PROBE AND CANONICALIZE: ACTIVE IDENTIFICATION OF HIDDEN PHYSICAL SCALE FOR OUT-OF-DISTRIBUTION WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical world models often fail to extrapolate when a physical factor leaves the training distribution, even when the governing interaction is unchanged. We study whether such a factor should be learned jointly with dynamics or identified separately and applied through known structure. Our *active hidden-physics canonicalization* predicts a factor-independent raw effect, estimates an unknown scale from a controlled visual probe, and applies it through a fixed analytical operator. In an object-centric environment where mass is absent from appearance, inverse mass is estimated from pixel displacement after a five-step probe calibrated only on training masses in $[1, 4]$. Across five backbone families and five seeds each, canonicalization reduces normalized final-position error by 58.8–60.9% on unseen masses in $[6, 10]$ and improves every evaluated seed. For the mirror-attention backbone, error falls from 1.670 ± 0.028 to 0.680 ± 0.004, matching a true-mass oracle at 0.679 ± 0.003. Crucially, supplying the same estimate as a learned input yields only 1.585 ± 0.086, and supplying oracle mass as input gives the same result. Fixed analytical application therefore reduces error by 57.1% relative to learned conditioning: access to the factor alone does not ensure extrapolation. Interventions further show robustness to estimation noise while revealing that this benchmark depends primarily on recovering the unseen physical scale rather than precise per-object attribution. These results isolate a specific OOD entanglement that is prevented by placing identified physical structure outside the learned predictor.

1 INTRODUCTION

A recurring failure mode of learned world models is systematic extrapolation: when a hidden physical factor governing an interaction — an object's mass, its friction coefficient, or its elasticity — takes a value never seen during training, learned dynamics may degrade even though the underlying interaction law has not changed. Learned simulators and object-centric models can capture rich in-distribution interactions (Battaglia et al., 2016; Watters et al., 2017; Sanchez-Gonzalez et al., 2020), but an end-to-end predictor must still learn both *how* an interaction unfolds and *how it scales* with the hidden factor. This can entangle a reusable computation with the numerical range observed during training.

We study a direct alternative: separate the two problems explicitly. Rather than asking a network to jointly learn the interaction and the hidden factor's effect, we let it predict a *factor-independent raw effect*, and apply the hidden factor's influence afterward, through a fixed, non-learned analytical operator whose form follows directly from the underlying physics (e.g., Newton's second law for mass). We call this mechanism *canonicalization*. The hidden factor itself is estimated by an *active probe*: a short, controlled physical interaction, observed only through rendered images, from which the factor is recovered by pixel-based measurement — never read from privileged simulator state.

We test this principle in a controlled object-centric environment where an object's mass is not encoded in its visual appearance, removing any shortcut between color and mass. Our contributions are:

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

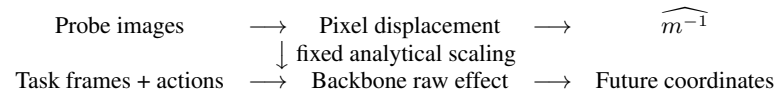


Figure 1: Canonicalization pipeline. Hidden mass is absent from appearance and from the backbone input. A separate visual probe estimates inverse mass, which enters prediction only through fixed multiplication.

1. We introduce active hidden-physics canonicalization and show that its fixed operator, rather than access to mass alone, is decisive: learned conditioning remains at 1.585 ± 0.086 OOD even with oracle mass, whereas explicit application reaches 0.680 ± 0.004.
2. We show the mechanism is *architecture-agnostic*: across five distinct backbone families, canonicalization yields a consistent 58.8–60.9% relative reduction in OOD error, with every seed of every backbone improved.
3. Through controlled interventions — object permutation, scene- and batch-level averaging, and injected estimation noise — we characterize *what* the mechanism actually exploits, finding that recovering the correct OOD physical *scale* matters far more than precise per-object attribution.

2 Related Work

World models and learned physical dynamics. Latent world models support prediction and planning by learning compact transition models (Watter et al., 2015; Hafner et al., 2019; 2020). Object- and relation-centric approaches instead model interacting entities explicitly (Battaglia et al., 2016; Chang et al., 2017; Watters et al., 2017), while graph-network simulators have scaled learned dynamics to complex particle systems (Sanchez-Gonzalez et al., 2020). Our question is orthogonal to backbone capacity: can a known scalar dependence be removed from the learned transition and reinstated analytically?

