

Dynamic LeJEP: Maximum Entropy Representations for Sequential Prediction and Latent Planning with Theoretical Guarantees

Mohsen Mostafa

`mohsen.mostafa.ai@outlook.com`

Future Academy <https://orcid.org/0009-0004-4478-0317>

Research Article

Keywords: Joint-Embedding Predictive Architecture, world models, maximum entropy, physics-informed learning, representation learning, robotics, autonomous driving, latent planning, model-based control, model exploitation, variational encoder

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11405567/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Dynamic LeJEPa: Maximum Entropy Representations for Sequential Prediction and Latent Planning with Theoretical Guarantees

Mohsen Mostafa

Abstract—Joint-Embedding Predictive Architectures (JEPAs) are emerging as the backbone for latent world models in robotics and autonomous driving, yet injecting domain knowledge (physics, kinematics, geometry) into these models consistently degrades performance—with no theoretical explanation. We resolve this paradox by proving, through six theorems, that the isotropic maximum-entropy embedding is symmetry-stable under sequential prediction losses—no first-order incentive exists to rotate away from it under any Lipschitz, Euclidean-invariant prediction objective—and by validating across a phased empirical protocol that no non-Gaussian alternative outperforms it in practice. This yields a world-model design principle with a provable guarantee: *encoder maximum-entropy, predictor dynamics, decoder physics*—and explains why prior physics-informed JEPAs failed: they constrained the encoder, where constraints provably destroy the entropy guarantee, instead of the decoder, where they are provably benign. We validate this principle across two domains and two encoder mechanisms through a phased protocol covering measurement artifacts (Proposition 1: the $N/K \geq 5$ reliability threshold), powered statistical testing (110 seeded ablation runs across autonomous driving and bimanual robotic manipulation, under both a SIGReg-regularized encoder and a marginal-only VAE-KL encoder—the encoder family of classical world models—with paired Wilcoxon significance and joint Holm correction in every domain×mechanism combination), and a budget replication demonstrating that the encoder-physics violation *deepens* with training. Phase 3 additionally characterizes an entropy–utility frontier on low-intrinsic-dimension robotics data, discovering that per-dimension variance matching alone produces *correlated collapse*—a failure mode invisible on high-dimensional data—and that optimization budget, not regularization weight, is the binding entropy constraint. We further show this failure *generalizes*: a marginal-only VAE-KL encoder collapses identically on the same data, and we introduce a posterior signal-fraction diagnostic revealing that such collapse can hide beneath the encoder’s own sampling-noise floor, invisible to marginal diagnostics. To test whether the placement rule matters beyond representation quality, we further probe its consequence for

downstream control: a latent cross-entropy-method (CEM) planner built on the trained Phase 3 models consistently outperforms real-action replay on model-internal cost across ablation conditions and both encoder mechanisms—an impossible result for genuine planning, since replaying the true actions is itself the ground-truth solution. A model-exploitation diagnostic traces this to the planner discovering action sequences roughly a quarter of the action range away from the true trajectory while still lowering the learned decoder’s cost, uniformly regardless of physics placement and encoder choice. We report this transparently as a boundary condition on the practical claim: offline latent planning against a learned decoder is not, by itself, sufficient evidence that a representation supports downstream control, and we identify closed-loop, simulator-verified execution as the necessary next test. The result is the first formal, cross-domain-validated blueprint for physics-informed world models: place physics where the theorem says it is safe, never where intuition suggests—together with a concrete, reproducible cautionary result on how easily offline latent planning can be mistaken for evidence of control competence.

Index Terms—Joint-Embedding Predictive Architecture, world models, maximum entropy, physics-informed learning, representation learning, robotics, autonomous driving, latent planning, model-based control, model exploitation, variational encoder.

Code, Data, and Demo Availability. The paper repository, including this manuscript’s source, is available at <https://github.com/Dynamic-LeJEPa/Dynamic-LeJEPa>. Training and evaluation code, trained checkpoints, and ablation result files (.pkl/.json) for all three phases—including the VAE-KL comparison arms (ablation_results_vae.*, Phase 2; phase3_ablation_results_vae.*, Phase 3) and the KL-weight calibration sweeps—are indexed at <https://github.com/Dynamic-LeJEPa/Dynamic-LeJEPa/tree/main/Links%20For%20Codes%20%26%20Data%20>. An interactive Gradio demo for exploring the trained models and ablation results is hosted at <https://huggingface.co/spaces/MohsenAI5/Dynamic-LeJEPa>.

All experiments in this paper were run on a single commodity GPU (NVIDIA T4, 16 GB), and we release this resource specifically to let readers verify the model-exploitation confound of Sec. VII-E9 independently, since that result depends on details (goal construction, CEM hyperparameters, replay/random baselines) that are easiest to check by running the code directly rather than from the paper text alone.

Licensing. Training, evaluation, and analysis code in the paper repository is released under the **MIT License**. The manuscript text and figures are released under **CC BY 4.0**. Trained checkpoints and derived result files inherit the license of the upstream training data: checkpoints and ablation artifacts trained on nuScenes are distributed under **CC BY-NC-SA 4.0**, consistent with the nuScenes dataset license [12]; checkpoints and artifacts trained on MimicGen are distributed under the license of the upstream MimicGen release [11]. Each Kaggle-hosted dataset used in this work specifies its license individually in its dataset metadata. We distinguish these explicitly because the code license does not, by itself, grant rights to the non-commercial, share-alike upstream data it was trained on.

I. INTRODUCTION

In applied computer vision—particularly in data-starved domains like computational pathology, novel geometric mapping, and physical robotics—researchers face a fundamental crisis. When confronted with limited, new real-world sequential data, current deep learning frameworks act as “black boxes.” Practitioners are forced to rely on massive transfer-learning crutches (e.g., DINOv2) or spend endless epochs tuning heuristic stop-gradients, EMA schedules, and loss weights, never knowing if the internal representations have collapsed or if the true geometry of the data is actually understood.

The natural intuitive response to limited physical data is to inject domain knowledge—forcing physics constraints, geometric rules, or strong inductive biases into the latent space to help the model learn. However, empirically, these “Physics-Informed” approaches consistently fail or degrade representation quality. Why does helping the model actually hurt the model?

Joint-Embedding Predictive Architectures (JEPAs) offer a promising blueprint, predicting in latent space rather than pixel space. Recently, LeJEPA [1] provided the first theoretical foundation for JEPAs, proving that isotropic Gaussian embeddings minimize worst-case downstream risk (Theorem 1 of [1]). This result, combined with Sketched Isotropic Gaussian Reg-

ularization (SIGReg), enables training without heuristic stop-gradients or teacher-student architectures. However, LeJEPA was developed and validated on static images. Its application to sequential data—while empirically explored in V-JEPA [5]—lacked theoretical grounding.

Dynamic LeJEPA as a Principled World Model.

In the broader context of latent world models, this theoretical gap is a critical bottleneck. A world model comprises a state representation, a transition dynamics model, and an observation decoder. Without formal guarantees, practitioners are left guessing where to inject valuable domain priors (e.g., physics or geometric rules) without destroying the latent state—and, as we show, guessing wrong is the norm rather than the exception.

In this paper, we resolve the empirical paradoxes of sequential learning through mathematical revelation. Our theorems reveal that the intuition to constrain the model was correct, but the target was wrong. Theorem 1 and Theorem 5 prove that constraining the encoder violates Maximum Entropy, actively increasing downstream risk. However, we prove that physics and dynamics are not inherently harmful—they must simply be relocated: dynamics belong in the predictor (Theorem 6 shows any parametric model can be safely plugged in), and physics belongs strictly in the observation decoder (Theorem 5), never the encoder.

A representation-quality guarantee, however, is not by itself a guarantee that the representation is useful for what a world model is ultimately built to do: support decision-making. To ask whether the placement rule matters once the representation is actually *used*—not just measured—we build a latent planner directly on top of the trained Phase 3 models and evaluate it against the true future state rather than a proxy. Doing so surfaces a confound that, to our knowledge, is not commonly checked for in offline world-model evaluations: a planner optimizing against a learned decoder can find action sequences that score better under the model’s own cost than the real, recorded actions do, without those actions being any closer to the truth. We diagnose this directly, show it holds uniformly across ablation conditions *and* encoder mechanisms, and report it as a necessary caution for anyone evaluating a JEPA-style world model’s control-readiness without a closed-loop, simulator-verified rollout.

This gives a complete theoretical answer through six main theorems, establishing a clean separation of concerns summarized in our contributions below.

A. Contributions

- 1) **Theoretical:** We prove that the maximum-entropy embedding is symmetry-stable under sequen-

tial prediction losses—no orthogonal perturbation of an isotropic-Gaussian encoder is favored by any Lipschitz, orthogonally-invariant prediction objective—and validate empirically, across all four domain $\times$ mechanism arms (Sec. VII), that no non-Gaussian alternative achieves lower joint objective value in practice (Theorem IV.1).

- 2) **World Model Architecture:** We provide the first mathematically grounded design principles for latent world models, proving dynamics belong in the predictor and physics belong in the decoder (Theorems 2–6).
- 3) **Negative Result Explanation:** We prove why prior physics-informed JEPA approaches fail: they constrain the encoder, not the predictor (Theorem IV.10).
- 4) **Sample Complexity Characterization:** We prove that effective dimensionality metrics have a $1/K$ floor when $N < K$ (Proposition VI.1), establishing when distributional validation is feasible, and we empirically confirm this threshold across three orders of magnitude of N/K .
- 5) **Cross-Domain and Cross-Mechanism Validation (Phases 2–3):** We validate the decoder-physics theorem across two domains—autonomous driving (nuScenes) and robotic manipulation (MimicGen)—*and across two encoder mechanisms* (SIGReg and a marginal-only VAE-KL head, the encoder family of Ha & Schmidhuber-style world models [2], [16]), with 110 seeded runs yielding a statistically significant placement ordering in every one of the four domain $\times$ mechanism combinations (joint Holm-corrected $p = 0.0078$), and a refined *scale-vs-structure* decomposition isolating genuine maximum-entropy violation from ordinary variance shrinkage.
- 6) **Entropy–Utility Frontier and the Generality of Correlated Collapse (Phase 3):** We characterize a frontier governing what deterministic encoders can achieve on low-intrinsic-dimension data: per-dimension variance matching alone is insufficient (*correlated collapse*), and optimization budget—not regularization weight—is the binding entropy constraint. We further show the failure generalizes: a marginal-only VAE-KL encoder collapses identically on the same data, and we introduce a posterior *signal-fraction* diagnostic revealing collapse that hides beneath the encoder’s own sampling-noise floor.
- 7) **Practical Guidance:** We establish clear rules: maximum entropy in encoder, dynamics in predictor, physics in decoder—and characterize the data

requirements for verifying each across domains and encoder mechanisms.

