WHERE STRUCTURE ENTERS A WORLD MODEL: FROM PREDICTIVE FIT TO PHYSICAL ROLLOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive world models learn to forecast observations, but accurate short-term predictions do not necessarily produce physically coherent behavior over longer horizons. We investigate how incorporating physical structure can strengthen world models while retaining their underlying predictive objectives. Across token-based models of planetary motion and JEPA-style visual models of mechanical systems, we study conserved-quantity supervision, symmetry-aware representations, temporal consistency, and learned structured dynamics. In planetary motion, angular-momentum supervision improves autonomous next-token forecasting, while the benefits of temporal regularization depend on the accompanying physical objective and representation. In visual models, generating rollouts through supervised energy-based dynamics heads produces more faithful trajectories and phase-space evolution. Together, these findings highlight the importance of how physical structure participates in prediction: improvements in an auxiliary physical readout or a local acceleration estimate do not automatically translate into more reliable long-horizon behavior. Evaluations across initial conditions and physical parameter shifts further reveal where these benefits persist and where generalization remains limited. Our study supports physical structure as a practical means of improving predictive world models and provides a unified framework for understanding its effects on representations, learned transitions, and autonomous rollouts. The central lesson is that effective physical inductive biases should help models both represent physical regularities and use them to generate future states.

1 INTRODUCTION

World models used for planning must predict what happens after observations stop. Small errors can accumulate until a pendulum loses its oscillation or a planet leaves its orbit. Can physical structure help a predictive model sustain the motion of a system while relying on its own predictions?

We add physical structure in four ways: supervising physical quantities, building relative geometry into the model, encouraging consistent representation changes over time, and learning mechanical equations for forecasting. These choices guide training or forecast generation. Prior work examines what models represent (Li et al., 2023, Mitchell & Krakauer, 2023) and how consistently they behave (Vafa et al., 2024). We ask how physical guidance can improve the motion they generate.

We study two model families. A joint-embedding predictive architecture (JEPA) infers motion from images, progressing from oscillating pendulums and springs to damping and coupled double-pendulum motion. For planetary motion, a separate autoregressive transformer receives positions directly and generates coordinates through next-token prediction (NTP). Together, these studies test physical guidance when the model must infer state and when state is observed.

These four forms of structure lead to three main findings:

- **Physical guidance improves generated trajectories.** Learned mechanics improves visual rollouts across four systems. Angular-momentum supervision improves planetary forecasts even though its auxiliary head is unused during generation.
- **Geometry improves forecasting within NTP.** A pairwise adapter improves the matched relative-coordinate model while retaining the next-token objective and decoder.

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.

- **Temporal benefits depend on the representation and task.** Controlled comparisons identify gains in forecasting and force learning, including force-direction transfer under eccentricity and stellar-mass changes.

2 Background and Related Work

Learning what changes over time. JEPA learns to predict features of observations, with V-JEPA extending this objective to masked video (Assran et al., 2023, Bardes et al., 2024). Temporal objectives guide how those features change. NextLat predicts the next latent, while Temporal Difference in Vision (TDV) uses observed image changes to predict representation changes (Teoh et al., 2025, Daithankar et al., 2026). We adapt TDV to planetary motion. Related work introduces spatial smoothness, stability, and temporal locality into GPT-2-scale models through continuous coordinate regression, noisy contexts, and restricted attention (Liu et al., 2026). Our models retain discrete NTP and its causal context.

Using physics to guide prediction. Physical information can shape a learning signal or the rule that advances state. Equation-residual penalties (Raissi et al., 2019), differentiable solvers (Hosseini Dashtbayaz et al., 2025), and physical controls (Gillman et al., 2025) offer different ways to make this connection. Hamiltonian and Lagrangian networks learn functions such as energy and derive motion through their derivatives (Greydanus et al., 2019, Cranmer et al., 2020). Dissipative extensions capture energy loss (Sosanya & Greydanus, 2022). Context-sensitive imagination (Röder et al., 2025) and physical-state recovery from predicted features (Rao et al., 2026) also link physical structure to forecasting. We compare physical targets used during NTP training with supervised mechanics used to generate visual rollouts.

Making geometry part of the model. For a planar position r_i, a rigid transform combines translation a with an orthogonal matrix Q. The set $O(2)$ contains all planar rotations and reflections:

$$r_i' = Qr_i + a, \qquad Q \in O(2), \quad a \in \mathbb{R}^2. \tag{1}$$

Invariant scalars stay unchanged. Equivariant vectors rotate or reflect with their inputs. Relative coordinates remove shared translation, while summed messages remove dependence on source ordering (Zaheer et al., 2017). Weighting relative vectors with invariant scalars gives equivariant updates (Satorras et al., 2021). This construction connects object interactions (Lei et al., 2025, Feng et al., 2025) and temporal differences (Yuan et al., 2025) to learned dynamics. We use it to propose displacements inside NTP, where the useful symmetries remain task-dependent (Vadgama et al., 2025).

Conservation means constancy along a trajectory. Signed planar angular momentum, for example, is conserved in our orbits yet reverses under reflection. We therefore test conservation supervision and geometry separately.

Testing what the model can do. Inductive-bias probes assess behavior after controlled adaptation (Vafa et al., 2025). We pair that perspective with autonomous forecasting. Forecasting tests generated motion, force adaptation tests learning from limited labels, and force transfer tests reuse on changed orbits.

3 Problem Setting

Both tasks initialize from observed history, then forecast without further observations or simulator access. Their models and objectives are separate. We ask: **RQ1:** How does learned mechanics affect autonomous visual forecasts? **RQ2:** Can physical supervision and pairwise geometry improve forecasting while retaining NTP? **RQ3:** Which combinations benefit from temporal supervision, and which gains persist under changed physical conditions?

Visual forecasting. Given two images x_{t-1}, x_t, the model must predict future physical states $y_{t+k} = (q_{t+k}, v_{t+k})$, where t indexes frames and k counts forecast steps. Here q is an angle or spring position and v its velocity. Encoder f_θ and predictor g_ϕ learn to predict next-image features.

An exponential-moving-average target encoder $\bar{f}_{\bar{\theta}}$ also uses the observed history $x_{\leq t}$. Subscripts $\theta, \phi, \bar{\theta}$ denote network parameters. Mask m selects patches in the future image:

$$\mathcal{L}_{\text{vid}} = \ell\big(g_\phi(f_\theta(x_{\leq t}), m), \text{sg}[\bar{f}_{\bar{\theta}}(x_{t+1} \mid x_{\leq t})_m]\big)\,, \tag{2}$$

where ℓ is Smooth-L1 loss on normalized features and sg stops gradients through the target. Both visual models also use numerical position and velocity labels to train the encoder, predictor, and state decoder. Physics-JEPA additionally trains a mechanical transition (Section 4.1). At inference, both models receive only the initial image pair.

Planetary forecasting. A separate transformer learns from sequences of N discretized coordinate rows. After a sampling-cadence header, each row z_i contains one body's coordinates. Rows enumerate bodies within a frame, then advance in time. The learned distribution p_θ predicts the next row from the preceding rows $z_{\leq i}$:

$$\mathcal{L}_{\text{NTP}} = -\sum_{i=1}^{N-1} \log p_\theta(z_{i+1} \mid z_{\leq i}). \tag{3}$$

Let $\mathcal{K} \subseteq \{F, H, L, \text{TD}\}$ index the force, orbital-energy, angular-momentum, and temporal objectives enabled in a planetary variant. With weights $\lambda_k \geq 0$, its training loss is

$$\mathcal{L}_{\text{planet}} = \mathcal{L}_{\text{NTP}} + \sum_{k \in \mathcal{K}} \lambda_k \mathcal{L}_k. \tag{4}$$

Physical labels are calculated from simulated states. The temporal objective uses coordinate displacements. All variants generate future coordinates through NTP, with auxiliary heads used only in training. Table 1 places these objectives alongside structure within the model's computation.

4 Structure within Predictive World Models

We first describe visual forecasting, then three additions to planetary NTP. Hats denote predicted quantities.

4.1 Visual prediction with learned mechanical transitions

Both visual models in Figure 1 start from the same initial image pair. Baseline image-JEPA repeatedly predicts visual features and decodes them into position and velocity. Physics-JEPA decodes the initial state once, then advances it with learned mechanics.

Both train from scratch using the visual feature loss in Equation 2 and mean squared errors on decoded positions and velocities, for both observed and predicted frames. The mechanical head learns a potential energy $V_\psi(q)$ and a dissipation function $R_\psi(q, v)$, where ψ denotes its parameters. Dots denote time derivatives and ∂_q, ∂_v derivatives with respect to position and velocity. For a single angle or spring coordinate, energy E_ψ uses fixed kinetic energy with unit inertia:

$$E_\psi(q, v) = \tfrac{1}{2}v^2 + V_\psi(q), \qquad \dot{q} = v, \qquad \dot{v} = -\partial_q V_\psi(q) - \partial_v R_\psi(q, v). \tag{5}$$

Table 1: Physical structure organized by its role in predictive computation. The remaining interaction functions and potential dynamics are learned.

Structure	Entry point	Supplied information	Capability tested
Physical supervision	Training objective	Force or conserved-quantity labels	Autonomous token forecasting
Temporal consistency	Representation objective	Same-body displacement or a predicted physical step	Temporal evolution
Relative and pairwise geometry	Input and residual branch	Relative vectors and invariant scalar features	Token rollout and frame behavior
Learned mechanics	Executed transition	Kinetic structure and state/acceleration supervision	Visual trajectories and phase space

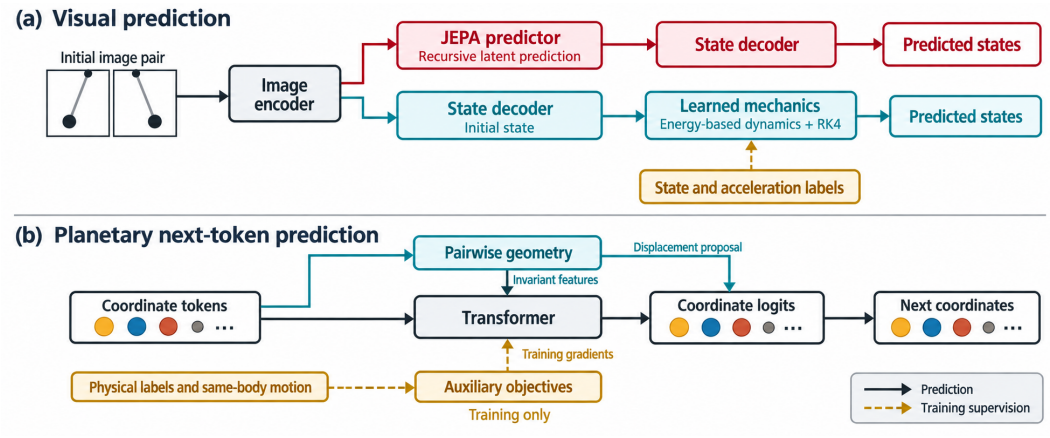

Figure 1: Visual (a) and planetary (b) forecasting. Baseline image-JEPA decodes predicted features. Physics-JEPA advances states through learned mechanics with fourth-order Runge–Kutta integration (RK4). Both visual models share initial images and state supervision. The planetary adapter changes embeddings and token scores, while auxiliary losses train the backbone. Solid arrows show prediction, dashed arrows supervision.