Physical-property identification. Prior work estimates latent physical properties from observed interactions, including mass and friction (Wu et al., 2015), and visual de-animation recovers physical scene parameters for forward simulation (Wu et al., 2017). Visual-physics benchmarks such as PHYRE and Physion test intuitive physical reasoning under richer visual variation (Bakhtin et al., 2019; Bear et al., 2021). Our controlled intervention is deliberately close to classical system identification (Ljung, 1999); the contribution is not a new estimator, but an intervention study of where the identified factor must enter a learned world model.

Physics-informed and adaptive learning. Physics-informed neural networks incorporate differential-equation residuals into training objectives (Raissi et al., 2019), while sparse system discovery and hybrid scientific models combine known structure with learned components (Brunton et al., 2016; Rackauckas et al., 2020). RMA similarly estimates hidden environment properties online to adapt a downstream policy (Kumar et al., 2021). Our narrower question is where a known algebraic dependence should be applied: the learned network predicts a canonical effect and a fixed operator applies the separately identified scale. The learned-input oracle in Section 4 directly tests whether unrestricted conditioning is sufficient.

3 Method

Figure 1 summarizes the information flow. The identification path observes only the controlled probe; the predictive path observes the task frames and actions; the two paths meet only at the analytical operator.

3.1 Canonicalization

Let an interaction produce a raw predicted effect $d \in \mathbb{R}^2$ (e.g., a displacement) that would be exact if the hidden physical factor m (e.g., mass) were equal to a canonical reference value. Under a

physical law of the form *effect* $\propto 1/m$ (as in $a = F/m$), we define the canonicalized effect as

$$d_{\text{final}} = d_{\text{raw}} \cdot \hat{m}^{-1}, \tag{1}$$

where $\hat{m}^{-1}$ is an estimate of the inverse hidden factor, produced by a component entirely separate from the network computing d_{raw}, and the multiplication is a fixed, non-learned operation. The network is never given m or $\hat{m}^{-1}$ as an input to be implicitly combined with other computation; it only ever sees the effect after canonicalization, and its only task is to produce d_{raw} correctly for a canonical, factor-independent regime.

3.2 Active visual probing

We estimate $\hat{m}^{-1}$ through a controlled, five-step physical probe: a known force is applied to the object, and its resulting displacement is observed through two rendered images (before and after the probe), never through privileged position or mass variables. The object's location in each image is measured by a pixel-intensity-weighted centroid, and the probe displacement is converted to $\hat{m}^{-1}$ through a single scalar linear mapping,

$$\hat{m}^{-1} = w \cdot \|\Delta_{\text{pixels}}\|, \tag{2}$$

calibrated by least-squares regression exclusively on training-range masses ($[1, 4]$); the fit is never exposed to out-of-distribution masses. This mapping is consistent with the underlying physics: for a fixed probing force applied over one step, displacement is proportional to $1/m$.

3.3 Object-centric world model

The base world model receives two rendered frames, the initial object coordinates, and an action sequence, and predicts per-object coordinates over a four-step horizon. Canonicalization is applied to the predicted displacement of the object whose mass is hidden. We evaluate the same interface across five interaction-processing families: a mirror-based block (patches route information through a small set of learned intermediaries), standard self-attention, a weight-tied attention variant, a GRU-based recurrent dynamics model, and an attention-based latent predictor. Hidden mass is never supplied to the standard models, and color/type is sampled independently of mass.

The design deliberately separates three information paths: appearance and actions determine the learned canonical effect; probe images determine $\hat{m}^{-1}$; and the only connection between them is fixed multiplication. Initial coordinates and object correspondence are provided rather than inferred; this controlled choice isolates extrapolation in dynamics from detection and tracking.

To test whether explicit application is necessary, we construct a parameter-matched learned-conditioning baseline. It linearly projects the complete estimated inverse-mass vector into the latent dimension and adds it to every patch embedding; the analytical operator is disabled. An oracle version receives the true inverse masses through the same learned path. All five conditions contain this projection, including standard and canonicalized variants where it receives zeros, giving exactly 995,936 parameters per model.