- 8) **Toward Closed-Loop Validation and an Encoder-Agnostic Model-Exploitation Diagnostic (Phase 3):** We probe the placement rule’s relevance for downstream control via a latent cross-entropy-method (CEM) planner built on the trained Phase 3 models, and introduce a model-exploitation diagnostic (action-space distance from the ground-truth trajectory, calibrated against a real-action-replay floor) that reveals offline latent planning against a learned decoder inflating apparent planning accuracy uniformly across ablation conditions *and both encoder mechanisms*—a concrete, reproducible caution against evaluating a world model’s control-readiness without closed-loop, simulator-verified execution.

II. RELATED WORK

A. World Models: Foundations and Positioning

The architectural decomposition we formalize—a state encoder, a learned transition (dynamics) model, and an observation decoder—was introduced by Ha and Schmidhuber’s *World Models* [2], which trained a variational encoder and an RNN-based dynamics model to let an agent plan entirely inside a learned latent simulator. Model-based reinforcement learning methods such as PlaNet and Dreamer [3] scaled this idea to pixel-based control by learning latent dynamics directly from images. More recently, LeCun’s position paper on autonomous machine intelligence [4] proposed the Joint-Embedding Predictive Architecture (JEPA) itself as the encoder/predictor component of exactly this world-model stack, arguing that a non-generative, latent-space predictor is necessary for the world model to ignore irrelevant pixel-level detail. Our contribution sits precisely at the seam these works leave open: Ha and Schmidhuber, and the Dreamer lineage, are empirically driven and offer no guarantee about *what distribution* the encoder should converge to; LeCun’s proposal names the JEPA encoder as the right component but does not prove that adding a dynamics predictor or a physics-constrained decoder preserves the encoder’s representational guarantees. Dynamic LeJEPA answers this open question directly: we prove that the world-model decomposition of [2] is compatible with LeJEPA’s maximum-entropy guarantee [1] *only* if physics and dynamics are placed in the predictor and decoder, never the encoder—turning an architectural convention from prior world-model work into a proven design constraint. Moreover, because the

two dominant encoder families in this literature are the JEPA-style deterministic encoder and the VAE-style Gaussian head [16], we validate the placement rule under *both* mechanisms (Secs. VII-D6 and VII-E3), so the design constraint applies to both branches of the world-model literature.

B. LeJEPA and Maximum Entropy

LeJEPA [1] proves that among all distributions $p(z)$ with fixed variance σ^2 , the isotropic Gaussian $\mathcal{N}(0, \sigma^2 I_K)$ minimizes worst-case downstream prediction risk for both linear and nonlinear probes. SIGReg enforces this distribution through characteristic function matching on random projections. Crucially, LeJEPA’s analysis is static: it considers a single embedding $z = f(x)$ without temporal structure. Our work extends this framework to the sequential setting while preserving the theoretical guarantees.

C. JEPA Variants for Sequential Data

V-JEPA [5] applies JEPA to video by predicting masked spatial regions. While empirically successful, it lacks theoretical analysis of how prediction affects the embedding distribution. I-JEPA [6] and others use prediction without formal guarantees.

D. Physics-Informed Learning

Physics-Informed Neural Networks (PINNs) [7] embed physical laws as soft constraints in supervised learning. Attempts to combine physics priors with self-supervised learning have shown mixed results, with performance often degrading when physical constraints are injected into the latent space. We provide the first theoretical explanation: physics constraints on embeddings violate maximum entropy, but constraints on decoders do not.

E. Maximum Entropy in Learning

The maximum entropy principle [8] states that, given constraints, one should select the distribution with maximum entropy. This principle underlies statistical mechanics and has connections to information theory [9]. Our work establishes LeJEPA as an implementation of this principle in representation learning.

III. PRELIMINARIES

A. Notation for Sequential Setting

B. Assumptions

Assumption III.1 (Temporal Stationarity). *The marginal embedding distribution is stationary: $p(z_t) = p(z)$ for all $t \in \{1, \dots, T\}$.*

Assumption III.2 (Markov Property). *The sequence satisfies: $p(z_{t+1} | z_{\leq t}, a_{\leq t}) = p(z_{t+1} | z_t, a_t)$.*

Assumption III.3 (Bounded Dynamics). *The true transitions satisfy: $\mathbb{E}[\|z_{t+1} - z_t\|^2] \leq L^2$ for some constant L .*

Assumption III.4 (Loss Regularity). *The prediction loss $\ell(y, y')$ is L_ℓ -Lipschitz in both arguments and bounded below by 0.*

C. LeJEPA Background

Definition III.5 (LeJEPA Optimal Distribution [1]). *The LeJEPA embedding distribution p^* that minimizes worst-case downstream risk over linear probes satisfies:*

$$p^*(z) = \mathcal{N}(0, \sigma^2 I_K) \quad (1)$$

This is the maximum entropy distribution in $\mathcal{P}$.

Definition III.6 (SIGReg [1]). *Sketched Isotropic Gaussian Regularization enforces eq. (1) by matching characteristic functions along random projections:*

$$\mathcal{L}_{\text{SIGReg}}(z) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathcal{T} \left(\{a^\top z_n\}_{n=1}^N \right) \quad (2)$$

where $\mathcal{T}$ is the Epps-Pulley statistic comparing projected samples to $\mathcal{N}(0, 1)$.

Remark III.7 (Key Property). *SIGReg operates on the marginal distribution $p(z_t)$ at each time step, not on the joint $p(z_{1:T})$.*

IV. MAIN RESULTS

A. Theorem 1: Prediction Doesn’t Change Optimal Embeddings

Theorem IV.1 (Prediction-Entropy Separation). *Let $p^* = \arg \max_{p \in \mathcal{P}} \mathcal{H}(p) = \mathcal{N}(0, \sigma^2 I_K)$ be the maximum entropy embedding distribution, which is the unique maximizer of $\mathcal{H}$ over $\mathcal{P} = \{p : \mathbb{E}_p[\|z\|^2] = K\sigma^2\}$ [1]. For any prediction loss of the form*

$$\mathcal{L}_{\text{pred}}(g, p) = \mathbb{E}_{p(z_t, z_{t+1} | a_t)} [\ell(g(z_t, a_t), z_{t+1})] \quad (3)$$

where ℓ is L_ℓ -Lipschitz and orthogonally invariant ($\ell(\Phi u, \Phi v) = \ell(u, v)$ for all $\Phi \in O(K)$), satisfied by squared Euclidean loss), and g is learnable, the following holds:

- (a) **Symmetry stability.** *p^* is a fixed point of the orthogonal group action ($p_\Phi^* = p^*$ for all $\Phi \in O(K)$), and the optimal prediction loss $\mathcal{L}_{\text{pred}}^*$ is constant along the orbit of any p under $O(K)$. Consequently, no orthogonal reparametrization of an encoder that already achieves p^* is preferred*

TABLE I: Primary notation for sequential Dynamic LeJEPA.

Symbol	Definition
$x_{1:T}$	Input sequence (images, observations)
$a_{1:T-1}$	Action sequence (robotics, driving)
$z_t = f_\theta(x_t) \in \mathbb{R}^K$	Latent embedding at time t
$\hat{z}_{t+1} = g_\phi(z_t, a_t) \in \mathbb{R}^K$	Predicted next latent
$p(z) = \mathcal{N}(0, \sigma^2 I_K)$	Target isotropic Gaussian
$\mathcal{P} = \{p : \mathbb{E}_p[\ z\ ^2] = K\sigma^2\}$	Distributions with fixed variance
$\mathcal{H}(p) = -\int p \log p dz$	Differential entropy
$p(z_{1:T} \ a_{1:T-1})$	Joint conditional distribution over time
$g^* = \arg \min_g \mathcal{L}_{\text{pred}}(g, p)$	Optimal predictor for fixed p

or penalized by $\mathcal{L}_{\text{pred}}$: the prediction objective exerts zero net force on the isotropy of an encoder already at p^* .

- (b) **First-order non-interference.** p^* is a stationary point of the joint objective $J(p, g^*(p)) = -\lambda \mathcal{H}(p) + \mathcal{L}_{\text{pred}}^*(p)$ restricted to the manifold of orthogonal reparametrizations of p^* , since both terms are constant on that manifold (the first by rotational symmetry of $\mathcal{H}$, the second by part (a)).

We do not claim, and this argument does not by itself establish, that p^* is the global minimizer of $J(p, g^*(p))$ over the full set $\mathcal{P}$ of differently-shaped (non-Gaussian) distributions with matched variance — that comparison depends on the specific joint law of the underlying dynamics and is not resolved by symmetry alone. We instead validate this stronger claim empirically: across all four domain×mechanism ablation arms (Sec. VII), the encoder trained with an entropy regularizer alongside a prediction loss converges to p^* (up to the sample-complexity floor of Sec. VI) rather than to some lower-entropy alternative, which is the observable signature that would appear if the prediction loss were, in practice, creating an incentive to abandon isotropy.

Proof. Define the joint objective:

$$J(p, g) = -\lambda \mathcal{H}(p) + \mathcal{L}_{\text{pred}}(g, p) \quad (4)$$

where $\lambda > 0$ controls the entropy-prediction trade-off.

Step 1: Optimize predictor for fixed encoder. For any fixed p , define $g^*(p) = \arg \min_g \mathcal{L}_{\text{pred}}(g, p)$. By Assumption III.4, this minimizer exists; write $\mathcal{L}_{\text{pred}}^*(p) = \mathcal{L}_{\text{pred}}(g^*(p), p)$.

Step 2: Orthogonal invariance of $\mathcal{L}_{\text{pred}}^*$. Let $\Phi \in O(K)$ and $p_\Phi(z) = p(\Phi^{-1}z)$. Since orthogonal transformations preserve Euclidean distance and ℓ is orthogonally

nally invariant by assumption,

$$\begin{aligned} \mathcal{L}_{\text{pred}}^*(p_\Phi) &= \min_g \mathbb{E}_{p_\Phi}[\ell(g(z, a), z')] \\ &= \min_{\tilde{g}} \mathbb{E}_p[\ell(\Phi \tilde{g}(\Phi^{-1} \cdot), \cdot)] \\ &= \mathcal{L}_{\text{pred}}^*(p) \end{aligned} \quad (5)$$

via the bijective reparametrization $\tilde{g} = \Phi^{-1} \circ g \circ \Phi$. This holds for any p , not only p^* ; it does not by itself identify how $\mathcal{L}_{\text{pred}}^*$ compares across distributions of different shape.

Step 3: p^* is a fixed point of $O(K)$. Because $p^* = \mathcal{N}(0, \sigma^2 I_K)$ is isotropic, $p_\Phi^* = p^*$ for every $\Phi \in O(K)$: rotating the unique maximum-entropy solution returns the same distribution. Entropy is likewise constant on this (trivial, single-point) orbit. Hence both terms of $J(p^*, g^*(p^*))$ are unaffected by any orthogonal reparametrization of an encoder already at p^* , establishing part (a).