The potential's slope determines acceleration. Conservative systems set $R_\psi = 0$. Damped systems learn a positive damping coefficient $d_\psi(q)$ and use $R_\psi(q, v) = \frac{1}{2} d_\psi(q) v^2$, so that $\dot{E}_\psi = -d_\psi(q) v^2$. For the double pendulum, q contains both angles and v both angular velocities. The head uses their known coupled kinetic energy (Appendix A.4). RK4 integrates these equations over each frame interval. Physics-JEPA trains the head on simulated states, penalizing errors in acceleration and energy balance along the motion. These losses update only the mechanics head. Separate gradient clipping preserves the visual updates. A second, Lagrangian head receives acceleration supervision during training (Appendix A.4). The baseline uses only feature and state losses. Appendix A.1 tests an additional loss that couples mechanics to visual prediction.

4.2 Physical supervision while retaining next-token prediction

For planetary models, auxiliary readouts predict force F, orbital energy H, and planar angular momentum L_z from each body's hidden state. Their mean squared error (MSE) losses are

$$\mathcal{L}_F = \mathrm{MSE}(\hat{F}, F), \quad \mathcal{L}_H = \mathrm{MSE}(\hat{H}, H), \quad \mathcal{L}_L = \mathrm{MSE}(\hat{L}_z, L_z). \tag{6}$$

Let $r_i, r_\star$ be absolute planet and star positions, $m_i, m_\star$ their masses, and G the gravitational constant. Newtonian force is

$$F_i = -G m_i m_\star \frac{r_i - r_\star}{\|r_i - r_\star\|_2^3}. \tag{7}$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm. The cubic denominator normalizes the displacement vector and gives an inverse-square force magnitude. For star-relative position $r = r_i - r_\star$, relative velocity v, and reduced mass $\mu = m_i m_\star / (m_i + m_\star)$, the remaining targets are $H = \frac{1}{2}\mu\|v\|^2 - G m_i m_\star / \|r\|$ and $L_z = \mu(r_x v_y - r_y v_x)$. Labels match hidden states by body and time. Each enabled loss in Equation 4 trains both readout and backbone. NTP alone generates coordinates.

4.3 Relative coordinates and symmetry-aware interactions

Plain NTP predicts absolute coordinates. Relative NTP predicts changes in a body's position relative to its star. Training and generation both compute and accumulate these displacements in integer coordinate bins.

Pairwise NTP estimates a body's next displacement from where other bodies are and how they have recently moved. The adapter uses the latest two complete prefix frames and makes this proposal rotate with the system. For body i at frame t, $\tilde{r}_{i,t}$ is its position divided by 50 astronomical units (AU). Its frame displacement is $\delta r_{i,t} = \tilde{r}_{i,t} - \tilde{r}_{i,t-1}$, without division by timestep. For another body

j, $r_{ij} = \tilde{r}_{j,t} - \tilde{r}_{i,t}$ is their separation and $w_{ij} = \delta r_{j,t} - \delta r_{i,t}$ their relative motion. Their squared lengths and dot product form three scalar features u_{ij}. A shared network with parameters ψ maps these features to a vector message γ_ψ and scalar weights α_ψ, β_ψ. Summing over other bodies gives hidden message η_i and displacement proposal d_i:

$$u_{ij} = [\|r_{ij}\|^2, \|w_{ij}\|^2, r_{ij}^{\mathsf{T}} w_{ij}], \qquad \eta_i = \sum_{j \neq i} \gamma_\psi(u_{ij}), \tag{8}$$

$$d_i = \sum_{j \neq i} \alpha_\psi(u_{ij}) r_{ij} + \beta_\psi(u_{ij}) w_{ij}. \tag{9}$$

The adapter adds η_i to the input embedding and biases output token scores toward the proposed displacement $50 d_i$ AU. One learned gate controls both additions, starting at zero. Both learn through the NTP loss, retaining the full causal history. Distances and dot products are rotation- and reflection-invariant, making the continuous proposal d_i equivariant (Appendix A.3). Ordered attention and quantization remain in the full token model.

4.4 Same-body temporal learning

Neighboring token rows can describe different bodies. We instead pair the same body across successive frames. Our adaptation of TDV (Daithankar et al., 2026) uses coordinate displacement to predict how its representation changes, replacing the original image differences. Let $h_{b,t}$ be the transformer's hidden state for body b at frame t, and $\xi_{b,t}$ its quantized star-relative coordinates. A small network M_{TD} predicts a feature change from the current hidden state and scaled displacement $\delta\xi_{b,t}$:

$$\delta\xi_{b,t} = (\xi_{b,t+1} - \xi_{b,t})/c_\xi, \tag{10}$$

$$\mathcal{L}_{\mathrm{TD}} = \rho(h_{b,t} + M_{\mathrm{TD}}(h_{b,t}, \delta\xi_{b,t}), \mathrm{sg}[h_{b,t+1}])\,. \tag{11}$$

Here ρ is Smooth-L1 loss, averaged over valid same-body pairs and hidden dimensions. The scale $c_\xi = 7002$ is the maximum token ID. The target is the same transformer's next hidden state, with gradients stopped. Original TDV uses an exponential-moving-average teacher. The loss trains the temporal network and NTP backbone without force or mass labels. Future displacement is training-only, leaving NTP generation unchanged. Appendix A.5 gives the settings.

5 Experimental Design

Visual training and evaluation. Each visual system trains on 200 simulated episodes of 150 images (64×64 pixels), sampled every $\Delta t = 0.05$ s. Each training window contains two observed frames and the next image as target. The encoder uses four transformer blocks, eight heads, width 256, and 8×8 patches. We evaluate the final epoch: 50 for simple pendulum and spring, 60 for damped pendulum, and 120 for double pendulum. Simple and damped pendulums have three independently initialized training runs (random seeds 42–44), and spring and double pendulum one (seed 42).

Root mean squared error (RMSE) pools 400 future steps (20 s) over 25 held-out initial conditions, excluding initialization. Angular errors take the shortest difference around the circle. Energy RMSE uses the simulator's energy function. Acceleration RMSE compares velocity changes per unit time. We report means and sample standard deviations across training runs. Appendix A.6 defines drift metrics.

Planetary training and data. All variants share 100,000 single-planet trajectories and 15,625 AdamW updates at learning rate 6×10^{-4}. Each update combines gradients from four batches of eight trajectories (effective batch size 32). The transformer has 12 layers, 12 attention heads, width 768, and 999 body-coordinate rows of context. Coordinates use 7,000 bins over $[-50, 50]$ AU plus three special tokens. Relative and pairwise models share integer-bin encoding. Auxiliary weights are 0.1. Velocities and masses are saved in physical-state records alongside coordinates to compute targets and evaluation metrics. They are not input tokens. Validation NTP loss selects the checkpoint. Three independent training runs use random seeds 1337–1339. The adapter increases parameter count from 108.53M to 109.72M. Absolute TDV uses the earlier auxiliary coordinate

encoding, so its ranking against corrected relative TDV compares complete configurations. Relative-model contrasts hold encoding fixed.

Each system draws two independent masses uniformly between 0.0001 and 0.0101 solar masses, designating the heavier body as its star. Orbit size (semimajor axis) is uniform between 0.3 and 32 AU. Elongation (eccentricity) follows a beta distribution with shape parameters 0.867 and 3.03. The cadence token specifies regular sampling every 0.0189 or 0.486 years, shared by both bodies.

Planetary forecasting and physics. Each model receives 32 frames and generates 128 on 16 held-out systems, choosing the most probable valid coordinate token at each step. Position MSE averages squared errors over future planet coordinates, then systems, excluding the star. Angular-momentum drift measures change from the last observed state, divided by a fixed physical scale from the true initial orbit. Equation 28 defines this scale, including near-zero angular momentum. Finite differences estimate velocities from predicted positions. Table 3 aggregates training runs. Appendix C gives individual scores and paired uncertainty intervals.

Sparse force adaptation. Each backbone receives a fresh linear force head and 1,000 full-backbone updates using 25 force labels per orbit, 50 for validation, and 424 for testing. Fits use eight named-planet systems at both cadences, with identical label splits and seed 1337. Figure 3 shows Earth at the finer cadence, with each condition fitted separately. Figure 4(a,b) averages all 16 orbits and reuses the nominal fit under both parameter changes without refitting. Appendices C.5 and C.6 report all nine variants. Observed orbital positions are inputs. Force-vector NMSE divides mean squared vector error by mean squared true-force magnitude.

Controlled changes. Paired planetary test sets change one parameter, retaining other random draws. Eccentricities 0.45 and 0.75 lie within training support. Tests span 64 frames, with a separate 400-frame test set. Further tests vary stellar mass, body count, order, and coordinates. Multi-planet systems combine independent star-planet orbits. Visual tests vary initial conditions, rendering, lengths, and masses. Changed parameters are withheld from the head. Forecasts use pretrained checkpoints without adaptation. Appendix D gives all scenarios.

6 Results

6.1 Learned mechanics improves visual dynamics

Physics-JEPA improves all four systems over baseline image-JEPA (Table 2). Both are trained to predict one step ahead and evaluated over 400 steps from the same image pair. Simple-pendulum angle RMSE falls from 0.8814 to 0.2212 rad, and damped-pendulum RMSE from 1.1261 to 0.0446 rad, improving in every seed.