4 Experiments

4.1 Protocol

Models are trained exclusively on masses in $[1, 4]$ and evaluated, without retraining, on unseen masses in $[6, 10]$. The reported metric is final-position error divided by the error of a stationary-object predictor; lower is better, and a ratio of one means equality with that baseline. Five independent training seeds are used per condition. Conditions use matched scene generation, optimization budgets, and evaluation scenes. Initial positions and object correspondence are provided to all models (Section 5); future positions are used only as training targets and evaluation labels, never as model inputs. Standard deviations are sample standard deviations across training seeds, not confidence intervals over all possible scenes.

Table 1: Learned conditioning versus explicit analytical application, mirror-attention backbone, 5 seeds. Lower is better.

Condition	ID ratio	OOD ratio
Standard (non-factorized)	0.781 ± 0.004	1.670 ± 0.028
Learned input, estimated mass	0.779 ± 0.003	1.585 ± 0.086
Learned input, oracle mass	0.779 ± 0.003	1.585 ± 0.086
Fixed operator, estimated mass	0.780 ± 0.002	$\mathbf{0.680 \pm 0.004}$
Fixed operator, oracle mass	0.781 ± 0.003	$\mathbf{0.679 \pm 0.003}$

Table 2: Canonicalization across five backbone families, 5 seeds each.

Backbone	Standard OOD	Canonicalized OOD	Oracle OOD	Reduction
Mirror-attention	1.670 ± 0.028	0.680 ± 0.004	0.679 ± 0.003	59.3%
Self-attention	1.684 ± 0.012	0.693 ± 0.007	0.693 ± 0.007	58.8%
Weight-tied attention	1.756 ± 0.023	0.690 ± 0.004	0.690 ± 0.004	60.7%
GRU dynamics	1.749 ± 0.028	0.685 ± 0.003	0.684 ± 0.003	60.9%
Attention latent predictor	1.753 ± 0.021	0.690 ± 0.002	0.689 ± 0.002	60.7%

4.2 Main result

Table 1 compares standard prediction, learned conditioning with estimated or oracle mass, and analytical canonicalization with estimated or oracle mass. All conditions use identical scenes, parameter counts, and optimization budgets; no OOD sample is used for probe calibration or model selection.

ID performance is nearly identical across conditions. Learned conditioning gives only a 5.1% reduction over the standard model and remains worse than the stationary baseline. Its estimated- and oracle-mass results are identical, ruling out estimator error as the explanation. Applying the same estimate through the fixed operator instead reduces error by 57.1% relative to learned conditioning and reaches within 0.001 of the analytical oracle. Thus, knowing the factor is insufficient when its use must itself extrapolate from the ID range.

4.3 Generalization across backbones

Table 2 repeats the same comparison across five backbone families. The relative OOD error reduction from canonicalization ranges from 58.8% to 60.9%, and every one of the 25 (backbone $\times$ seed) runs improves under canonicalization.

4.4 Estimator quality and probe length

The visual estimator reaches an OOD mass MAE of approximately 0.066 (ID: 0.007) and an OOD mass correlation of 0.997. Table 3 shows that probe length matters: increasing from one to five steps substantially improves OOD identification, indicating that the five-step probe is a deliberate design choice, not an arbitrary one.

4.5 What does the mechanism actually exploit?

We run controlled interventions on the estimated mass to isolate what drives the OOD gain (Table 4). Permuting mass estimates across objects within a scene, or across scenes entirely, causes only mild degradation despite destroying the estimate-to-object correspondence (correlation with true mass drops to near zero). Injecting substantial noise (10–25%) barely affects performance. In sharp contrast, replacing the estimate with an in-distribution-calibrated constant collapses performance to near-baseline failure.

Table 3: Sensitivity to probe length.

Probe steps	MAE (ID)	MAE (OOD)	Correlation (OOD)
1	0.0356	0.3466	0.926
3	0.0116	0.1042	0.993
5	0.0072	0.0648	0.997

Table 4: Interventions on the estimated mass, mirror-attention backbone, OOD.