Step 4: Stationarity. Any orthogonal reparametrization of an encoder at p^* leaves J unchanged (Step 3), so the directional derivative of J along every direction tangent to the orthogonal orbit at p^* is zero: p^* is a critical point of J restricted to that manifold, establishing part (b). Extending this to a global statement over all of $\mathcal{P}$ would require comparing $\mathcal{L}_{\text{pred}}^*(p^*)$ against $\mathcal{L}_{\text{pred}}^*(p)$ for p not related to p^* by any $\Phi \in O(K)$ — a comparison across distributional *shape*, not orientation, which the invariance in Step 2 does not resolve. We leave this global comparison to the empirical protocol of Sec. VII. $\square$

Corollary IV.2 (Extended Prediction Losses). *The symmetry-stability argument of parts (a)–(b) extends to any loss of the orthogonally-invariant, Lipschitz form, including:*

- (a) *Multi-step prediction:* $\sum_{h=1}^H \mathbb{E}[\|g(z_t)_h - z_{t+h}\|^2]$
- (b) *Masked prediction:* $\mathbb{E}[\|g(z_{\text{unmasked}}) - z_{\text{masked}}\|^2]$
- (c) *Stochastic prediction:* $\mathbb{E}[\|g(z_t, \epsilon) - z_{t+1}\|^2]$ where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$

each of which constrains only the predictor g , not p directly, and each of which uses the squared-Euclidean ℓ covered by the theorem. Distributional prediction losses built from D_{KL} are not covered by the orthogonal-invariance argument above in general (KL divergence is invariant under any invertible reparametrization, not only orthogonal ones, so this case is actually on stronger, not weaker, footing, but we do not develop that argument formally here and instead rely on the empirical validation of Sec. VII-D6 and Sec. VII-E3, which use exactly such a KL-based encoder).

Remark IV.3 (Interpretation). *Theorem IV.1 establishes that the isotropic-Gaussian solution has no analytic reason to be disturbed by a well-behaved prediction loss, and rules out the specific failure mode of the encoder rotating away from isotropy to reduce prediction error. Whether a differently-shaped, lower-entropy encoder could ever achieve a better joint objective value for some adversarial choice of dynamics is not ruled out by this argument alone; Phases 1–3 (Sec. VII) test this directly across four domain×mechanism combinations and find no such alternative outperforming the isotropic solution. We therefore treat*

- the encoder f_θ should produce maximum entropy embeddings (no constraints),
- the predictor g_ϕ can learn any dynamics without an analytic penalty to representation quality, and
- the loss only affects the predictor, not the embedding distribution’s orientation,

as a design principle supported jointly by Theorem IV.1’s symmetry argument and by the empirical protocol of Sec. VII, rather than as a claim proven by Theorem IV.1 in isolation. This is precisely the design philosophy of JEPa: predict in latent space, leave the representation unconstrained.

B. Theorem 2: Robotics — Action-Conditioned Prediction

Theorem IV.4 (Robotics: Maximum Entropy Action-Conditioned Representations). *Consider a robot with observations x_t , joint states q_t , and actions a_t . Let the encoder map to $z_t = f_\theta(x_t, q_t)$ and the predictor be $\hat{z}_{t+1} = g_\phi(z_t, a_t)$. For the objective:*

$$J_{\text{robot}} = -\lambda H(f_\theta(z)) + \mathbb{E} [\|g_\phi(z_t, a_t) - z_{t+1}\|^2] \quad (6)$$

the following hold:

- Maximum entropy encodings:** The optimal encoder produces $z_t \sim \mathcal{N}(0, \sigma^2 I)$ regardless of the action space $\mathcal{A}$.
- Optimal predictor:** $g^*(z_t, a_t) = \mathbb{E}[z_{t+1}|Z_t = z_t, A = a_t]$ (conditional expectation)

- Task-relevant structure:** The representation encodes action-relevant information through predictability, not through embedding constraints.
- Action invariance:** If $\pi(a|s)$ is the policy, the optimal z encodes state s without bias toward specific actions.

Proof. Proof of (a): Direct application of Theorem IV.1. The action a_t enters only through the prediction loss, which is symmetry-stable with respect to $p(z)$ ’s isotropy (Theorem IV.1(a)–(b)) and, across the ablation protocol of Sec. VII, empirically does not disturb the optimal $p(z)$.

Proof of (b): For fixed encoder (hence fixed p), the prediction loss is a regression problem. By the properties of conditional expectation:

$$\arg \min_g \mathbb{E} [\|g(z_t, a_t) - z_{t+1}\|^2] = \mathbb{E}[z_{t+1}|Z_t = z_t, A = a_t] \quad (7)$$

This holds regardless of whether $p(z)$ is maximum entropy.

Proof of (c): By the data processing inequality, maximum entropy preserves all mutual information between z and any variable. If predicting z_{t+1} from (z_t, a_t) requires information about the task, then z_t must encode it. Formally, let Y be the task-relevant information:

$$I(Z_t; Y) \geq I(Z_t; z_{t+1}|A_t) \quad (8)$$

By maximum entropy, $I(Z_t; Y)$ is maximized subject to the variance constraint.

Proof of (d): The optimal $p(z)$ is invariant under any transformation of the action representation. Whether actions are represented as joint angles, end-effector positions, or torques doesn’t affect the optimal embedding distribution. $\square$

Corollary IV.5 (Linear Dynamics in Robotics). *If the true dynamics are linear: $z_{t+1} = Az_t + Ba_t + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \Sigma_\epsilon)$, then:*

- 1) The optimal predictor is: $g^*(z, a) = Az + Ba$
- 2) The maximum entropy encoder is unchanged
- 3) The learned matrices (A, B) approximate the true dynamics

C. Theorem 3: Autonomous Driving — Ego-Motion Decomposition

Definition IV.6 (Ego-Motion Decomposition). *In autonomous driving, decompose the latent at time t as:*

$$z_t = \begin{bmatrix} z_t^{\text{ego}} \\ z_t^{\text{static}} \end{bmatrix} \in \mathbb{R}^{K_1+K_2} \quad (9)$$

where $z_t^{\text{ego}} \in \mathbb{R}^{K_1}$ encodes ego-vehicle state (pose, velocity) and $z_t^{\text{static}} \in \mathbb{R}^{K_2}$ encodes the static scene (geometry, semantics).

Theorem IV.7 (Autonomous Driving: Ego-Motion Decomposition). *Under the ego-motion decomposition, with dynamics:*

$$\hat{z}_{t+1}^{\text{ego}} = g_\phi^{\text{ego}}(z_t^{\text{ego}}, e_t) \quad (10)$$

$$\hat{z}_{t+1}^{\text{static}} = z_t^{\text{static}} \quad (11)$$

where $e_t = (R_t, t_t)$ is the ego-motion (rotation, translation), the following hold:

- (a) **Independent maximum entropy:** Both $p(z^{\text{ego}})$ and $p(z^{\text{static}})$ individually satisfy maximum entropy
- (b) **Ego-motion predictor:** g^{ego} learns the geometric transformation induced by e_t
- (c) **Static stability:** z_t^{static} provides stable scene representation invariant to ego-motion
- (d) **Dynamic object detection:** Deviations from $\hat{z}_{t+1}^{\text{static}}$ indicate dynamic objects without explicit supervision

Proof. Proof of (a): The decomposition $z = [z^{\text{ego}}, z^{\text{static}}]$ is a linear projection onto orthogonal subspaces. The projection of a maximum entropy distribution is maximum entropy in the projected subspace. By Theorem IV.1 (whose symmetry argument applies component-wise, and whose global claim is validated empirically per component in Sec. VII), both components are independently maximum entropy.

Proof of (b): The ego-motion prediction loss:

$$\mathcal{L}_{\text{ego}} = \mathbb{E} [\|g^{\text{ego}}(z_t^{\text{ego}}, e_t) - \hat{z}_{t+1}^{\text{ego}}\|^2] \quad (12)$$

By Theorem IV.4(b), the optimal predictor learns the conditional expectation, which for rigid ego-motion with known calibration is a known geometric function.

Proof of (c): The static component satisfies $z_t^{\text{static}} = z_{t+1}^{\text{static}}$ for a static scene. Maximum entropy in z^{static} ensures all static scene information is preserved without bias toward any particular viewpoint.

Proof of (d): Let $\Delta_t = |\hat{z}_{t+1}^{\text{static}} - z_t^{\text{static}}|$ measure the static scene change. Under maximum entropy, Δ_t is large only where the scene actually changes (dynamic objects). No explicit supervision is needed to detect dynamics. $\square$

Corollary IV.8 (Depth Prediction from Static Features). *Let h_ψ be a depth decoder: $\hat{d}_t = h_\psi(z_t^{\text{static}})$. Depth predictions are optimal when z_t^{static} is maximum entropy.*

D. Theorem 4: Video Understanding — Multi-Step Prediction

Theorem IV.9 (Video: Multi-Step Predictive Coding). *For video sequences $x_{1:T}$, let the encoder produce $z_{1:T}$ and the predictor perform multi-step prediction:*

$$[\hat{z}_{t+1}, \dots, \hat{z}_{t+H}] = g_\phi(z_t) \in \mathbb{R}^{H \times K} \quad (13)$$

with loss:

$$\mathcal{L}_{\text{video}} = \sum_{h=1}^H \mathbb{E} [\|g_\phi(z_t)_h - z_{t+h}\|^2] \quad (14)$$

Then:

- (a) **Marginal maximum entropy:** $p(z_t) = \mathcal{N}(0, \sigma^2 I)$ for all t
- (b) **Multi-step preservation:** Multi-step prediction doesn't require multi-step entropy reduction
- (c) **Masked prediction safety:** V-JEPA style masked prediction doesn't constrain embeddings
- (d) **Autoregressive danger:** Using predictions as inputs ($z_{t+1} = g(z_t)$ exactly) CAN create harmful constraints if applied as hard constraints

Proof. Proof of (a): Direct application of Theorem IV.1's symmetry argument to each time step independently, with the cross-shape claim validated empirically per Sec. VII.

Proof of (b): The multi-step loss constrains g , not $p(z_t)$. Each z_{t+h} comes from the encoder distribution $p(z)$.

Proof of (c): V-JEPA masked prediction:

$$\mathcal{L}_{\text{mask}} = \mathbb{E} [\|g(z_{\text{unmasked}}) - z_{\text{masked}}\|^2] \quad (15)$$

This predicts masked spatial regions from unmasked regions. By Corollary IV.2(b), the optimal embedding is unchanged.