Table 2: Baseline image-JEPA versus Physics-JEPA, separately trained with matched seeds. Forecasts: 400 steps on 25 initial conditions. State RMSE uses radians (spring: meters). Shared cells give total-system energy RMSE in joules for the double pendulum's upper and lower links. Single pendulums report mean ± sample SD over three seeds, and † marks one seed.

System	Baseline RMSE	Physics RMSE	Baseline energy	Physics energy
Simple pendulum	0.8814 ± 0.1892	0.2212 ± 0.0492	3.6523 ± 0.5831	0.4106 ± 0.0760
Damped pendulum	1.1261 ± 0.0174	0.0446 ± 0.0005	5.0440 ± 0.2389	0.0987 ± 0.0004
Spring–mass	$3.1486^{\dagger}$	$0.6905^{\dagger}$	$5.7768^{\dagger}$	$3.9664^{\dagger}$
Double pendulum, upper θ_1	$0.5709^{\dagger}$	$0.1882^{\dagger}$	$5.1660^{\dagger}$	$0.5058^{\dagger}$
Double pendulum, lower θ_2	$1.1174^{\dagger}$	$0.2600^{\dagger}$		

In phase space, sustained oscillation traces a closed loop, while damping draws the trajectory inward as energy is lost. Figure 2 shows baseline image-JEPA losing these patterns. Learned mechanics more closely preserves the pendulum and spring's cyclic motion and the damped pendulum's inward spiral. Simple-pendulum acceleration RMSE falls from 33.9603 to 1.8284 rad/s^2, and energy-value RMSE improves across all systems (Table 2).

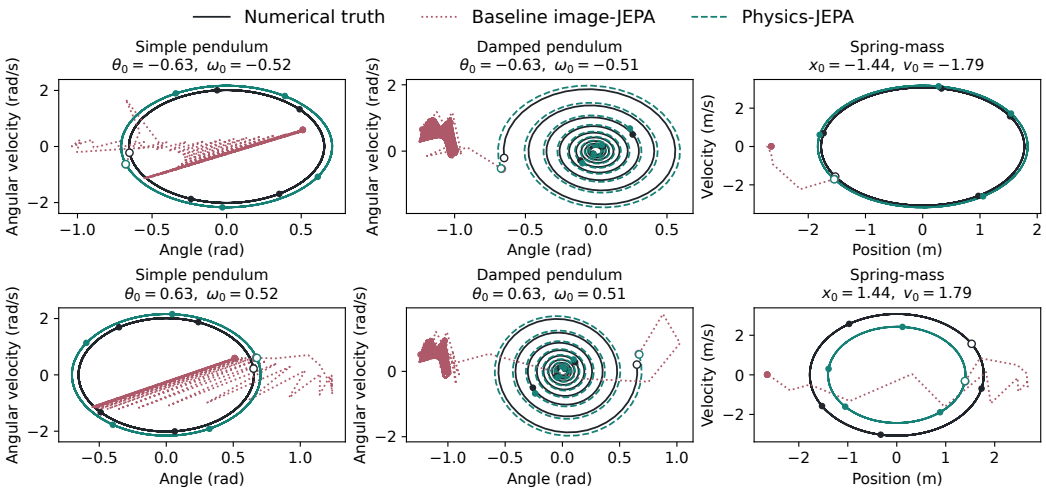

Figure 2: Autonomous forecasts from the same images. The separately trained baseline image-JEPA (red) predicts visual features, while Physics-JEPA (green) follows learned mechanics. Black curves show truth, open circles the first forecast, and dots 5 s intervals. Titles give the last observed states for two conditions and seed 42. Aggregate scores include all 25 conditions and 400 forecast steps.

Matched baseline and Physics-JEPA models produce identical latent forecasts. Physics-JEPA improves trajectories by using its supervised mechanics head to advance the decoded state. Latent agreement holds across eight nominal pairs and 116 parameter tests. The fresh-image comparison in Appendix B further shows why accurate state estimates need a reliable transition to sustain motion.

6.2 Physical supervision and geometry improve token forecasting

Can structure also improve the existing predictor? Planetary motion tests this question with NTP generating every future coordinate.

Angular-momentum supervision improves autonomous NTP. Plain NTP's position MSE falls from 1.0755 to 0.1437 AU^2, with lower angular-momentum drift (Table 3). Both improve in every matched seed. The auxiliary head is unused during generation, so the token predictor itself improves.

Table 3: Autonomous planetary forecasts without adaptation. Mean $\pm$ sample SD across three seeds on 16 systems, with 32 observed and 128 future frames. Both metrics are lower-is-better. Drift uses a fixed physical scale. Relative and pairwise models share the coordinate encoding.

Model	Position MSE (AU^2)	Angular-momentum drift
Plain NTP	1.0755 ± 1.6393	0.9196 ± 0.6293
Plain + angular momentum	0.1437 ± 0.0741	0.4662 ± 0.0929
Relative NTP	0.2034 ± 0.0864	0.4537 ± 0.0178
Relative + TDV	0.1395 ± 0.0542	0.7274 ± 0.0898
Pairwise NTP	0.0630 ± 0.0292	0.6078 ± 0.1268
Pairwise + TDV	0.1439 ± 0.0444	0.5279 ± 0.1198

Pairwise geometry improves relative NTP. Pairwise NTP lowers relative NTP's position MSE from 0.2034 to 0.0630 AU^2 and eight-frame displacement-vector normalized mean squared error (NMSE) from 0.1155 to 0.0301. NMSE scales error by the true vectors' mean squared magnitude. Both improve in every seed with the same coordinates and token encoding.

TDV depends on the representation. TDV improves relative NTP's position accuracy in every seed with encoding fixed. At the same weight, it increases pairwise NTP's position error while

reducing its angular-momentum drift (Table 3). Better conservation and better forecasting therefore need to be checked separately when combining temporal and geometric guidance.

Autonomous token forecasts under shifts. Relative + TDV reduces relative NTP's position MSE from 1.6329 to 0.7586 AU^2 at eccentricity 0.75 and from 16.6076 to 4.9973 on the separate 400-frame test set. Pairwise geometry is strongest on nominal forecasts, while relative + TDV leads these controlled changes, using pretrained models without adaptation.

6.3 Force learning and physical parameter transfer

We next test force learning from limited labels, adapting each backbone with the same budget. These tests use observed orbital positions. Correct force arrows should turn with the planet and point toward its star. Figure 3 compares equal-budget fitting within each condition. Figure 4(a,b) tests transfer of the nominal fit under parameter changes.

Temporal learning improves force adaptation in relative coordinates. Among the displayed models in Figure 3, relative + TDV has the highest direction agreement and lowest vector error in all three Earth conditions. Adding TDV to relative NTP increases direction cosine from 0.56 to 0.86, from 0.26 to 0.78, and from 0.40 to 0.79, respectively. The relative-coordinate control isolates this temporal benefit.

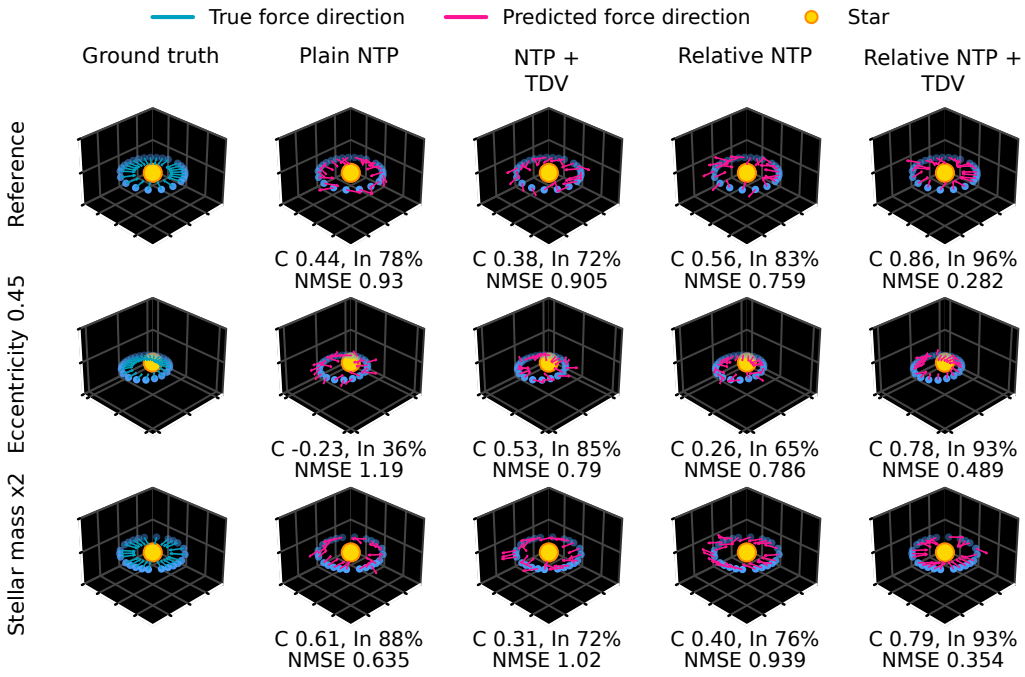

Figure 3: Force directions after equal-budget adaptation on Earth trajectories, using the orbital rendering of Figure 1 in Vafa et al. (2025). Each row is fitted separately from its pretrained checkpoint. Arrows show unit directions at 24 matched positions, while metrics use all 424 held-out labels. C is direction cosine and In is inward fraction, both higher-is-better. Physical vector NMSE is lower-is-better. Models use seed 1337. Appendix C.5 gives the full supervision comparison and selection protocol.

Nominal force learning and parameter transfer. Across all 16 orbit and cadence combinations, relative + TDV improves direction cosine over plain NTP from 0.337 to 0.544. Its directional advantage persists on changed trajectories without refitting: 0.308 to 0.508 at eccentricity 0.45 and 0.271 to 0.413 with doubled stellar mass (Figure 4a,b). It has the highest mean among all nine configurations, beating plain NTP in 13, 12, and 12 of the 16 cases.