Intervention	OOD ratio	Correlation with true mass
Correct per-object association	0.680 ± 0.003	$+0.998$
Object permutation	0.685 ± 0.003	$+0.004$
Scene permutation	0.692 ± 0.004	-0.013
Inverted attribution	0.691 ± 0.003	-0.692
In-distribution constant	1.365 ± 0.051	0.000
10% noise	0.680 ± 0.003	$+0.822$
25% noise	0.691 ± 0.004	$+0.498$
Oracle	0.680 ± 0.003	$+1.000$

This pattern is consistent with a complementary ablation replacing per-object estimates with scene- or batch-level averages (Table 5): all such coarser estimates perform within 0.005 of the per-object oracle, while an in-distribution constant fails outright.

We conclude that, in this benchmark, the extrapolation gain is driven primarily by recovering the correct *unseen physical scale* rather than fine-grained per-object attribution. We view this as a strength for robustness and a scope limitation to be addressed by future benchmarks with greater intra-scene heterogeneity (Section 5).

Tables 4 and 5 use separate, matched evaluation draws and should be compared within, not across, tables. In both experiments the qualitative conclusion is the same: the OOD-scale intervention is necessary, while per-object precision is not the limiting factor in this environment.

5 Discussion and Limitations

Synthetic, controlled environment. Our environment is deliberately simple: rigid two-dimensional objects, a single hidden factor, and known initial positions and object identity. This isolates the causal-factorization question cleanly but does not yet constitute an end-to-end pixels-to-control system; extending the probe and identity correspondence to be inferred purely from pixels is left to future work.

Scale versus attribution. As shown in Section 4, our results primarily demonstrate identification of the correct OOD physical scale rather than precise per-object mass attribution, since coarser (scene- or batch-level) estimates perform nearly as well as per-object ones. A benchmark with greater intra-scene mass heterogeneity would be needed to stress-test attribution specifically.

Learned access is not explicit use. The oracle learned-input condition is central to interpreting our result: it receives the exact hidden factor yet fails OOD almost as severely as the standard model. This does not prove that every possible conditioning architecture must fail, but it shows that equal information, capacity, training scenes, and supervision do not reproduce the benefit. The fixed operator changes the extrapolation burden rather than merely adding information.

Architectural structure is not the source of the gain. A branch ablation (Appendix A.2) finds no robust OOD advantage from two-branch processing. The primary mirror implementation also contains an optional geometric delta-cache. Disabling it improves standard OOD error from 1.670 to 1.575 and canonicalized error from 0.680 to 0.668; canonicalization therefore reduces error by

Table 5: Granularity of mass identification, OOD.

Condition	OOD ratio	Inverse-mass MAE
Correct, per object	0.661 ± 0.002	0.0010
Scene-level average (probed)	0.663 ± 0.001	0.0154
One object for entire scene	0.662 ± 0.002	0.0198
Batch-level average (probed)	0.665 ± 0.001	0.0165
In-distribution constant	1.339 ± 0.049	0.2722
Oracle, per object	0.661 ± 0.002	0.0000

57.6% without the cache and 59.3% with it. We treat both mechanisms as implementation details, not sources of the reported extrapolation gain.

External transfer is unresolved. After fixing the protocol, we ran an exploratory, single-seed diagnostic on the Mass-Collision subset of Physion++ using reference instance segmentation. A mass estimator trained only on light objects did not extrapolate reliably to heavy objects, and estimated-mass canonicalization failed to beat a stationary baseline (Appendix A.4). Because this pilot changes the sensing, dynamics, dimensionality, and intervention assumptions simultaneously, we do not count it as evidence for or against the controlled benchmark result. It does show that the present method is not a plug-and-play video predictor: transfer requires both a valid identification intervention and an operator matched to the physical interaction.

Excluded development claims. During development we identified and corrected two data-leakage bugs (Appendix A.3) and did not reproduce several informal exploratory estimates under a matched, leak-free protocol. Those estimates are excluded from all claims and tables.

6 Conclusion

Explicitly separating identification of a hidden physical factor from factor-independent dynamics substantially improves extrapolation in our controlled visual benchmark. The decisive result is not simply that the probe recovers mass: even oracle mass supplied through a learned input fails OOD, whereas fixed analytical application matches the oracle operator. The effect is consistent across five backbone families and persists without branch specialization or geometric caching. It supports delegating known algebraic structure to explicit computation when the law is known, while leaving robust tracking, uncontrolled identification, richer interaction laws, and downstream control as open problems.

