarantee/invertibility

tension (with proof), and the factored-drive generalisation. We classify evidence as **directional** when only a single configuration was evaluated or the per-seed separation is not clear. Findings in this class are the reversibility effect on the engine; the Coulomb margin ($3\times$ vs 15–$40\times$ for gravity); N-scaling ($N{=}32$); the necessity dose-response; in-regime control (overlapping CIs); the perception partial-recovery; and the engine energy-shed and noise robustness. Two further tags complete the vocabulary used in the findings table and claim map: **supported** marks effects that replicate but with a scoped configuration, a qualified control, or only some of their pre-registered gates passed (stated in the row); and **boundary** marks a mapped limit, a place where the approach demonstrably stops working, reported as a result of equal standing. A **pre-registered negative** is a pre-registered hypothesis whose gates failed, kept at full strength because the failure is the finding (the energy-aux unification is the instance).

F Pre-registration and reproducibility

The protocol, including the metrics, the equal-capacity rule, the non-finite cap, and the permitted conclusions, was fixed in dated documents ahead of the corresponding runs, and serves as a prospective specification and as provenance for the results. The following were recorded before the corresponding experiments: the fixed-kinetic and decorrelation designs; the perception and canonical-latent design; the pixel-based comparison and its success criterion; the necessity analysis; and the stochastic-dissipative design, including the initial go/no-go experiment and temperature as the primary metric. The guarantee–invertibility trade-off and the anti-friction boundary interpretation in the engine and driven experiments were specified in advance. For the later arc, the specifications are dated ahead of their results: the oracle-grounding and refusal arms of the video dissipation experiment (the first of which *refuted our initial grounding-noise attribution*); the energy-auxiliary unification (all three predictions failed under the pre-registered thresholds); and the final estimator and its gates, including one physics amendment recorded *before* the full run (the kinetic-envelope form fails on self-gravitating worlds, and the energy-balance form replaced it with the gates unchanged) and the exponential-twin control whose outcome retired the training-floor instrument. The video world-model characterisation was specified in advance, with its thresholds fixed before any run; its positive control was registered as the experiment's power condition, and its firing is the reported finding. The frozen public components used across the external-encoder experiments are DINOv2-small, VideoMAE-base, V-JEPA2 (ViT-L/H/g), and V-JEPA2-AC (ViT-g), each from its official public release pinned to a fixed version. *Statistics.* Aggregates are the mean over at least three seeds, and at least five seeds at scale. When five or more seeds are available, the reported bracket [lo, hi] is a normal-approximation 95% confidence interval, mean $\pm$ 1.96 SE, a normal-approximation interval that we label as such wherever it appears; it is not an observed range, and five seeds do not by themselves justify the approximation, so for the core comparisons we report the per-seed values and, where the design is paired, the paired difference with its interval. Significance statements rest on exact tests with a stated null: the binomial regime-match test has null $\Pr(\text{match}) = \frac{1}{2}$ per seed, and the paired sign-flip tests of the factorial have the null of exchangeable signs of the per-seed differences, Holm-corrected across the declared family. Display clipping must be distinguished from numerical failure. A finite value above the temperature-error threshold of 100 or drift threshold of 10^4 remains a finite observation. A non-finite state or diagnostic is a numerical failure and has no measured finite error. We do not interpret ratios to substituted failure sentinels quantitatively. Recomputing finite-only aggregate statistics requires the underlying per-run records. *Horizons.* Stability is evaluated at 2000 steps, $100\times$ the training horizon. Inversion error is evaluated at 150 steps, within the pre-Lyapunov window used for all models. At longer horizons, trajectory MSE saturates even for accurate chaotic-system models. *Regime-match* is therefore the primary inversion metric and nMSE the secondary metric. The released reproducibility materials comprise per-experiment model checkpoints, aggregate result tables, the scripts that regenerate the figures, and a fixed software environment.

Table 23: Compute and hyperparameters (defaults; per-experiment overrides in the released configs). All runs use Adam and one-step prediction loss unless noted. Each run uses one GPU, with runs distributed across eight GPUs.

setting	value
optimiser / learning rate	Adam / 10^{-3}
batch size	1024 (one-step transitions)
epochs	150–300 (200 typical)
seeds	≥ 3 (≥ 5 at scale)
capacity (MLP / HNN / factored)	$\sim$138k parameters (matched)
transformer	2 layers, 4 heads, width 96 ($\sim$90k)
decode-free JEPA pair	shared 383k encoder; unstructured dynamics $36{,}748$ vs factored $4{,}417$ params
learned binders	tracker $4{,}609$ (assignment only) vs slot-attention $105{,}728$ params
frozen encoders	DINOv2 22M / VideoMAE 86M / V-JEPA2 ViT-L/H/g 0.3–1B frozen; head + dyn. trained
integrator	leapfrog (dt=0.002); Tao (Tao, 2016) for the engine
rollout horizons	stability 2000 steps; inversion 150 steps
hardware / time	one GPU per cell; $\sim$1–4 min/cell (oracle), longer on pixels