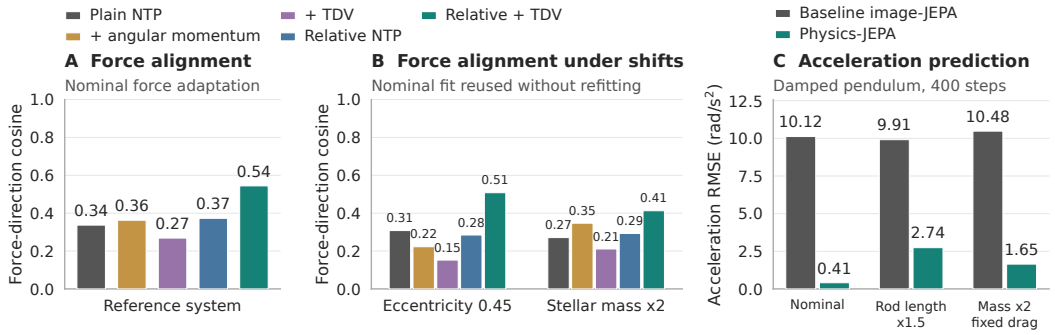

Figure 4: Force-direction cosine after nominal adaptation (a) and parameter changes without refitting (b), higher-is-better. Means cover eight planets and two cadences, with 424 held-out labels per orbit, seed 1337. (c) Damped-pendulum acceleration RMSE over 400 autonomous steps, 25 initial conditions and three training seeds, lower-is-better.

Visual acceleration under parameter shifts. For the damped pendulum, Physics-JEPA reduces acceleration RMSE from 10.12 to 0.41 rad/s^2 at nominal parameters, from 9.91 to 2.74 with a longer rod, and from 10.48 to 1.65 with doubled mass at fixed drag (Figure 4c). These forecasts use the trained head directly, without the changed parameters as inputs.

7 Discussion and Limitations

Match the structure to the prediction task. Visual mechanics generates the next state directly. Planetary supervision and geometry guide the token predictor that generates coordinates. Both require checking trajectories alongside physical scores. Energy can stay constant at the wrong amplitude, and a phase portrait can trace the right curve at the wrong speed. Joint trajectory and physical metrics reveal both errors.

Scope and next steps. Ordered attention and quantization limit end-to-end equivariance. Capacity-matched adapters and equally supervised unconstrained dynamics would further isolate geometric and mechanical choices. Body-count shifts and parameters ambiguous from two images remain difficult. Spring, double-pendulum, and force adaptation use one seed. Force direction is mass-independent at fixed position. Further tests must assess mass-scaling of force magnitude and integration error. Training and data efficiency require matched learning curves.

8 Conclusion

When observations stop, a predictive model must sustain motion using what it has learned. Our results show three ways that physical structure can help: supervising physical quantities, building geometry into prediction, and learning mechanical transitions. Across visual systems of increasing complexity, supervised energy-based transitions improve trajectories, acceleration, and energy accuracy. In planetary motion, angular-momentum supervision improves autonomous forecasting and momentum stability. Pairwise geometry improves the matched relative model across training seeds while retaining next-token prediction. Temporal consistency adds useful guidance, with benefits that depend on the representation and the task.

The phase-space plots make these improvements visible, while trajectory and physical metrics measure their extent. Together they show why the location of an inductive bias matters. A learned mechanical head improves visual rollout by advancing the decoded state. An auxiliary objective or geometric adapter improves planetary rollout by changing the token predictor itself. Controlled parameter and horizon tests show where these gains persist and which combinations remain useful. The practical lesson is to connect physical guidance to the computation that produces future states, then evaluate the motion it generates. This connects learning physical regularities with using them to make reliable predictions over time.

AI USE STATEMENT

Generative AI tools assisted with experimental design, implementation and debugging, analysis of results, literature search, figure preparation, and manuscript drafting and editing. Verification includes automated tests, checks of evaluation artifacts and source records, and visual inspection of figures. The authors are responsible for the final methods, interpretation, citations, and presentation.

REPRODUCIBILITY STATEMENT

Section 5 specifies the data, training budgets, evaluation horizons, and aggregation rules. The appendix provides objective definitions, physical metrics, complete comparison tables, and additional implementation details. Figures and tables are generated from saved predictions with recorded configurations and source hashes.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020. URL https://arxiv.org/abs/2003.04630.

Ninad Daithankar, Alexi Gladstone, Yann LeCun, and Heng Ji. You don't need strong assumptions: Visual representation learning via temporal differences. *arXiv preprint arXiv:2606.15956*, 2026. URL https://arxiv.org/abs/2606.15956.

Fan Feng, Phillip Lippe, and Sara Magliacane. Learning interactive world model for object-centric reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8187faaf6759ef6d4e93293339bc656e-Abstract-Conference.html.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/953dcf20cbd275465066ad63ea111f13-Abstract-Conference.html.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html.

Nima Hosseini Dashtbayaz, Hesam Salehipour, Adrian Butscher, and Nigel Morris. Physics-informed reduced order modeling of time-dependent PDEs via differentiable solvers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ae0f838dc2854f893ec9acbdca41dc59-Abstract-Conference.html.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. SPARTAN: A sparse transformer world model attending to what matters. In *Advances in Neural Information Processing Systems*, volume 38,

2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e2c33c1753e1f1924f5560b11fc4444f-Abstract-Conference.html.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Ziming Liu, Sophia Sanborn, Surya Ganguli, and Andreas Tolias. From Kepler to Newton: Inductive biases guide learned world models in transformers. *arXiv preprint arXiv:2602.06923*, 2026. URL https://arxiv.org/abs/2602.06923.

Melanie Mitchell and David C. Krakauer. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023. doi: 10.1073/pnas.2215907120. URL https://doi.org/10.1073/pnas.2215907120.

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. URL https://doi.org/10.1016/j.jcp.2018.10.045.

Pratyaksh Rao, Wancong Zhang, Randall Balestriero, Yann LeCun, and Giuseppe Loianno. Sky-JEPA: Learning long-horizon world models for zero-shot sim-to-real control of quadrotors. *arXiv preprint arXiv:2606.23444*, 2026. URL https://arxiv.org/abs/2606.23444.

Frank Röder, Jan Benad, Manfred Eppe, and Pradeep Banerjee. Dynamics-aligned latent imagination in contextual world models for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/696996164c52a52b8a162e62574fc9f9-Abstract-Conference.html.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. $E(n)$ equivariant graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Andrew Sosanya and Sam Greydanus. Dissipative hamiltonian neural networks: Learning dissipative and conservative dynamics separately. *arXiv preprint arXiv:2201.10085*, 2022. URL https://arxiv.org/abs/2201.10085.

Jayden Teoh, Manan Tomar, Kwangjun Ahn, Edward S. Hu, Tim Pearce, Pratyusha Sharma, Akshay Krishnamurthy, Riashat Islam, Alex Lamb, and John Langford. Next-latent prediction transformers learn compact world models. *arXiv preprint arXiv:2511.05963*, 2025. URL https://arxiv.org/abs/2511.05963.

Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracas, Christian A. Shewmake, Artem Moskalev, and Erik J. Bekkers. Probing equivariance and symmetry breaking in convolutional networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/afb4b81632998eed0357939db825aa0d-Abstract-Conference.html.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26941–26975, 2024. doi: 10.52202/079017-0846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60727–60747. PMLR, 2025. URL https://proceedings.mlr.press/v267/vafa25a.html.

Chaohao Yuan, Maoji Wen, Ercan E. Kuruoglu, Yang Liu, Jia Li, Tingyang Xu, Deli Zhao, Hong Cheng, and Yu Rong. Non-stationary equivariant graph neural networks for physical dynamics simulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e372bfe389adbf742a759396f0170356-Abstract-Conference.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/6931-deep-sets.

A Mathematical and Implementation Details

The appendix specifies the training objectives, physical metrics, and evaluation scenarios, followed by the complete model comparisons and seed-level measurements supporting the main text. Tables and figures are generated from saved numerical forecasts, with recorded configurations, checkpoint hashes, and checks for missing or nonfinite values. Aggregates include all evaluation systems and available training seeds for each comparison.

A.1 Gradient flow and visual consistency

The main text defines the predictive objectives. Here we detail how the visual models generate forecasts and how their losses update each component. Let h_t be encoded image features at frame t, and let P denote the autoregressive use of predictor g_ϕ. Decoder D maps two successive feature vectors to physical state $y = (q, v)$. The operator $\Phi^{\psi}_{\Delta t}$ is one fourth-order Runge–Kutta (RK4) step of the learned mechanical equations with parameters ψ over frame interval Δt. Hats mark predictions and k counts future steps:

$$\begin{aligned} \text{JEPA:} \quad & \hat{h}_{t+1} = P(\hat{h}_{t-1}, \hat{h}_t), \qquad \hat{y}_{t+1} = D(\hat{h}_t, \hat{h}_{t+1}), \\ \text{Mechanics:} \quad & \hat{y}_t = D(h_{t-1}, h_t), \qquad \hat{y}_{t+k} = \Phi^{\psi}_{\Delta t}(\hat{y}_{t+k-1}). \end{aligned}$$

The first predicted features use the observed image pair as their history. Each later step uses previous predictions.

Baseline image-JEPA (condition A) trains visual prediction and state decoding. Physics-JEPA (condition B) also trains a mechanical head. Its mechanical losses leave the visual branch unchanged, with separate gradient clipping preserving the visual updates. The consistency-coupled model (condition C) adds a stopped-gradient one-step mechanical target to visual prediction:

$$\mathcal{L}_{\text{cons}} = \left\| \hat{y}^{\text{latent}}_{t+1} - \text{sg}\left[\Phi^{\psi}_{\Delta t}(\text{sg}[\hat{y}_t])\right] \right\|_2^2. \tag{12}$$

Here $\hat{y}^{\text{latent}}_{t+1}$ is the state decoded from predicted features. This argument carries gradients to the visual predictor. The stopped-gradient target leaves the mechanics head's supervised update unchanged.

A.2 Conserved quantities and force supervision

For relative displacement r, velocity v, planet mass m, and stellar mass M, define reduced mass $\mu = mM/(m + M)$. The finite-mass two-body training targets are

$$H = \tfrac{1}{2}\mu\|v\|^2 - GmM/\|r\|, \tag{13}$$

$$L_z = \mu(r_x v_y - r_y v_x), \tag{14}$$

$$F = -GmMr/\|r\|^3. \tag{15}$$

The fixed-star limit gives $\mu \simeq m$ when $m \ll M$. Targets use aligned state records. A masked regression loss averages only available physical labels. Force-law supervision calculates the force target from physical state, while force auxiliary supervision reads the corresponding numerical target. These objectives supervise physical values during training. At generation, NTP uses its learned token distribution without an analytic conservation constraint.