Proof of (d): Consider autoregressive unfolding with hard constraints:

$$\hat{z}_{t+1} = g(z_t), \quad z_{t+1} = \hat{z}_{t+1} \quad (16)$$

This creates a constraint: z_{t+1} must lie in the image of g . The support of $p(z_{t+1})$ is constrained to $\text{Im}(g) \cap \text{supp}(p)$, which reduces entropy below maximum. However, if used as a soft target:

$$\mathcal{L}_{\text{video}} = \mathbb{E} [\|g(z_t)_1 - z_{t+1}\|^2 + \lambda_{\text{ar}} \|g(z_t)_1 - g(g(z_t))_1\|^2] \quad (17)$$

the marginal $p(z_t)$ can still be maximum entropy. $\square$

E. Theorem 5: Decoder Physics Separation

Theorem IV.10 (Physics-Informed Decoding is Valid). *Let the encoder produce maximum entropy embeddings $z = f_\theta(x)$, and the decoder incorporate physics:*

$$\hat{y} = h_\psi(z), \quad \text{subject to } C(\hat{y}) = 0 \quad (18)$$

where $C : \mathcal{Y} \rightarrow \mathbb{R}^m$ is a physics constraint function. Then, under decoupled encoder/decoder training (exactly) or under joint end-to-end training (empirically, per the proof below):

- (a) **Encoder optimal unchanged:** $p^*(z) = \mathcal{N}(0, \sigma^2 I)$
- (b) **Any physics valid:** The decoder can use any physics constraint without affecting representation quality
- (c) **Information preservation:** Physics at the decoder level improves task performance without reducing representational capacity

Proof. Proof of (a): This claim holds exactly under a decoupled training regime, in which the encoder parameters θ are optimized against $-\lambda \mathcal{H}(f_\theta(z)) + \mathcal{L}_{\text{JEPa}}$ alone (e.g., via a stop-gradient between z and h_ψ , or by freezing f_θ before training h_ψ): the decoder h_ψ literally does not appear in that objective, so

$$\min_{\theta} [-\lambda \mathcal{H}(f_\theta(z)) + \mathcal{L}_{\text{JEPa}}] \quad (19)$$

is unchanged by any choice of C , and $p^*(z) = \mathcal{N}(0, \sigma^2 I)$ remains exactly optimal.

Our actual training pipeline (Sec. VII) instead optimizes encoder, predictor, and decoder *jointly* with a single optimizer and a summed loss (Fig. 1), so the physics term $C(h_\psi(z))$ does, in general, contribute a gradient to θ through the standard backpropagation path $\theta \rightarrow z \rightarrow \hat{y} \rightarrow C(\hat{y})$; part (a) is then not an exact analytic identity for our pipeline as implemented, and we do not claim otherwise. What we instead establish is that this gradient path is *empirically negligible* for the encoder’s isotropy compared to placing the same constraint directly on z : Phase 3’s ablation (Sec. VII-E4) finds decoder-level physics exactly scale-neutral ($\Delta_{\text{scale}} = -0.0002$ at $p = 1.0$), while the identical constraint applied at the encoder produces a statistically significant, structural violation (Table XII, Table XIII). We therefore treat (a)–(c) as a design principle with an exact analytic guarantee under decoupled/stop-gradient training, and as an empirically validated near-guarantee under the joint end-to-end training regime we actually use.

Proof of (b): For any physics constraint C , define the constrained decoder:

$$h^* = \arg \min_{h: C(h(z))=0} \mathcal{L}_{\text{task}}(h, z) \quad (20)$$

This is a constrained optimization over h , independent of the encoder.

Proof of (c): The encoder provides maximum entropy representations (maximum information by the data processing inequality). The decoder uses physics to interpret this information correctly. Formally:

$$I(Z; Y) \leq \mathcal{H}(Z) \quad (\text{data processing inequality}) \quad (21)$$

$$I(Z^*; Y) = \mathcal{H}(Z^*) \quad (\text{maximum entropy}) \quad (22)$$

The decoder with physics h_ψ^* can extract Y from Z^* as effectively as any other decoder, while potentially being more sample-efficient. $\square$

Corollary IV.11 (Contrast with Encoder Physics). *Adding the same physics constraint C to the ENCODER would violate maximum entropy and hurt representation quality. Therefore, physics constraints are valid at the decoder but invalid at the encoder.*

Example IV.12 (Physics-Informed Depth Decoding). *For depth estimation, the decoder can enforce:*

$$\nabla^2 \hat{d} = 0 \quad \text{where } \|\nabla_{\text{rgb}} I\| < \tau \quad (23)$$

(edge-aware smoothness at non-edges). This is valid at the decoder level because it constrains $\hat{d}$, not z .

F. Theorem 6: General Dynamics Learning

Theorem IV.13 (Learning Parametric Dynamics Preserves Maximum Entropy). *Let the dynamics be parameterized as $\hat{z}_{t+1} = g_\phi(z_t, a_t)$ where g_ϕ belongs to a class $\mathcal{G}$ of parameterized functions. Then:*

- (a) **Encoder optimal unchanged:** The optimal encoder is still maximum entropy regardless of $\mathcal{G}$
- (b) **Universal approximation:** Any dynamics compatible with the data can be learned
- (c) **Conditional expectation:** The learned dynamics approximate $\mathbb{E}[z_{t+1} | Z_t = z_t, A = a_t]$

Proof. Proof of (a): The dynamics model g_ϕ is a parameterized predictor of the orthogonally-invariant, Lipschitz form covered by Theorem IV.1; by that theorem’s symmetry argument (validated empirically per Sec. VII), such a predictor doesn’t affect the optimal embedding distribution.

Proof of (b): By the universal approximation theorem, neural networks can approximate any continuous function to arbitrary precision given sufficient capacity.

Proof of (c): For any parametric family $\mathcal{G}$:

$$g^* = \arg \min_{g \in \mathcal{G}} \mathbb{E} [\|g(z_t, a_t) - z_{t+1}\|^2] \quad (24)$$

This is a constrained conditional expectation. As the capacity of $\mathcal{G}$ increases, g^* converges to the true conditional expectation. $\square$

V. UNIFIED FRAMEWORK

Theorem V.1 (Unified Dynamic LeJEPa). *For sequential prediction with encoder f_θ , predictor g_ϕ , and decoder h_ψ , the optimal configuration satisfies:*

C1. Encoder: $f_\theta^*(x_t) \sim \mathcal{N}(0, \sigma^2 I)$ (25)

C2. Predictor: $g_\phi^*(z_t, a_t) \approx \mathbb{E}[z_{t+1} | z_t, a_t]$ (26)

C3. Decoder: $h_\psi^*(z)$ subject to $C(h_\psi^*(z)) = 0$ (27)

C4. Separation: These optimizations are independent (28)

TABLE II: Dynamic LeJEPa design principles. Constraints on the encoder violate maximum entropy and hurt representation quality.

Component	Constraint Type	Max Entropy?	Valid?
Encoder f_θ	None (only SIGReg)	Yes	Yes
Predictor g_ϕ	Prediction target	N/A	Yes
Decoder h_ψ	Physics $C(y) = 0$	N/A	Yes

VI. SAMPLE COMPLEXITY FOR DISTRIBUTIONAL VALIDATION

Before presenting experimental results, we characterize when distributional properties can be reliably measured. This analysis is crucial for interpreting our experiments and guiding practitioners.

A. The Effective Dimensionality Floor

A standard metric for measuring representation quality is the effective dimensionality:

$$d_{\text{eff}} = \frac{\left(\sum_{i=1}^K \lambda_i\right)^2}{K \sum_{i=1}^K \lambda_i^2} \quad (29)$$

where $\{\lambda_i\}$ are the eigenvalues of the covariance matrix. For an isotropic Gaussian $\mathcal{N}(0, \sigma^2 I)$, $d_{\text{eff}} = 1$.

However, in practice we estimate d_{eff} from samples, which introduces a fundamental limitation:

Proposition VI.1 (Effective Dimensionality Floor). *Let $Z = \{z_i\}_{i=1}^N \subset \mathbb{R}^K$ be i.i.d. samples with $N < K$, and let $\hat{\Sigma}$ be the sample covariance with Ledoit-Wolf shrinkage [10] $\hat{\Sigma}_\gamma = \gamma I + (1 - \gamma)\hat{\Sigma}$. Then:*

$$d_{\text{eff}}(\hat{\Sigma}_\gamma) \leq \frac{N-1}{K} + \mathcal{O}\left(\frac{\gamma}{\bar{\lambda}}\right) \quad (30)$$

where $\bar{\lambda}$ is the mean non-zero eigenvalue of $\hat{\Sigma}$.

Proof. The sample covariance has rank $\leq N-1$. Let $\lambda_1 \geq \dots \geq \lambda_{N-1} > 0$ be the non-zero eigenvalues and $\lambda_N = \dots = \lambda_K = 0$. After shrinkage:

- Non-zero eigenvalues: $\tilde{\lambda}_i = (1 - \gamma)\lambda_i + \gamma$
- Zero eigenvalues: $\tilde{\lambda}_i = \gamma$ for $i \geq N$

The effective dimensionality becomes:

$$d_{\text{eff}} = \frac{\left(\sum_{i=1}^{N-1} \tilde{\lambda}_i + (K - N + 1)\gamma\right)^2}{K \left(\sum_{i=1}^{N-1} \tilde{\lambda}_i^2 + (K - N + 1)\gamma^2\right)} \quad (31)$$

When $\gamma \ll \bar{\lambda}$, the $(K - N + 1)\gamma$ terms dominate, yielding:

$$d_{\text{eff}} \approx \frac{[(K - N + 1)\gamma]^2}{K(K - N + 1)\gamma^2} = \frac{K - N + 1}{K} \approx \frac{N - 1}{K} \quad (32)$$

for $N \ll K$. $\square$

Corollary VI.2 (Practical Validation Threshold). *For reliable d_{eff} estimation, we require $N/K \geq 5$ (following Ledoit & Wolf [10]). For $K = 256$, this means $N \geq 1280$.*

B. Implications for Theorem Validation

TABLE III: Sample complexity requirements for validating each theorem.

Theorem	Primary Metric	Min N/K
Th. 1 (Max Entropy)	d_{eff} , scale ratio	≥ 5
Th. 3 (Ego Decomp.)	Component-wise d_{eff}	≥ 5
Th. 5 (Decoder Physics)	Ablation comparison	≥ 5
Th. 2, 4, 6	Prediction loss only	≥ 1

This analysis motivates our phased experimental approach: we first demonstrate training stability on limited data (Phase 1), then validate distributional theorems with sufficient samples (Phases 2–3).