AI assistance disclosure

Generative AI tools, including OpenAI and Anthropic language models, were used for research ideation, literature discovery, experimental design, drafting and editing prose, and code generation. A substantial portion of the experimental and diagnostic code was AI-generated from author specifications, but the codebase was not generated in its entirety: the author developed the core research idea, physical factorization, environment logic, and several implementation components. AI-assisted code also helped generate deterministic synthetic data by implementing simulator and experiment scripts; no generative model produced the reported observations or labels directly. The author executed every reported experiment, inspected outputs, selected the final protocol, and takes responsibility for all claims.

AI-generated code was not accepted without verification. During manual line-by-line review, the author identified multiple leakage or evaluation defects, corrected them manually, discarded affected results, and reran the relevant experiments. Subsequent matched ablations also overturned favorable development interpretations and exposed a failed external-transfer pilot; the manuscript reports the validated evidence rather than the earlier estimates. Language models assisted with mathematical exposition but were not used to prove a formal theorem. This disclosure is intended to describe both the extent of AI assistance and the human verification that determined the final evidence.

Reproducibility statement

Main, learned-conditioning, delta-cache, and multi-backbone results use five independent training seeds; the branch ablation uses three seeds and the external diagnostic uses one, as labeled. Raw per-seed logs are retained. The probe, fixed operator, data ranges, metrics, and principal hyperparameters are specified in Sections 3–4 and Appendix A.1.

References

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, 2019.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix Binder, Hsiao-Yu Fish Tung, Pramod RT, Cindy Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems*, 2021.

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*, 2016.

Michael B. Chang, Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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.

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*, 2020.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems*, 2015.

Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems*, 2017.

Jiajun Wu, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Joshua B. Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, 2015.

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*, 113(15):3932–3937, 2016.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, second edition, 1999.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.

Jiajun Wu, Erika Lu, Pushmeet Kohli, William T. Freeman, and Joshua B. Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*, 2017.

A Appendix

A.1 Full experimental protocol

Training uses 600 steps with batch size 32. The main evaluation uses 256 episodes per seed. The probe uses 5 steps with before/after rendered frames and pixel-centroid displacement measurement.

The learned-conditioning ablation uses the same mirror-attention model class and 995,936 parameters in every condition. A linear map projects the N inverse-mass values to the latent dimension and adds the result to every patch embedding. Standard and fixed-operator conditions retain this module but feed it zeros; oracle conditions alone access true inverse mass. Training and OOD evaluation scenes are regenerated from identical seeds across paired conditions.

A.2 Two-branch architecture ablation

Table 6 compares a single-branch architecture to two two-branch variants (separate action handling for the actor only, versus shared action encoding), with and without canonicalization, 3 seeds each, parameter-matched to within 2.64%.

Table 6: Branch structure ablation, without canonicalization (pixel error) and with canonicalization (OOD ratio), 3 seeds.

Architecture	OOD error, no canon. (px)	OOD ratio, with canon.
Single branch	10.681 ± 0.041	0.689 ± 0.009
Two branches (actor-only action)	10.839 ± 0.164	0.690 ± 0.013
Two branches (shared action)	10.713 ± 0.168	0.686 ± 0.009

A.3 Data leakage bugs identified and corrected

During development, two data-leakage issues were identified through independent code review and corrected prior to producing any result reported in this paper: (1) a velocity variable that was copied from ground truth but never reassigned from an estimated value before being used by the planner, and (2) a mass-estimation fallback path that read a true position variable instead of the corresponding estimated one. Both were caught by manual line-by-line review, not automated tooling, and all affected results were discarded and recomputed.

A.4 Exploratory Physion++ stress test

We ran a post-hoc, single-seed diagnostic on Physion++ Mass-Collision using reference instance segmentation and no future frames as input. The estimator and predictor were trained on $m \leq 2$ and evaluated on $m \geq 4$; identification used only pre-prediction trajectories. Unlike our benchmark, this setting has no standardized probe, is three-dimensional, and contains only 12 readout and 22 test OOD scenes.

Mass estimation extrapolated poorly (MAE 11.82, Pearson $r = 0.208$). Test trajectory errors for stationary, standard, estimated-mass, and oracle predictors were respectively 2.348, 5.963, 5.949, and 3.124 pixels. The oracle reduced standard-model error by 47.6% but did not beat stationarity because heavy targets barely move. We exclude this pilot from the main evidence and leave uncontrolled identification to future work.