A.3 EQUIVARIANCE OF THE GEOMETRIC BRANCH

Under $r'_i = Qr_i + a$, Q is a rotation or reflection ($Q^\mathsf{T}Q = I$) and a a translation. Relative vectors satisfy $r'_{ij} = Qr_{ij}$. Relative motion features transform in the same way. Their squared norms and inner products are invariant, so $u'_{ij} = u_{ij}$. The scalar coefficient networks therefore return the same values and

$$d'_i = \sum_{j \neq i} \alpha_\psi(u_{ij}) Qr_{ij} + \beta_\psi(u_{ij}) Qw_{ij} = Qd_i. \tag{16}$$

Summation is invariant to reordering source messages. This proof concerns the continuous proposal with matched source identities. The surrounding ordered transformer, feature injection, quantization, and autoregressive decoding are evaluated empirically as a complete system. The construction specifies transformation behavior without specifying an inverse-square exponent or a conserved energy function. The shared message MLP has one hidden layer of width 1536 with a GELU activation, producing one embedding-width message and two scalar coefficients per pair. For coordinate a, proposal bias at bin k is

$$B_{a,k}(d_i) = -\frac{1}{2}\left(\frac{c_k - 50 d_{i,a}}{64\,\Delta c}\right)^2, \qquad \Delta c = 100/7000 \text{ AU}, \tag{17}$$

where c_k is the signed displacement bin center. Special-token entries receive the coordinate-bias floor, and the proposal mask disables the bias when no complete frame is available. This bias and the hidden message share the learned scalar gate in the following interface. For row n predicting body i, e_n is the original input embedding, e'_n the updated embedding, and ℓ_n the output logits:

$$e'_n = e_n + g\eta_i, \qquad \ell_n = W\ \mathrm{GPT}(e'_{\leq n}) + gB(d_i). \tag{18}$$

Here GPT maps the causal embedding history $e'_{<n}$ to a hidden state, W is the output projection, and g is a learned scalar gate initialized to zero. The message η_i has the embedding dimension. The vector $B(d_i)$ contains the biases defined above for the displacement proposal d_i. Both connections learn through NTP.

A.4 LEARNED MECHANICS AND SUPERVISED DYNAMICS

Hamiltonian mechanics in canonical coordinates and Lagrangian mechanics in velocity coordinates satisfy the equations below. Here q is position, p canonical momentum, v velocity, H the Hamiltonian energy, and L the Lagrangian:

$$\dot{q} = \partial_p H, \qquad \dot{p} = -\partial_q H, \tag{19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\partial_v L - \partial_q L = 0. \tag{20}$$

The implementation uses velocity state (q, v). For one degree of freedom it uses unit-inertia kinetic energy and a learned potential. The double-pendulum head uses

$$T(q,v) = \tfrac{1}{2}(m_1 + m_2) l_1^2 v_1^2 + \tfrac{1}{2} m_2 l_2^2 v_2^2 \tag{21}$$

$$+ m_2 l_1 l_2 \cos(q_1 - q_2) v_1 v_2. \tag{22}$$

Here q_1, q_2 are upper- and lower-link angles, v_1, v_2 their angular velocities, m_1, m_2 their masses, and l_1, l_2 their lengths. With learned potential V_ψ and dissipation function R_ψ, the Euler–Lagrange equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}\partial_v[T(q,v) - V_\psi(q)] - \partial_q[T(q,v) - V_\psi(q)] = -\partial_v R_\psi(q,v). \tag{23}$$

Writing $L = T - V_\psi$, automatic differentiation constructs inertia matrix $M = \partial_v^2 L$ and mixed-derivative matrix $C = \partial_q \partial_v L$, then solves

$$M\dot{v} = \partial_q L - Cv - \partial_v R_\psi. \tag{24}$$

The coupled kinetic matrix and training-system parameters are supplied. The potential remains learned. Damping uses positive quadratic Rayleigh coefficients, ensuring nonpositive dissipative power. The heads receive numerical acceleration supervision and a power-balance objective:

$$\mathcal{L}_{\mathrm{dyn}} = \|f_\psi(y) - \dot{y}\|^2, \tag{25}$$

$$\mathcal{L}_{\mathrm{power}} = |\nabla_y E_\psi(y)^\mathsf{T} \dot{y} - \dot{E}_\psi^{\mathrm{expected}}|^2. \tag{26}$$

Here $f_\psi(y)$ is the learned state derivative, $\dot{y}$ the simulated derivative, and $\nabla_y E_\psi$ the energy gradient with respect to state. The conservative expected power is zero, and the damped expected power follows the Rayleigh term. A separate Lagrangian head receives acceleration supervision. Every reported assisted trajectory integrates the energy-based head with RK4 from an image-derived initial state.

A.5 Training settings

Visual training initial conditions are independent uniform draws. The simple pendulum uses $\theta_0 \in [-1.2, 1.2]$ rad and $\omega_0 \in [-1.6, 1.6]$ rad/s. The damped pendulum uses $[-1.5, 1.5]$ rad and $[-2, 2]$ rad/s. Both have nominal length 1 m and gravity 9.81 m/s^2, with damped angular coefficient 0.3 s^{-1}. The spring uses mass 1 kg, stiffness 3 N/m, $x_0 \in [-2.5, 2.5]$ m, and $v_0 \in [-3, 3]$ m/s. Each double-pendulum arm has length 1 m and mass 1 kg, with independent initial angles in $[-1.2, 1.2]$ rad and angular velocities in $[-1.5, 1.5]$ rad/s. Entire simulated episodes remain within their split when extracting overlapping three-frame windows.

Visual JEPA loss has weight 1. Position and velocity losses each have weight 5, or 10 per coordinate for the double pendulum. The energy-based and Lagrangian derivative losses each have weight 0.05 in B and C. Power loss has weight 0.05 for conservative systems and 0.3 for the damped pendulum. These terms ramp linearly from epochs 5 to 20, or 10 to 40 for the double pendulum. C adds consistency with weight 2 under the same ramp. The potential MLP has three width-256 tanh layers. The single-DOF head feeds position and a zero velocity into this network, ensuring potential energy depends on position alone. Context-encoder targets use EMA decay 0.999. Training predicts one future frame per window, while evaluation recursively generates the entire forecast without fresh observations. Visual optimization uses AdamW with learning rate 3×10^{-4}, cosine decay, and batch size 32.

Planetary auxiliary regressions attach to the currently observed body row and exclude star and metadata rows where physical labels are masked. The main temporal runs use first differences of quantized star-relative coordinates, omit star pairs, and retain the full causal attention context. All combinations use the common auxiliary weight 0.1. Planetary validation selects the best base-task checkpoint within the fixed training budget, using 10 validation batches every 250 updates. The adapter adds exactly 1,189,635 parameters. Visual evaluation uses the final epoch, consistently across A, B, and C.

A.6 Metrics and aggregation

Planetary position MSE excludes the star and averages over the forecast window, planet coordinates, and systems. For an angular prediction, visual RMSE uses $\operatorname{atan2}(\sin\delta, \cos\delta)$, with $\delta = \hat{q} - q$. Observed initialization frames are omitted. The main visual energy-value error compares physical energy evaluated on the predicted and numerical trajectories:

$$\mathrm{RMSE}_E = \sqrt{\frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T}[E_{\mathrm{phys}}(\hat{y}_{s,t}) - E_{\mathrm{phys}}(y_{s,t})]^2}. \tag{27}$$

Here S counts evaluation systems and T future steps for one trained model. State RMSE follows the same pooling order, using circular angular errors where appropriate. Training seeds are aggregated only afterward. Double-pendulum energy is computed jointly from both arms. Drift instead measures changes relative to the predicted cutoff energy, including accumulated dissipation for damped motion. Both are needed because a conserved but incorrect orbit can have low drift.

For planetary angular momentum, the evaluator forms relative velocities by finite differences and computes $\hat{L}_t = \mu(\hat{r}_{x,t}\hat{v}_{y,t} - \hat{r}_{y,t}\hat{v}_{x,t})$. The reported drift for one system is

$$D_L = \frac{1}{T}\sum_t \frac{|\hat{L}_t - \hat{L}_0|}{S_L}, \qquad S_L = \max\left(|L_0^{\mathrm{ref}}|, \mu\sqrt{G(m+M)\|r_0^{\mathrm{ref}}\|}, \epsilon\right). \tag{28}$$

The fixed scale uses the true initial context state. Here L_0^{ref} and r_0^{ref} are the true initial angular momentum and star-relative position, m, M the planet and stellar masses, $\mu = mM/(m+M)$ their reduced mass, and ϵ a positive numerical floor. Centered differences estimate interior velocities,

with second-order one-sided differences at the endpoints. Evaluation averages system drifts within each trained model before aggregating seeds. Low drift measures stability around the predicted initial value, so it is interpreted alongside trajectory error.

For star-to-planet vector r_i, true force F_i, and estimate $\hat{F}_i$, the supplementary force geometry diagnostics include

$$C_{\mathrm{dir}} = \frac{1}{n} \sum_i \frac{\hat{F}_i^{\mathsf{T}} F_i}{\|\hat{F}_i\| \|F_i\|}, \tag{29}$$

$$P_{\mathrm{in}} = \frac{1}{n} \sum_i \mathbb{1}[\hat{F}_i^{\mathsf{T}} r_i < 0]. \tag{30}$$

Direction and inward fraction are higher-is-better. Vector error is lower-is-better. Zero-vector and nonfinite cases are counted explicitly. Normalized acceleration error uses a zero-acceleration reference. Forecasting tables report variation across training seeds. The force-adaptation table reports variation across orbit and cadence cases for one pretrained seed. Captions identify the aggregation unit.

B Additional Visual Results

B.1 Physical metrics and consistency coupling

The nominal results use the same numerical truth across matched conditions. The main table compares baseline image-JEPA (A) with Physics-JEPA (B). Phase portraits use initial-condition grid indices 6 and 18 at seed 42, chosen before examining model performance. The table below also compares learned mechanics with latent autoregression within Physics-JEPA, from the same initial images. The consistency-coupled model (C) tests additional dynamics-to-visual consistency. It improves latent angle forecasting in one of three seeds for each single-pendulum system. Its structured forecasts remain close to B's.

Table 4: Acceleration RMSE and energy drift/balance RMSE for the same 400-step forecasts. Acceleration is in rad/s^2 (spring in m/s^2). Energy is in joules. The damped system includes dissipated energy. Double-pendulum acceleration pools both arms, and its energy drift is a whole-system metric. † indicates one seed.