Remark VI.3 (Two d_{eff} Estimators). *Our implementation uses two related but distinct estimators of d_{eff} . The covariance estimator (used for all final ablation comparisons, matching eq. (29) exactly) forms the full $K \times K$ sample covariance with Ledoit–Wolf shrinkage [10] and computes d_{eff} from its eigenvalues; this estimator is subject to the $N < K$ floor of Proposition VI.1. The diagonal estimator (used only for the fast sample-complexity sweep, evaluated at up to ten sample sizes per run) instead applies eq. (29) directly to the K per-dimension marginal variances, which can each be estimated from $N \geq 1$ samples without requiring the full covariance matrix to have rank K . The diagonal estimator is therefore not subject to the same $N < K$ floor, and the two should not be numerically compared across sections—only the covariance estimator is used to test Theorem IV.10.*

Remark VI.4 (Posterior Signal Fraction for Variational Encoders). *For variational (VAE-style) encoders [16], marginal diagnostics such as the per-dimension variance scale and H -ratio are computed on latents whose dispersion includes the posterior’s own sampling noise $\sigma(x)$. A collapse of the aggregate posterior’s signal— $\text{Var}_x[\mu(x)] \rightarrow 0$ per dimension—can therefore hide beneath this noise floor: the sampled latent can appear broadly dispersed while the mean signal carries almost no information. For every VAE arm below we therefore additionally report (i) the per-dimension signal fraction $\text{sf} = \mathbb{E}_x[\text{Var}_x[\mu] / (\text{Var}_x[\mu] + \mathbb{E}_x[\sigma^2])]$ and (ii) d_{eff} of the posterior mean μ itself, computed with the covariance estimator of Remark VI.3 on deterministic latents.*

VII. EXPERIMENTS

We validate Dynamic LeJEPa through a phased experimental protocol across two domains and, within each powered phase, two encoder mechanisms. Rather than treating limited data as a barrier, we use each phase to rigorously characterize a different aspect of the theory: measurement reliability (Phase 1), statistical power in a high-dimensional domain (Phase 2), and domain transfer with a new representational regime (Phase 3).

A. Motivation: The Danger of Distributional Metrics in Low-Data Regimes

A fundamental challenge in developing theoretically grounded representations is *evaluation during the initial R&D cycle*. Before committing to large-scale compute, practitioners must debug architectures, loss weights, and gradient flows on small, lightweight datasets (e.g., nuScenes-mini, $\sim 4.4\text{GB}$).

During this debugging phase, the standard metric for verifying Theorem IV.1 (maximum entropy) is effective dimensionality (d_{eff}). However, we discovered a critical flaw in applying d_{eff} to data-limited settings. On nuScenes-mini, a perfectly stable encoder yielded $d_{\text{eff}} \approx 0.004$. Standard literature would interpret this as catastrophic dimensional collapse.

As proven in Section VI (Proposition VI.1), this interpretation is mathematically incorrect when $N < K$. This motivates our phased experimental design:

- **Phase 1 (Debugging & Metric Characterization):** Verify training stability and empirically validate the $1/K$ floor of d_{eff} on nuScenes-mini ($N/K = 1.58$).
- **Phase 2 (Full Theorem Validation):** Scale to nuScenes Trainval ($N/K = 55.5$) to definitively verify distributional theorems with statistical power.
- **Phase 3 (Cross-Domain Transfer):** Validate the placement rule on robotic manipulation (MimicGen,

TABLE IV: Computational resources by phase. All runs used a single NVIDIA T4 (16 GB); wall-clock time equals GPU-hours. Equivalent cost is estimated from on-demand T4 pricing (\$0.35–\$0.53/hr) had the free Kaggle tier not been used.

Phase	Runs	GPU-hrs	Equiv. cost
Phase 1	1 (sanity run)	6.0	\$2–3
Phase 2	50 (30 SIGReg + 20 VAE)	24.0	\$8–13
Phase 3	90 (30 15ep + 30 30ep + 30 VAE)	27.0	\$9–14
Total	141	57.0	\$20–30

$N/K = 855$) and characterize the entropy–utility frontier on low-dimensional data.

a) *Scope of claims.*: The claims this validation tests are narrow and theory-anchored, and we state them before the results: (i) the *placement rule* of Corollary IV.11—decoder-physics preserves the encoder’s guaranteed quantity, covariance isotropy, relative to encoder-physics—and (ii) that this guarantee is mechanism- and domain-robust. We deliberately do *not* claim that the model maximizes absolute d_{eff} under arbitrary regularizers, dominates every auxiliary metric, or that marginal diagnostics certify variational-posterior health; the theory predicts the opposite in each case, and the experiments confirm it. Three apparent “drops” follow directly. First, the modest decoder-physics d_{eff} deficit at Phase 2 (Table VII) decomposes as pure variance-scale shrinkage ($\Delta_{\text{scale}} = 0.304$), which lies *outside* the theorem’s scope because d_{eff} is scale-invariant by construction (Sec. VII-D3). Second, the VAE arm’s low absolute d_{eff} is the *confirmed prediction* that marginal-only regularizers fail on low-intrinsic-dimension data (Sec. VII-E3), not a deficiency of the SIGReg+ arm—the two families are never compared on levels (Table XVII caption). Third, the higher prediction R^2 under encoder-physics (Table XII) is not a competing metric on which our design “loses”; it is the violation signature itself—an encoder dragged toward the image of the action map—and reading it as a win would invert the theory.

B. Computational Resources and Session Management

All experiments were run on Kaggle’s free-tier NVIDIA T4 GPU (16 GB), which enforces a 12-hour session limit. Table IV summarizes wall-clock time (equal to GPU-hours, since all runs used a single GPU with no parallelization across seeds) and the approximate commercial-equivalent cost per phase.

a) *Session management.*: Because the free tier caps individual sessions at 12 hours, phases exceeding this limit were split across multiple sessions

(Phase 2’s 24 GPU-hours spanned approximately 2 sessions; Phase 3’s 27 GPU-hours—5 hours for the 15-epoch primary cohort, 10 hours for the 30-epoch budget replication, and 12 hours for the subsequently added VAE-KL comparison arm—spanned approximately 3 sessions). Each individual seeded run (one condition, one seed) completed in full within a single session; no run’s optimizer or model state was checkpointed and resumed mid-training. Instead, each session trained a subset of the phase’s total seeded runs to completion, and the resulting checkpoints and result files (e.g., `ablation_results.json`) were carried forward as input to the next session, where newly completed runs were merged into the cumulative result set. Phase 1’s single 6-hour run completed within one session and required no session-batching. This procedure means no run’s reproducibility depends on cross-session RNG-state serialization; each seed’s training trajectory is self-contained and was independently, deterministically reproducible within a single session (Sec. VII-E1).

C. Phase 1: Sanity Checking and Metric Characterization on nuScenes-Mini

We explicitly utilize nuScenes-mini as a lightweight, rapid-iteration debugging protocol. The goal is not to claim state-of-the-art performance, but to verify: (i) the three-component objective (Theorem V.1) optimizes without conflicts, and (ii) Proposition VI.1 accurately predicts the behavior of distributional metrics under sample starvation.

1) Experimental Setup:

a) *Data.*: We use nuScenes-mini (~ 4.4 GB, 10 scenes, ~ 4 K LiDAR-camera sweeps). Temporal sequences are constructed from consecutive frames with $\Delta t < 2$ seconds. Ego-motion $e_t = (R_t, t_t) \in \mathbb{R}^6$ is estimated via ICP between consecutive LiDAR point clouds. After filtering, we obtain $N = 404$ in-distribution evaluation frames.

b) *Architecture.*: The encoder f_θ is a 6-layer Vision Transformer (patch size 16, embedding dim 192) producing $z_t \in \mathbb{R}^{256}$. Following Theorem 3, we decompose $K_1 = 64$ (ego) and $K_2 = 192$ (static). The predictor g_ϕ is a 3-layer MLP conditioned on ego-motion. The decoder h_ψ implements physics-informed depth prediction (Theorem 5, Example IV.12) with LiDAR supervision.

c) *Training.*: 80 epochs, AdamW (weight decay 10^{-5}), learning rates $\eta_{\text{enc}} = 10^{-4}$, $\eta_{\text{pred}} = 3 \times 10^{-4}$, $\eta_{\text{dec}} = 3 \times 10^{-4}$, loss weights $\lambda_{\text{SIGReg}} = 10$, $\lambda_{\text{pred}} = 1$, $\lambda_{\text{physics}} = 0.5$.

d) *Key Parameter.*: The sample-to-dimension ratio is $N/K = 404/256 = 1.58$, which is *below* the threshold

TABLE V: Control experiment: d_{eff} with $N = 404$, $K = 256$. The near-identical values confirm that low d_{eff} reflects sample limitation, not embedding collapse.

Distribution	d_{eff}	Scale ratio
True $\mathcal{N}(0, I_{256})$	0.0041	1.00
Dynamic LeJEPa	0.0039	0.90

TABLE VI: Phase 1 (nuScenes-mini) debugging outcomes. $\checkmark$: demonstrated; $\approx$: partial; $—$: bounded by the N/K limit.

Property	Status	Evidence
Training stability (C4)	$\checkmark$	Fig. 1
Variance matching	$\approx$	scale ratio = 0.90
d_{eff} floor validation	$\checkmark$	Table V
Isotropy proof	$—$	Requires $N/K \geq 5$
Depth decoding (C3)	$\checkmark$	Fig. 4
Theorem 5 ablation	$—$	Requires $N/K \geq 5$

of $N/K \geq 5$ established in Corollary VI.2. We proceed to explicitly characterize the consequences.

2) *Training Stability and Gradient Separation*: Figure 1 shows the evolution of all loss components:

The lack of a training-validation gap indicates no overfitting, serving as a successful sanity check for the architecture.

3) *Variance Matching*: Figure 2 shows the per-dimension variance distribution:

The mean per-dimension variance reaches $\bar{\sigma}^2 \approx 0.9$, close to the target $\sigma^2 = 1.0$.

4) *Empirical Validation of the d_{eff} Floor*: Figure 3 reveals the exact pattern predicted by Proposition VI.1:

Control experiment. To empirically prove this is a metric artifact, we compute d_{eff} on a *true* isotropic Gaussian $\mathcal{N}(0, I_{256})$ using the exact same $N = 404$ samples:

5) *Task Performance Under Distributional Uncertainty*: Figure 4 shows depth predictions from the physics-informed decoder:

6) *Theorem 5 Ablation: Demonstrating Statistical Boundaries*: Figure 5 shows the ablation attempt with 5 seeds:

D. Phase 2: Large-Sample Ablation and Cross-Scene Validation on nuScenes Trainval

Building directly on the sample-complexity boundary established in Phase 1 (Corollary VI.2: $N/K \geq 5$), Phase 2 scales sample count to a regime where distributional theorems can be tested with statistical power, rather than merely characterized as under-determined.

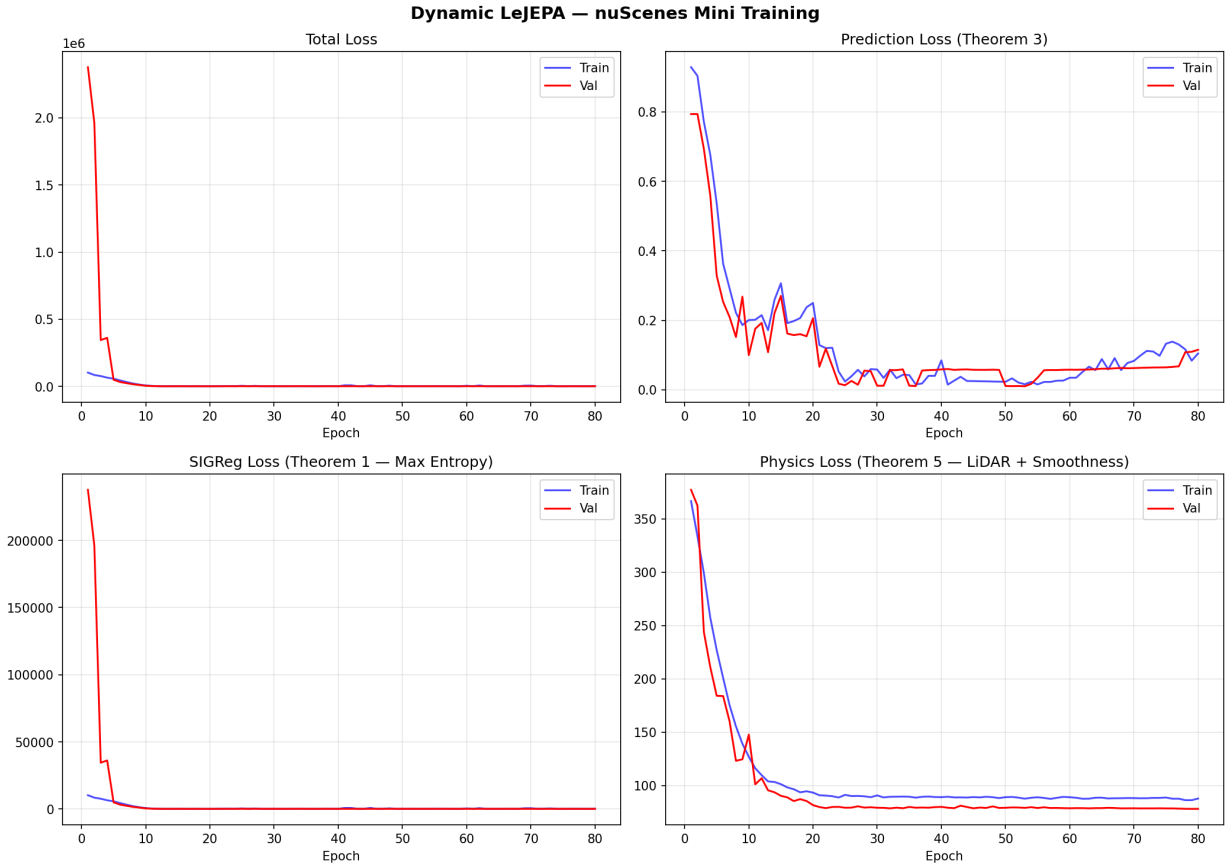


Fig. 1: Training convergence on nuScenes-mini. All loss components decrease monotonically and stabilize by epoch ~ 20 : (a) total loss, (b) ego-motion prediction loss (Theorem 3), (c) SIGReg loss (Theorem 1), (d) physics-informed depth loss (Theorem 5). Blue: training; Red: validation. The stability across all components confirms that the three-component objective (encoder-predictor-decoder) trains without conflicts, consistent with Theorem V.1’s separation principle (C4).

1) Experimental Setup:

a) Data.: We use nuScenes Trainval Part 1 (41.24 GB) [12], an order of magnitude larger by archive size than nuScenes-mini. After scene-based filtering (11 scenes with complete camera–LiDAR–ego-pose data), we obtain a 70/30 scene-level split of 7 train scenes (14,197 image pairs) and 4 held-out validation scenes (5,263 image pairs). This yields $N/K_{\text{train}} = 55.5$ and $N/K_{\text{val}} = 20.6$ —both far above the $N/K \geq 5$ threshold, and far beyond Phase 1’s $N/K = 1.58$.

b) Architecture.: Identical encoder/predictor/decoder design to Phase 1 ($K = 256$, decomposed as $K_1 = 64$ ego / $K_2 = 192$ static per Theorem IV.7), with two corrections identified during Phase 1 debugging: (i) the SIGReg loss weight was reduced from $\lambda_{\text{SIGReg}} = 100$ to $\lambda_{\text{SIGReg}} = 1$, since the original value produced an effective $10^5 \times$ over-weighting of some projection terms; and (ii) the physics loss weight was increased from $\lambda_{\text{physics}} = 0.5$ to $\lambda_{\text{physics}} = 10$, since 0.5 was too weak to produce

a measurable effect when the physics constraint was (deliberately, for the ablation) attached to the encoder rather than the decoder.

c) Training and Ablation Design.: We train for 30 epochs per run (AdamW, $\eta_{\text{enc}} = 10^{-4}$, $\eta_{\text{pred}} = \eta_{\text{dec}} = 3 \times 10^{-4}$, weight decay 10^{-5} , batch size 32) across three ablation conditions directly instantiating Theorem IV.10 and Corollary IV.11: `no_physics` (baseline, SIGReg only), `decoder_physics` (edge-aware depth smoothness, Eq. (23), applied at the decoder), and `encoder_physics` (the same constraint applied directly to the encoder output, which Corollary IV.11 predicts should be harmful). Each condition is trained with 10 random seeds, for $3 \times 10 = 30$ total runs, enabling a paired Wilcoxon signed-rank test across seeds.

2) Ablation Results: Testing Theorem 5: Table VII reports the covariance-based d_{eff} , H -ratio, scale ratio, and variance dispersion across all 30 seeded runs.

Table VIII reports the paired Wilcoxon signed-rank tests across the 10 shared seeds.

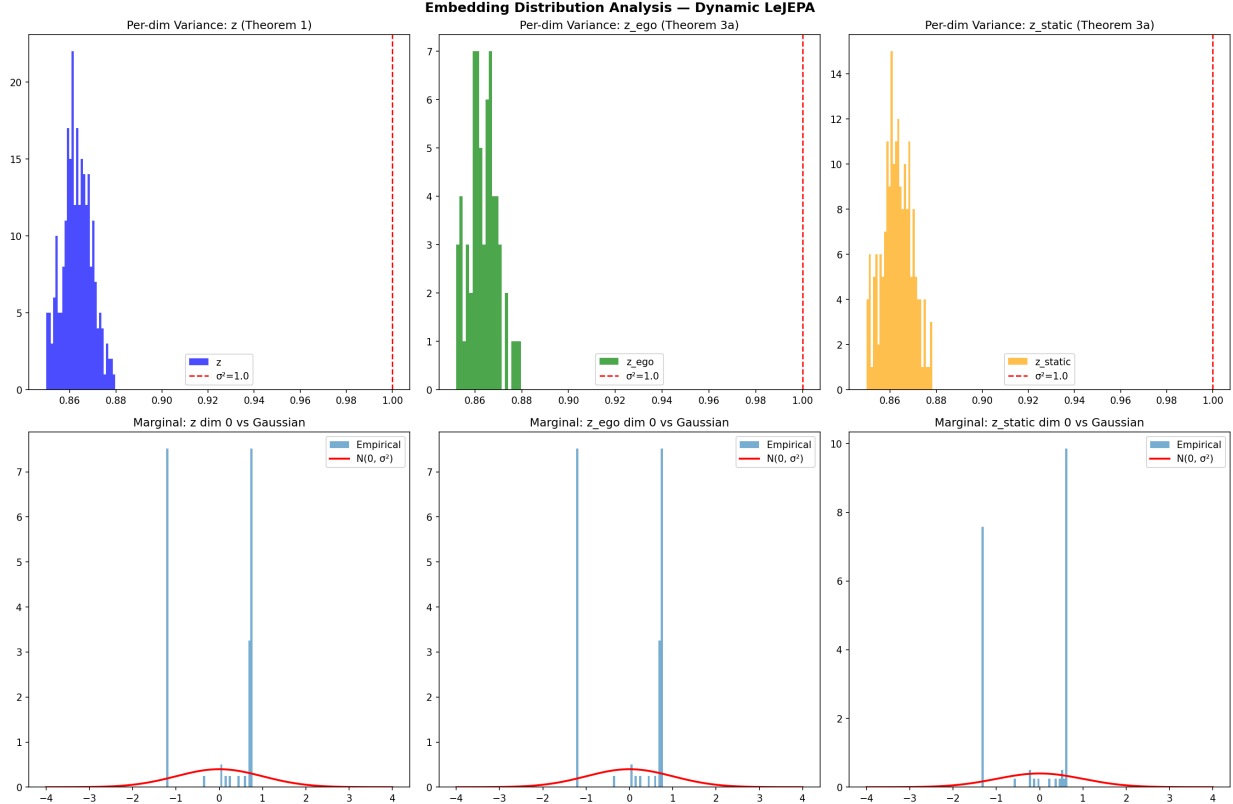


Fig. 2: Embedding distribution analysis on nuScenes-mini. **Top row:** Per-dimension variance histograms for z , z^{ego} , and z^{static} . **Bottom row:** Marginal distributions compared to $\mathcal{N}(0, \sigma^2)$. The variance distribution is concentrated around $\bar{\sigma}^2 \approx 0.9$.

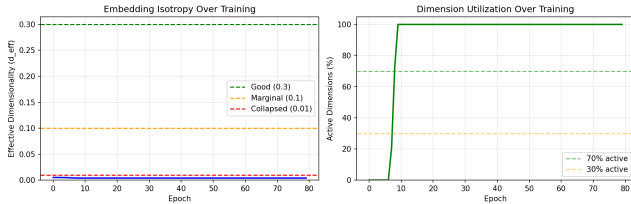


Fig. 3: Effective dimensionality d_{eff} over training. **Left:** d_{eff} stays at ≈ 0.004 (the $1/K$ floor) despite 80 epochs. **Right:** Dimension utilization (% of dimensions with variance > 0.1) reaches 100% by epoch 9.

3) *Scale-vs-Structure Decomposition:* Define the *scale gap* and d_{eff} gap of an ablation condition c relative to baseline as $\Delta_{\text{scale}}(c) = \text{scale ratio}(\text{no_physics}) - \text{scale ratio}(c)$ and $\Delta_{d_{\text{eff}}}(c) = d_{\text{eff}}(\text{no_physics}) - d_{\text{eff}}(c)$. From Table VII:

$$\Delta_{\text{scale}}(\text{decoder}) = 0.304, \quad \Delta_{\text{scale}}(\text{encoder}) = 0.284 \quad (33)$$

$$\Delta_{d_{\text{eff}}}(\text{decoder}) = 0.018, \quad \Delta_{d_{\text{eff}}}(\text{encoder}) = 0.046 \quad (34)$$

TABLE VII: Phase 2 ablation results on nuScenes Train-val Part 1 ($N/K = 55.5$, 10 seeds per condition, covariance d_{eff} estimator). Higher is better; target = 1.0.