System	Latent accel.	Structured accel.	Latent drift	Structured drift
Simple pendulum	33.9603 ± 24.5465	1.8284 ± 0.3823	3.4920 ± 0.5602	$0.0002 \pm 1.61 \times 10^{-5}$
Damped pendulum	10.1198 ± 5.8680	0.4098 ± 0.0040	3.8393 ± 0.4601	$0.0010 \pm 4.24 \times 10^{-5}$
Spring–mass	$7.3425^{\dagger}$	$2.0733^{\dagger}$	$4.9880^{\dagger}$	$9.81 \times 10^{-5\,\dagger}$
Double pendulum	$36.2834^{\dagger}$	$6.2625^{\dagger}$	$5.0567^{\dagger}$	$0.1434^{\dagger}$

Table 5: Complete nominal coupling comparison. A and B latent predictions are identical. Position RMSE units and seeds follow Table 2.

System	A latent	C latent	B structured	C structured
Simple pendulum	0.8814 ± 0.1892	0.9495 ± 0.1089	0.2212 ± 0.0492	0.2250 ± 0.0445
Damped pendulum	1.1261 ± 0.0174	1.0968 ± 0.2974	0.0446 ± 0.0005	0.0449 ± 0.0003
Spring–mass	$3.1486^{\dagger}$	$3.2259^{\dagger}$	$0.6905^{\dagger}$	$0.6879^{\dagger}$
Double pendulum, arm 1	$0.5709^{\dagger}$	$0.5621^{\dagger}$	$0.1882^{\dagger}$	$0.1782^{\dagger}$

B.2 Matched local and autonomous prediction

This test supplies true image pairs before each fresh-frame prediction, while autonomous prediction receives only the initial pair. Both paths score the same 160 targets in 25 systems. Mean fresh-frame versus autonomous latent RMSE is $0.0968/0.8812$ for simple pendulum, $0.0247/1.1571$ for

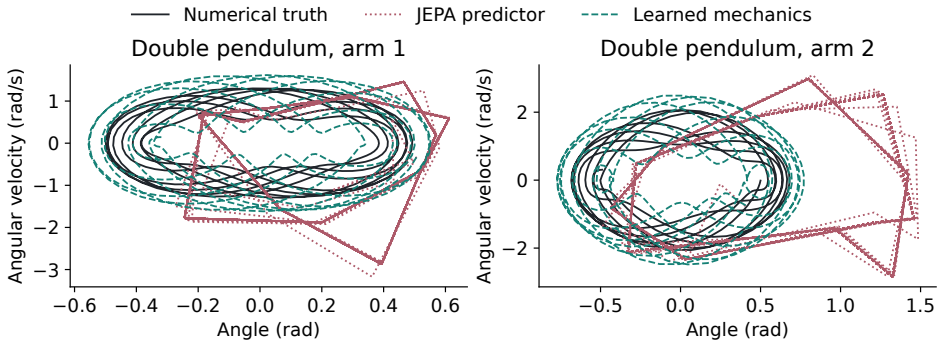

Figure 5: Both double-pendulum coordinates for held-out grid index 18, condition B, seed 42. The trajectories use the same initial image pair and 400 future targets. The nominal table aggregates all 25 initial conditions.

damped pendulum, $0.2044/3.0893$ for spring, and $0.0339/0.5657$ for double-pendulum arm one. Fresh observations correct state-estimation as well as transition error, so this test measures their combined effect.

C Additional Planetary Results

C.1 Autonomous trajectories and motion vectors

Figure 7 shows all training seeds and the growth of cumulative position error. Figure 6 uses the first held-out system and the middle listed training seed. No model-dependent choice of system, seed, or time window is made. The entire autonomous forecast is displayed, with unnormalized arrows $d_t = x_{t+8} - x_t$ at every sixteenth frame. No smoothing is applied.

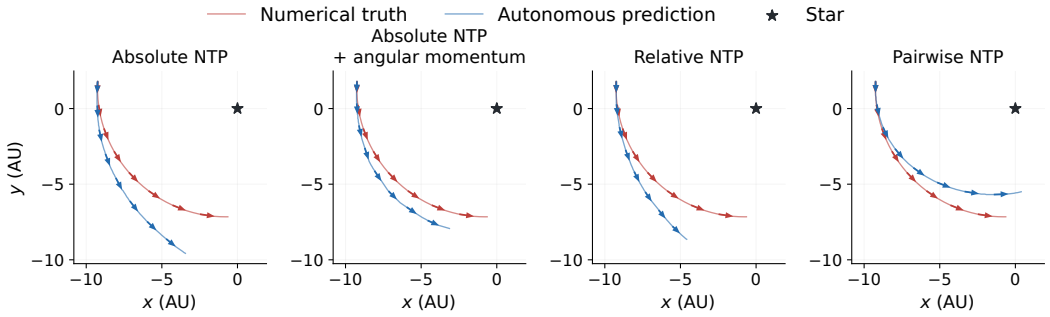

Figure 6: Predicted planetary motion after 32 observed frames. Red and blue curves show numerical truth and autonomous forecasts over 128 future frames. Arrows show eight-frame displacement in AU at every sixteenth frame, anchored on each trajectory with identical scales across panels. All panels use held-out system 0 and training seed 1338, without post-training adaptation. Table 3 and Table 6 aggregate all 16 systems and three seeds.

Table 6 evaluates all 120 overlapping eight-frame displacements in each of the 16 systems. For each seed,

$$E_{\text{motion}} = \frac{\sum_{s,b,t} \|\hat{d}_{s,b,t} - d_{s,b,t}\|^2}{\sum_{s,b,t} \|d_{s,b,t}\|^2}, \qquad d_{s,b,t} = x_{s,b,t+8} - x_{s,b,t}. \tag{31}$$

The star is excluded, and only future frames enter the calculation. The fixed eight-frame interval is the shortest horizon reported in the existing forecast evaluation. Cadence is identical across models for each system, though it differs between systems. These displacement vectors are measured in AU over eight frames. Direction cosine is computed only for nonzero predicted and reference

vectors, with coverage reported. Normalized vector error includes zero predictions and measures magnitude as well as direction. This diagnostic describes finite-interval motion and is distinct from the acceleration and frozen-readout tests below.

Table 6: Eight-frame displacement vectors from autonomous forecasts, all 16 identity systems, mean ± sample SD across three seeds. Lower vector NMSE and higher cosine are better. Cosine excludes zero-length vectors, with valid coverage shown. NMSE includes all vectors. These quantities describe displacement over eight frames.

Model	Vector NMSE	Direction cosine	Coverage
Absolute NTP	1.8469 ± 3.1033	0.9456 ± 0.0175	81.3%
Absolute + angular momentum	0.0648 ± 0.0250	0.9516 ± 0.0116	79.1%
Relative NTP	0.1155 ± 0.0599	0.9617 ± 0.0072	86.5%
Relative + TDV	0.1081 ± 0.0559	0.9633 ± 0.0020	90.6%
Pairwise NTP	0.0301 ± 0.0148	0.9673 ± 0.0044	89.7%
Pairwise + TDV	0.0897 ± 0.0321	0.9660 ± 0.0076	86.7%

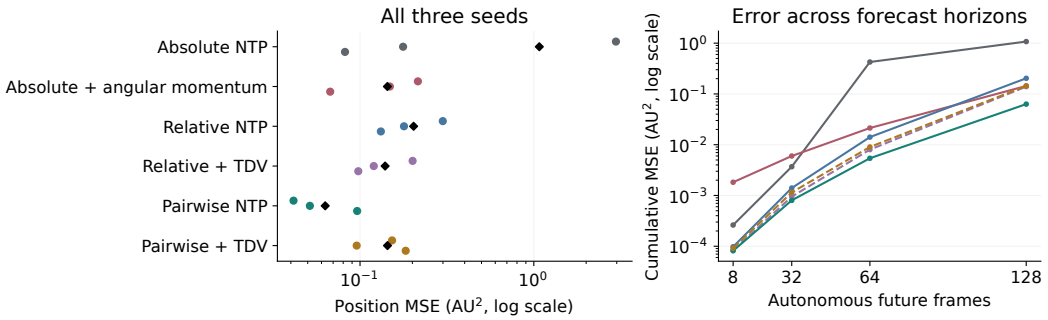

Figure 7: Autonomous planetary forecasting. Left: all three training seeds, with black diamonds indicating means. Right: cumulative error at the four evaluated horizons, averaged over seeds. Colors match across panels. The error axes are logarithmic. Angular-momentum supervision and pairwise geometry improve their respective matched baselines in every seed.

C.2 Complete objective comparison

Table 7 includes every current objective condition. The primary contrasts have three seeds. Additional force, energy, and pairwise-invariant combinations have one seed and supply exploratory ablations. Plain NTP conditions with TDV use the original auxiliary coordinate encoding. The relative and pairwise comparisons hold the revised encoding fixed.

C.3 Conservation and frozen readouts

Table 8 gives the matched-seed conservation evidence for angular-momentum supervision. All three seeds improve relative to absolute NTP, matching the direction of the position-error improvement.

Pairwise NTP's mean normalized energy and angular-momentum drift are 0.7696 and 0.6078. Adding TDV lowers these to 0.3880 and 0.5279, while increasing position MSE. Frozen linear acceleration readouts are trained on disjoint systems without backbone updates and evaluated on 100 held-out systems with 25 positions per system. Their normalized MSEs for relative NTP, relative + TDV, pairwise NTP, and pairwise + TDV are respectively 0.9866, 0.9935, 0.9876, and 0.9942. Direction cosines are 0.2194, 0.2124, 0.1919, and 0.2004. These results distinguish forecasting improvement from recovery of a reusable acceleration field.

Table 7: All planetary objective conditions on the original-distribution test set. Single-seed conditions (†) are exploratory. Plain NTP conditions containing TDV retain their original auxiliary coordinate encoding.

Model and supervision	Seeds	Position MSE (AU^2)
Plain NTP + angular momentum	3	0.1437 ± 0.0741
Plain NTP + angular momentum + TDV	3	0.1240 ± 0.0448
Plain NTP + force supervision	3	4.8811 ± 6.2799
Plain NTP + force supervision + TDV	3	3.1857 ± 4.6176
Plain NTP + force-law supervision	1	$0.4366^{\dagger}$
Plain NTP + force-law supervision + TDV	1	$0.1326^{\dagger}$
Plain NTP + energy supervision	1	$0.4465^{\dagger}$
Plain NTP + energy supervision + TDV	1	$0.0898^{\dagger}$
Plain NTP	3	1.0755 ± 1.6393
Plain NTP + TDV	3	0.3851 ± 0.3196
Pairwise NTP + angular momentum	1	$0.0568^{\dagger}$
Pairwise NTP + angular momentum + TDV	1	$0.0984^{\dagger}$
Pairwise NTP	3	0.0630 ± 0.0292
Pairwise NTP + TDV	3	0.1439 ± 0.0444
Relative NTP	3	0.2034 ± 0.0864
Relative NTP + TDV	3	0.1395 ± 0.0542

Table 8: Normalized angular-momentum drift on the same 16 identity systems and 128-step forecasts. Lower is better. Drift measures temporal variation from predicted initial angular momentum, normalized by a fixed physical scale from the true initial state. The final column reports the mean and sample SD over training seeds.

Model	1337	1338	1339	Mean $\pm$ SD
Absolute NTP	1.6417	0.4882	0.6290	0.9196 ± 0.6293
Absolute + angular momentum	0.4102	0.4148	0.5734	0.4662 ± 0.0929

C.4 PRETRAINED FORCE-HEAD CONTROL

Force-supervised models also predict forces directly from their pretrained heads, without further fitting. Each head receives the observed planet coordinates and causal history. A control on an Earth-mass orbit with eccentricity 0.45, a one-solar-mass star, and cadence 0.0189 years illustrates why direction scores require care. This case was selected after inspecting 48 planet, cadence, and parameter combinations for consistent improvements from TDV across three seeds. It contains 499 evaluated positions.

For the force-supervised TDV models, replacing predictions by each model's mean force on this test orbit gives direction cosine 0.3882, compared with 0.3729 for the original predictions. Magnitude calibration is also poor. Thus a fixed orientation can explain the directional score on this selected orbit. These pretrained heads are evaluated separately from the fresh heads and full-backbone adaptation used in the main force comparison.

C.5 MATCHED SPARSE FORCE ADAPTATION

Figure 3 compares four pretrained configurations after the same force-adaptation procedure. Each checkpoint receives a fresh linear force head, with the retained backbone trainable. Existing auxiliary heads are removed and their losses disabled. The pairwise variants retain their embedding branch, while the coordinate-logit proposal is inactive for force regression. AdamW uses learning rate 6×10^{-4}, batch size 8, gradient clipping 1, and 1,000 updates. All configurations use pretrained seed 1337 and the same head initialization and label splits.

Each fit contains eight named-planet two-body orbits at both 0.0189 and 0.486 years per frame. Of 499 available labels per orbit, 25 fit the model, 50 select the checkpoint, and 424 provide held-out evaluation. One scalar force normalization is computed from fitting labels only and applied to both coordinates. Metrics use physical vectors. For each orbit, vector NMSE is

$$E_F = \frac{\sum_{t \in \mathcal{H}} \|\hat{F}_t - F_t\|^2}{\sum_{t \in \mathcal{H}} \|F_t\|^2}, \tag{32}$$

where $\mathcal{H}$ is the held-out label set. The direction cosine and inward fraction follow Appendix A.6. The arrows alone show orientation, while vector NMSE also evaluates magnitude.

There are three fitting conditions: reference parameters, eccentricity 0.45, and doubled stellar mass. Each starts from the original pretrained checkpoint. The reference and doubled stars have one and two solar masses, outside the pretraining mass range. These tests measure force adaptation on observed orbits. Autonomous forecasting and fixed-model parameter transfer are separate evaluations. The two relative variants share the corrected integer-coordinate encoding. Absolute TDV retains its original auxiliary encoding, so comparisons across representations also include that implementation choice.

The main figure uses Earth at the finer cadence and shows all three conditions with relative-coordinate and temporal controls. It was chosen after examining the results. Twenty-four displayed positions are selected for spatial coverage from the shared held-out set, starting at its first point. Metrics always include all 424 held-out positions. Table 9 includes every pretraining variant on these same Earth cases. Table 10 checks the direction of improvement across all 48 planet, cadence, and condition cases. These related cases supply descriptive comparisons from one training seed.

C.6 AGGREGATE FORCE DIRECTIONS AND PARAMETER TRANSFER

Figure 4(a,b) uses the model fitted on reference parameters for every condition. This differs from the separately fitted rows of Figure 3. Each mean includes all eight planet identities and both cadences with equal weight. The 424 held-out labels per orbit are unchanged across models. The shifted conditions use no additional fitting or checkpoint selection. Table 11 reports all nine variants. The variation shown is across the 16 related orbit/cadence cases, conditional on pretraining seed 1337.

Relative + TDV improves direction cosine over plain NTP on 13 of 16 reference cases, 12 of 16 eccentricity cases, and 12 of 16 stellar-mass cases. The corresponding median paired improvements are 0.2464, 0.1417, and 0.1921. These tests support improved force directions after adaptation. Force magnitude rankings depend on the condition and aggregation. For example, pooled reference

Table 9: Complete Earth force-adaptation comparison, seed 1337 and cadence 0.0189 years. Each cell reports direction cosine / physical vector NMSE on 424 held-out labels. Higher cosine and lower NMSE are better. Every condition is fitted separately.

Pretraining	Reference	$e = 0.45$	$2M_\star$
Plain NTP	0.444 / 0.930	-0.234 / 1.192	0.614 / 0.635
NTP + force supervision	0.211 / 1.202	0.467 / 0.968	0.222 / 1.362
NTP + force law	-0.042 / 1.741	0.301 / 1.246	0.198 / 1.241
NTP + angular momentum	0.490 / 0.741	0.399 / 1.238	0.445 / 0.820
NTP + TDV	0.384 / 0.905	0.530 / 0.790	0.309 / 1.021
Relative NTP	0.560 / 0.759	0.263 / 0.786	0.400 / 0.939
Relative + TDV	0.858 / 0.282	0.776 / 0.489	0.786 / 0.354
Pairwise NTP	0.422 / 0.911	0.436 / 0.970	0.676 / 0.582
Pairwise + TDV	0.561 / 0.827	0.797 / 0.783	0.651 / 0.626

Table 10: Paired improvements over plain NTP across eight named planets, two cadences, and three separately fitted conditions. These are related diagnostic cases from one pretrained seed, not independent replications.

Pretraining	Direction wins	Vector-error wins
NTP + force supervision	21/48	23/48
NTP + force law	19/48	18/48
NTP + angular momentum	25/48	31/48
NTP + TDV	25/48	24/48
Relative NTP	33/48	23/48
Relative + TDV	34/48	36/48
Pairwise NTP	28/48	26/48
Pairwise + TDV	31/48	18/48

Table 11: Force-direction cosine after nominal force adaptation, higher-is-better. Values are mean $\pm$ sample SD across all 16 orbit/cadence cases, with 424 held-out labels per orbit and pretraining seed 1337. The same nominal-adapted model is evaluated in every column without refitting. The spread describes orbit variation, not independent training seeds.

Model	Reference	Eccentricity 0.45	Stellar mass $\times 2$
Plain NTP	0.3370 ± 0.3007	0.3081 ± 0.2393	0.2710 ± 0.2163
NTP + angular momentum	0.3627 ± 0.2732	0.2228 ± 0.3047	0.3470 ± 0.2170
NTP + force supervision	0.2802 ± 0.4269	0.1508 ± 0.4302	0.2153 ± 0.3765
NTP + force law	0.0938 ± 0.3962	0.1162 ± 0.4126	0.0633 ± 0.3850
NTP + TDV	0.2685 ± 0.3399	0.1524 ± 0.2876	0.2109 ± 0.3068
Relative NTP	0.3728 ± 0.4578	0.2846 ± 0.4595	0.2929 ± 0.4164
Relative + TDV	0.5441 ± 0.3112	0.5081 ± 0.2957	0.4128 ± 0.2990
Pairwise NTP	0.2462 ± 0.4159	0.3412 ± 0.2851	0.2352 ± 0.3074
Pairwise + TDV	0.2854 ± 0.4193	0.3702 ± 0.2717	0.2560 ± 0.3403

vector NMSE is 0.4465 for plain NTP and 0.6638 for relative + TDV, whereas their means of per-orbit NMSE are 112.25 and 19.55. The pooled score gives more weight to orbits with larger true forces.

C.7 Uncertainty across held-out systems

The main tables report training-seed variation. Table 12 separately measures variation across the 16 nominal evaluation systems. For each system, we average the paired treatment-minus-baseline position MSE across the three fixed training seeds. We then resample those 16 system-level differences with replacement 10,000 times and report percentile 95% bootstrap intervals for their mean. Negative differences favor the treatment. This preserves the pairing and the within-trajectory dependence across frames. The intervals are conditional on the trained checkpoints and do not replace replication over additional training seeds.

Table 12: Paired position-MSE differences in AU^2 on the original-distribution test set. Each system's difference is first averaged over the three fixed training seeds. Intervals use 10,000 paired bootstrap resamples of the 16 complete systems. Negative differences favor the first model. These intervals describe evaluation-system variation conditional on the trained checkpoints.

Contrast	Mean difference	95% interval
Angular momentum - absolute NTP	-0.9318	[-2.7508, 0.0248]
Pairwise - relative NTP	-0.1403	[-0.3914, -0.0027]
Relative + TDV - relative NTP	-0.0639	[-0.1836, 0.0006]
Pairwise + TDV - pairwise NTP	0.0808	[0.0058, 0.1946]

C.8 Coordinate encoding and sampling

Relative and pairwise models use integer-bin displacements and accumulation throughout training and generation. Quantization alone produces position MSE of approximately 0.000034 AU^2 on the original test set, below current model errors.

The training generator defines a parametric population by sampling orbital parameters independently for each system. Each training system contains one planet. All bodies in a system share a regular physical-time grid. The cadence varies across systems, so equal token counts need not cover equal fractions of an orbit. Instantaneous distance occupancy also differs from the semi-major-axis distribution, since eccentric orbits spend unequal time at different radii. Distance-error comparisons therefore require stratifying by cadence and forecast horizon. Controlled eccentricity interventions are reported separately from broader distribution changes.

D Complete Generalization Matrices

D.1 Seed-level controlled comparisons

Figure 8 shows seed-level forecasting errors for the controlled eccentricity and horizon tests, together with visual position errors under parameter changes. These supplement the force-direction and visual acceleration metrics in Figure 4. The visual comparison uses independently trained baseline and Physics-JEPA models with matched initial images and numerical truth.