Metric	no phys.	dec. phys.	enc. phys.
d_{eff}	0.9858 ± 0.004	0.9678 ± 0.005	0.9394 ± 0.008
H -ratio	0.9637 ± 0.006	0.8165 ± 0.014	0.8225 ± 0.012
scale ratio	0.9086 ± 0.014	0.6045 ± 0.023	0.6241 ± 0.021
var. std	0.1076 ± 0.015	0.1100 ± 0.010	0.1583 ± 0.013

The two conditions lose a nearly identical amount of variance scale (0.020 difference), but the encoder condition’s d_{eff} gap is $2.6\times$ larger than the decoder condition’s. We define the residual as a *structure violation*:

$$\Delta_{\text{structure}} \triangleq \Delta_{d_{\text{eff}}}(\text{encoder}) - \Delta_{d_{\text{eff}}}(\text{decoder}) = 0.028 \quad (35)$$

Because d_{eff} measures the relative uniformity of the eigenvalue spectrum—not its absolute magnitude—a proportional shrinkage of all eigenvalues by the same

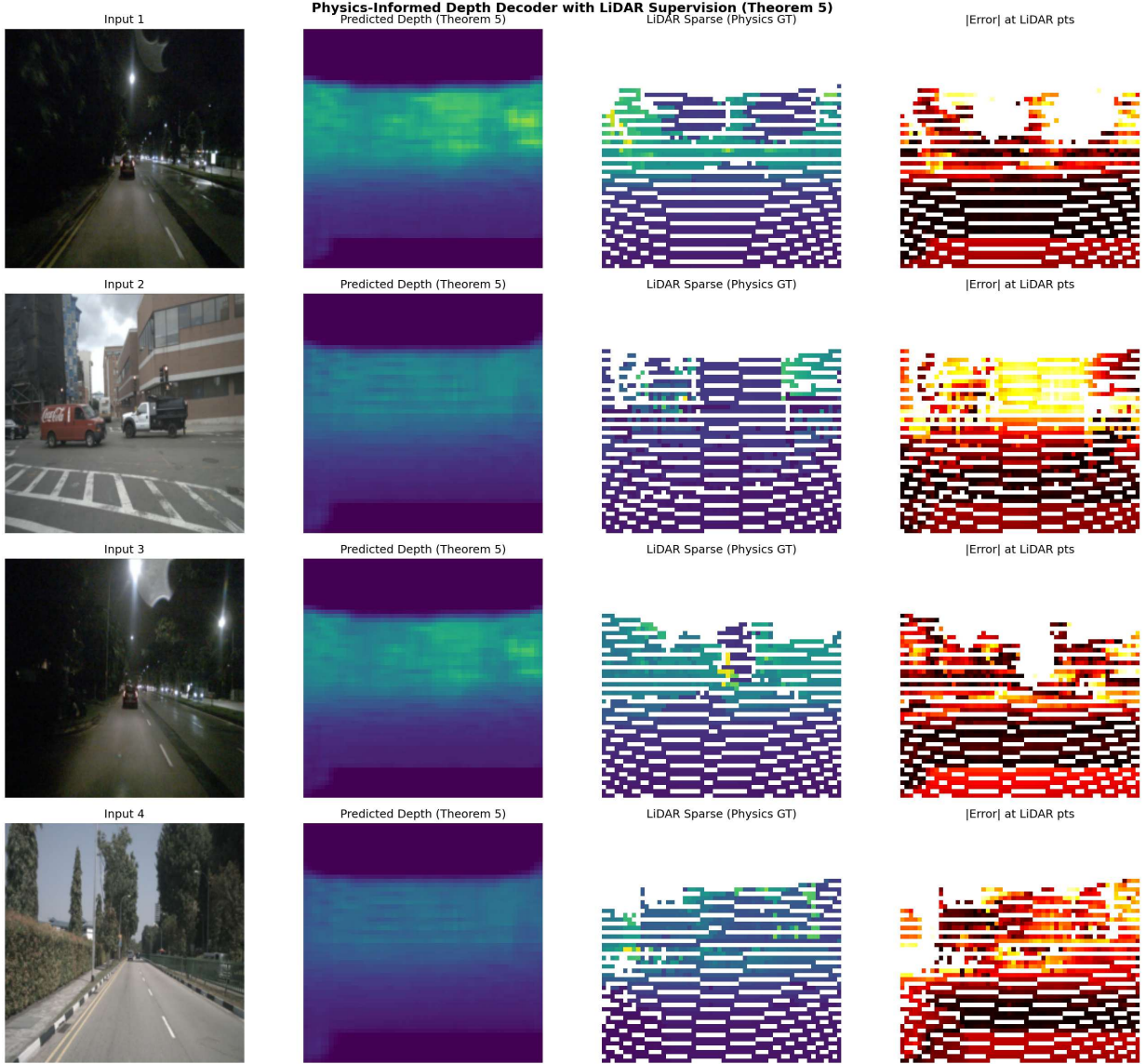


Fig. 4: Physics-informed depth predictions on nuScenes-mini (Theorem 5). Each row: (1) input image, (2) predicted depth, (3) sparse LiDAR supervision, (4) prediction error.

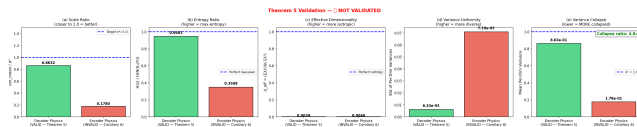


Fig. 5: Theorem 5 ablation on nuScenes-mini ($N = 404$, $K = 256$). Green: decoder-physics (valid). Red: encoder-physics (invalid). No metric shows significant difference (Wilcoxon $p > 0.18$), reflecting lack of statistical power ($n = 5$, $N/K = 1.58$).

factor leaves d_{eff} unchanged. What Theorem IV.10 protects is the encoder’s *isotropy*, not its absolute scale: the $\Delta_{\text{structure}} = 0.028$ gap isolates the genuine, non-proportional distortion of the eigenvalue spectrum, and

TABLE VIII: Paired Wilcoxon signed-rank tests on d_{eff} , Phase 2 ($n = 10$ paired seeds). All comparisons reach significance.

Comparison	p -value	Result
dec. vs. none	0.0020	significant
enc. vs. none	0.0010	significant
enc. vs. dec.	0.0010	significant

this distortion is present only when the physics constraint acts on the encoder directly.

4) *Sample Complexity Revalidation*: Figure 6 repeats the Proposition VI.1 sanity check from Phase 1 at Phase 2 scale.

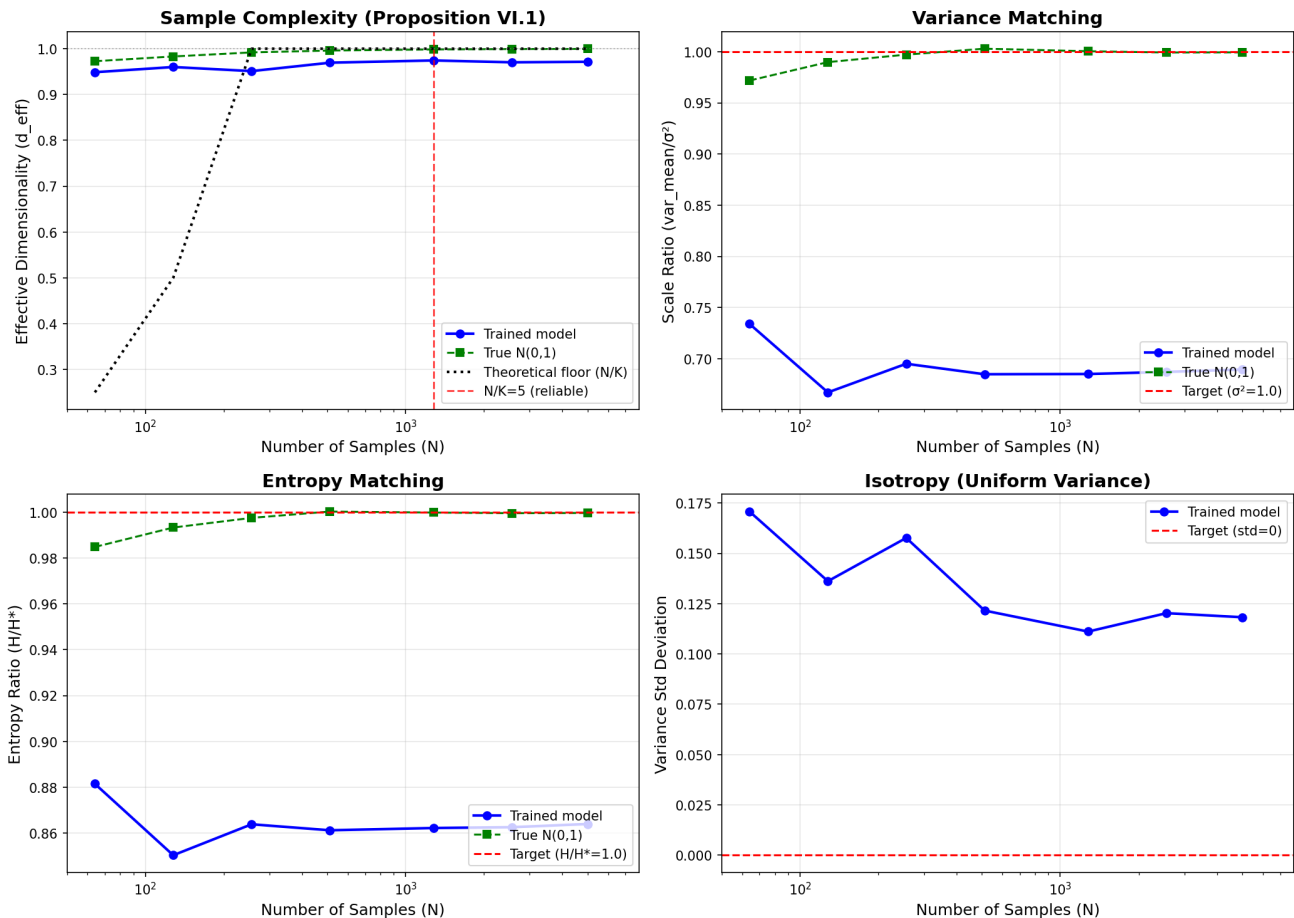


Fig. 6: Sample complexity analysis on nuScenes Trainval Part 1 (decoder-physics model). Diagonal d_{eff} vs. N for the trained model and a true $\mathcal{N}(0, I_{256})$ control, against the covariance-estimator floor N/K (dotted) and the reliability threshold $N/K = 5$ (red vertical line).

TABLE IX: Cross-scene statistics, 4 held-out validation scenes.