D.2 Planetary interventions

The paired suite reuses random seeds and nuisance draws while intervening on mass, eccentricity, horizon, or body composition. Physical states and system identifiers can change under these interventions. Its reference test set differs from the original-distribution test set in the main table. Matched forecasts normally span 64 frames, with 400 for the long-horizon condition. The standard suite includes broader, compound shifts and coordinate changes. Both suites are reported below without pooling their means.

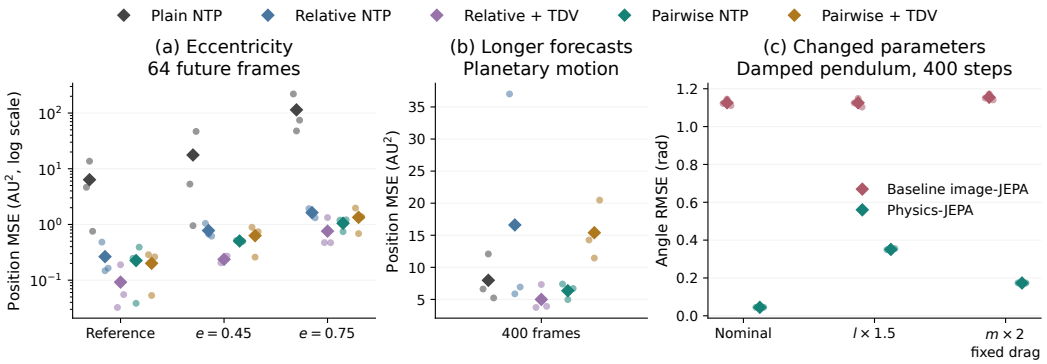

Figure 8: Seed-level forecasting errors. Circles show the three individual training seeds and diamonds show their arithmetic mean. Planetary panels measure position MSE, and the visual panel measures angle RMSE. Panel (a) uses a log scale. The 400-frame planetary test set is separate from the 64-frame eccentricity sets. Each visual score covers 400 future steps on 25 initial conditions.

Relative + TDV performs well on several controlled shifts. Larger body-count changes remain difficult. The matched four-planet case also changes mass distributions, and persistence scores 97.1951 AU^2, below the mean error of all four displayed models. These tests compose independent Kepler orbits without interplanetary forces.

Table 13: Complete matched scenario matrix for four models with matched coordinate encoding and three training seeds. Entries are mean autonomous planet-only MSE in AU^2. Each scenario is evaluated separately at its specified horizon.

Scenario	Relative	+ TDV	Pairwise	+ TDV
High eccentricity	1.6329	0.7586	1.0505	1.3436
Moderate eccentricity	0.7782	0.2363	0.5053	0.6298
400-frame horizon	16.6076	4.9973	6.3577	15.3980
Higher mass	0.4404	0.2772	0.2251	0.5579
Lower mass	0.0988	0.0356	0.0618	0.0812
Four planets	104.3827	103.6287	50222.0733	64140.1797
Four planets, reversed order	111.1045	101.8626	52268.0930	71078.8401
Matched reference	0.2648	0.0922	0.2253	0.2007

Table 14: Complete standard scenario matrix for four models with matched coordinate encoding and three training seeds. Entries are mean autonomous planet-only MSE in AU^2. Each scenario is evaluated separately at its specified horizon.

Scenario	Relative	+ TDV	Pairwise	+ TDV
Broad eccentricity extrapolation	67.4016	72.7884	66.9587	72.6492
Broad eccentricity interpolation	66.7530	69.5950	15219.3458	67.4005
New nominal cohort	119.4595	120.2306	99.9035	124.1985
Six planets	104.3705	106.7289	84825.8963	88123.6396
Historical-distribution reference	0.2034	0.1395	0.0630	0.1439
Higher stellar-mass extrapolation	119.7096	123.2674	12041.7800	16215.2600
Lower stellar-mass extrapolation	40.5245	49.6876	5007.6794	32.8092
Four planets	104.3827	103.6287	50222.0733	64140.1797
Reflection	1.3358	1.1612	1.7078	1.1384
Rotation, 45 degrees	0.2821	0.0622	0.0961	0.0832

Scenario	Relative	+ TDV	Pairwise	+ TDV
Shifted system distribution	110.8523	106.8372	63675.5283	72313.6830
Translation, (8, -6)	0.2034	0.1395	0.0630	0.1439

D.3 VISUAL INTERVENTIONS

Each visual table reports every condition-B scenario. All forecasts span 400 future steps and use the numerical truth associated with that scenario. Mass, gravity, and length values are not explicitly passed to the learned dynamics at evaluation. A visible-length shift changes the rendering as well as the dynamics, while a hidden-length shift preserves nominal visual length. Mass-null and matched scaling controls test evaluator consistency. They are not counted as independent demonstrations of new physical behavior.

The structured head retains useful advantages in several damped and spring conditions. Higher gravity for simple pendulum and a longer first rod for double pendulum illustrate cases where its nominal advantage does not persist. These tests show how particular physical changes affect the learned transition.

Table 15: Simple pendulum: condition-B 400-step angle RMSE in radians. Mean and sample SD across available seeds, with one seed marked †.

Scenario	Latent	Structured
Changed background	0.8802 ± 0.1873	0.2340 ± 0.0671
Changed bob color	0.8823 ± 0.1872	0.8999 ± 0.1009
Gravity scaled by 0.5	0.9206 ± 0.1782	0.9942 ± 0.0052
Gravity scaled by 1.5	0.8681 ± 0.1911	0.9087 ± 0.0002
Higher initial velocity	1.0800 ± 0.1540	0.9183 ± 0.0922
Larger initial angle	1.4310 ± 0.1045	1.3632 ± 0.0929
Hidden length scaled by 0.5	0.8608 ± 0.1925	0.9230 ± 0.0051
Hidden length scaled by 1.5	0.9002 ± 0.1835	0.9927 ± 0.0054
Visible length scaled by 0.5	0.8650 ± 0.1982	0.9360 ± 0.1636
Visible length scaled by 1.5	0.9042 ± 0.1858	0.9403 ± 0.0322
Half-mass null control	0.8814 ± 0.1892	0.2212 ± 0.0492
Double-mass null control	0.8814 ± 0.1892	0.2212 ± 0.0492
Nominal	0.8814 ± 0.1892	0.2212 ± 0.0492
Translated pivot	0.8850 ± 0.1872	0.9710 ± 0.1187
Reflected scene	0.8815 ± 0.1891	0.4203 ± 0.0993
Zoom control	0.8846 ± 0.1904	0.8145 ± 0.0915

Table 16: Damped pendulum: condition-B 400-step angle RMSE in radians. Mean and sample SD across available seeds, with one seed marked †.

Scenario	Latent	Structured
Changed background	1.1246 ± 0.0193	0.0450 ± 0.0004
Changed bob color	1.1322 ± 0.0122	0.1658 ± 0.0828
Gravity scaled by 0.5	1.1305 ± 0.0191	0.3701 ± 0.0001
Gravity scaled by 1.5	1.1245 ± 0.0172	0.3274 ± 0.0006
Higher initial velocity	1.1708 ± 0.0183	0.1934 ± 0.0099
Larger initial angle	1.2401 ± 0.0082	0.5350 ± 0.0237
Hidden length scaled by 0.5	1.1241 ± 0.0169	0.3516 ± 0.0005
Hidden length scaled by 1.5	1.1281 ± 0.0184	0.3354 ± 0.0002
Visible length scaled by 0.5	1.1220 ± 0.0212	0.3577 ± 0.0543
Visible length scaled by 1.5	1.1258 ± 0.0229	0.3507 ± 0.0037

Scenario	Latent	Structured
Half mass, fixed drag	1.1102 ± 0.0184	0.1105 ± 0.0005
Double mass, fixed drag	1.1547 ± 0.0161	0.1732 ± 0.0006
Nominal	1.1261 ± 0.0174	0.0446 ± 0.0005
Translated pivot	1.1319 ± 0.0122	0.4325 ± 0.2594
Reflected scene	1.1304 ± 0.0122	0.1048 ± 0.0371
Zoom control	1.1219 ± 0.0239	0.1256 ± 0.0694

Table 17: Spring–mass: condition-B 400-step position RMSE in meters. Mean and sample SD across available seeds, with one seed marked †.

Scenario	Latent	Structured
Changed background	$3.1494^{\dagger}$	$0.6923^{\dagger}$
Higher initial velocity	$4.2494^{\dagger}$	$1.9549^{\dagger}$
Higher mass	$3.3534^{\dagger}$	$2.7069^{\dagger}$
Lower mass	$3.0442^{\dagger}$	$2.1237^{\dagger}$
Mass/stiffness scaling control	$3.1486^{\dagger}$	$0.6905^{\dagger}$
Nominal	$3.1486^{\dagger}$	$0.6905^{\dagger}$
Higher stiffness	$3.0442^{\dagger}$	$2.1237^{\dagger}$
Lower stiffness	$3.3534^{\dagger}$	$2.7069^{\dagger}$

Table 18: Double pendulum, arm 1: condition-B 400-step angle RMSE in radians. Mean and sample SD across available seeds, with one seed marked †.

Scenario	Latent	Structured
Higher gravity	$0.5697^{\dagger}$	$0.6518^{\dagger}$
Lower gravity	$0.5691^{\dagger}$	$0.6310^{\dagger}$
Longer first rod	$0.5458^{\dagger}$	$0.6366^{\dagger}$
Shorter first rod	$0.6256^{\dagger}$	$0.6873^{\dagger}$
Longer second rod	$0.6035^{\dagger}$	$0.6081^{\dagger}$
Shorter second rod	$0.5451^{\dagger}$	$0.5840^{\dagger}$
Common-mass scaling control	$0.5709^{\dagger}$	$0.1882^{\dagger}$
Higher initial velocity	$0.7736^{\dagger}$	$0.6343^{\dagger}$
Larger initial angles	$1.0552^{\dagger}$	$1.0248^{\dagger}$
Higher mass ratio	$0.5904^{\dagger}$	$0.5051^{\dagger}$
Lower mass ratio	$0.5482^{\dagger}$	$0.5013^{\dagger}$
Nominal	$0.5709^{\dagger}$	$0.1882^{\dagger}$