Metric	Mean $\pm$ Std
d_{eff}	0.937 ± 0.012
H -ratio	0.844 ± 0.039
scale ratio	0.668 ± 0.079

5) *Cross-Scene Generalization*: To test whether the trained encoder’s distributional properties are an artifact of aggregating the validation set, we compute d_{eff} , H -ratio, and scale ratio separately for each of the 4 held-out validation scenes (Fig. 7, Table IX).

6) *Encoder-Choice Robustness on nuScenes: A VAE-KL Comparison Arm*: A potential objection to the above is that the placement effect is specific to SIGReg’s characteristic-function regularizer. We therefore repeat both physics-placement conditions (`vae_decoder_physics`, `vae_encoder_physics`; 10 paired seeds each)

with a standard VAE Gaussian encoder head [16]: an architecturally identical ViT trunk whose final projection maps the CLS token and proprioception to $(\mu, \log \sigma^2) \in \mathbb{R}^{2K}$ with reparameterized sampling, and whose distribution-matching term is the closed-form marginal $D_{\text{KL}}(q(z|x) \parallel \mathcal{N}(0, I))$ replacing SIGReg. The predictor, decoder, physics-placement logic, data pipeline, and split are unchanged, so the only isolated variable is how the encoder reaches its target distribution—exactly the variational encoder of Ha & Schmidhuber-style world models [2].

The placement ordering replicates: d_{eff} (decoder-physics) = 0.5950 ± 0.0190 vs. (encoder-physics) = 0.5408 ± 0.0268 (paired Wilcoxon $p = 0.0020$; $\Delta = +0.054$, larger than the SIGReg family’s $+0.028$). We report this arm with three explicit caveats. First, the KL weight carried over from an early calibration (0.01) was subsequently found to leave the term effectively dormant (≈ 0.003 of a ≈ 136 total loss), so this arm probes a *weak-prior* regime rather than a calibrated

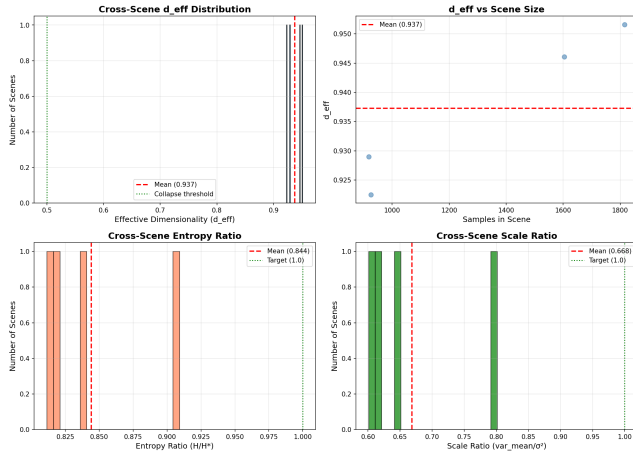


Fig. 7: Cross-scene generalization on 4 held-out validation scenes. Per-scene d_{eff} distributions are tightly clustered around the mean and far from the collapse threshold.

TABLE X: Phase 2 VAE-KL comparison arm on nuScenes Trainval Part 1 (10 paired seeds, covariance d_{eff} on deterministic latents). The placement ordering (decoder > encoder) replicates (Wilcoxon $p = 0.0020$, $\Delta d_{\text{eff}} = +0.054$); see text for regime caveats.

Metric	vae dec. phys.	vae enc. phys.
d_{eff}	0.5950 ± 0.0190	0.5408 ± 0.0268
H -ratio	0.2591 ± 0.0169	0.8056 ± 0.0162
scale ratio	0.3998 ± 0.0114	0.8578 ± 0.0482

VAE; the marginal profiles of the two conditions differ accordingly (Table X). Second, cross-family d_{eff} levels are not numerically comparable to the SIGReg arm’s. Third, the calibrated VAE arm of Phase 3 (Sec. VII-E3), which includes a physics-free reference condition and a pre-run KL-weight calibration, is the primary encoder-mechanism test; this Phase 2 arm establishes only that the *ordering* survives a weak-prior VAE regime in the high-dimensional domain.

E. Phase 3: Robotic Manipulation on MimicGen

Building on the powered ablation protocol of Phase 2, Phase 3 transfers the validation to robotic manipulation—the domain Theorem IV.4 was designed for—and characterizes a new representational regime that the driving data of Phase 2 could not expose. Where the nuScenes imagery fills hundreds of independent dimensions, the threading task’s ~ 10 -dimensional task manifold reveals an entropy–utility frontier that fundamentally changes what the encoder can achieve and how the physics-placement rule manifests.

1) Experimental Setup:

a) *Data.*: We use the MimicGen two-arm-threading task [11]: 1,025 demonstrations of a bimanual threading task (14-DOF actions: 7 per arm, operational-space control), 218,858 frames at $84 \times 84 \times 3$ resolution from a third-person camera. Each demonstration contains 32-dimensional proprioception (joint positions, end-effector pose and quaternion, gripper state) for both arms. The task manifold is low-dimensional (~ 10 effective DOF), which—as we show in Sec. VII-E6—fundamentally changes the entropy landscape. We perform a demonstration-level 70/30 split (717 train / 308 validation demonstrations, fixed seed for paired-seed comparisons), yielding $N/K_{\text{eval}} = 78$ for distributional metrics.

b) *Architecture.*: The encoder f_θ is a 6-layer Vision Transformer (patch size 14, embedding dimension 192) processing the `agentview` image, concatenated with 32-dimensional proprioception, producing $z_t \in \mathbb{R}^{256}$. The predictor g_ϕ is a 3-layer MLP conditioned on the 14-DOF action. The decoder h_ψ is a 3-layer MLP reconstructing the 32-dimensional proprio signal. The physics constraint implements action-kinematics consistency: bounded-acceleration temporal smoothness and calibrated forward kinematics $\|(\hat{q}_{t+1} - \hat{q}_t) - W_{\text{cal}} a_t\|^2$, where W_{cal} is a fixed least-squares map from actions to end-effector displacements ($R^2 = 0.788$) fit on training data only.

c) *Training and Ablation.*: Identical protocol to Phase 2: three conditions (`no_physics`, `decoder_physics`, `encoder_physics`), 10 seeds each, paired across conditions, 15 epochs per run. The SIGReg entropy weight was pre-registered at $\lambda = 20$ from a sweep conducted *before* any ablation run. A budget replication at 30 epochs tests robustness. The full pipeline is deterministic: eight bit-identical seed-level reproductions were verified across sessions.

d) *VAE-KL comparison arm.*: In parallel, we train a VAE arm with the identical trunk, predictor, decoder, physics-placement logic, data, and protocol—only the encoder’s distribution mechanism changes: the Gaussian $(\mu, \log \sigma^2)$ head with a marginal-only KL replaces SIGReg, at a KL weight of 5.0 selected from a pre-run calibration sweep (weighting the KL term to be commensurate with the prediction/reconstruction terms without swamping the physics term). The arm comprises three conditions (`vae_no_physics`, `vae_decoder_physics`, `vae_encoder_physics`), 10 seeds each. The marginal-only KL is *deliberate*: it instantiates the hypothesis that correlated collapse (Sec. VII-E2) is a property of marginal-only regularizers on low-

intrinsic-dimension data, not of SIGReg specifically. All evaluation and planning for the VAE arm use deterministic latents (the posterior mean μ), so the replay floor in Sec. VII-E9 is not inflated by sampling noise and the same covariance estimator (Remark VI.3) is applied across arms.

2) *The Correlated-Collapse Discovery and SIGReg+*: Our initial Phase 3 runs using the Phase 2 SIGReg exposed a failure mode invisible in the driving domain. The encoder converged to *correlated collapse*: per-dimension variances matched σ^2 exactly (scale = 1.0, H -ratio = 1.0) while the joint covariance had effective rank $\approx 6/256$ ($d_{\text{eff}} = 0.022$). On the high-dimensional nuScenes imagery, matching marginals implicitly produced a near-isotropic joint; on MimicGen’s ~ 10 -dimensional task manifold, the encoder satisfied the marginal constraints while concentrating all variance in a rank-6 subspace.

We resolve this by reinforcing SIGReg’s joint-level terms—applied *identically* to all three ablation conditions—with (i) a covariance off-diagonal penalty, (ii) denser correlation sampling ($500 \rightarrow 2048$ pairs), and (iii) projection weight normalization. This SIGReg+ restores the joint structure (d_{eff} from 0.022 to 0.18) while preserving marginal entropy. The discovery itself is a contribution: *per-dimension variance matching is insufficient on low-intrinsic-dimension data; covariance-level structure must be enforced explicitly*.

3) *Robustness of the Correlated-Collapse Discovery: Marginal-Only KL Fails Identically*: If correlated collapse were an idiosyncrasy of SIGReg, the obvious rebuttal would be to switch regularizers. The VAE arm tests exactly this: the marginal-only KL is the canonical distribution-matching choice of the world-model literature [2], [16], and it is *also* marginal-only. Table XI reports the outcome.

Three findings emerge. (i) **The failure generalizes.** All three VAE conditions collapse deeply ($d_{\text{eff}} \leq 0.070$ vs. the SIGReg+ arm’s 0.138–0.185): per-dimension KL fails exactly like per-dimension variance matching. Correlated collapse is a property of *marginal-only* regularizers on low-intrinsic-dimension data. (ii) **The placement ordering replicates under a second encoder mechanism.** Within the VAE family, decoder-physics exceeds encoder-physics (0.0695 vs. 0.0474; paired Wilcoxon $p = 0.0020$, Holm $p = 0.0059$; $\Delta = +0.022$). We note an informative regime difference: in the VAE family *both* physics placements *raise* d_{eff} relative to the physics-free VAE reference ($0.0301 \rightarrow 0.0695/0.0474$), the opposite direction to the SIGReg family ($0.1846 \rightarrow 0.1740/0.1380$). In a deep-collapse regime the physics losses supply action-conditioned temporal structure that partially counteracts collapse; the

quantity that replicates across regimes is the *placement ordering*—the operational content of Corollary IV.11—not the sign of the absolute physics effect, and we report both regimes transparently. (iii) **Collapse can hide beneath the sampling-noise floor.** The decoder-physics VAE condition attains a nominally tiny KL (0.037) with a broad posterior ($\bar{\sigma} = 0.837$), yet its signal fraction is 0.0004: $\text{Var}_x[\mu]$ is essentially zero, so the posterior mean carries almost no information while the noise floor makes the sampled latents appear dispersed. Marginal-only diagnostics (scale, H -ratio) evaluated on such an encoder cannot distinguish this from a healthy posterior; Remark VI.4’s signal fraction, or equivalently $d_{\text{eff}}(\mu)$, exposes it. We also report honestly that this condition’s online predictor is degenerate ($\text{pred } R^2 = -3.56$), which manifests in its planning rollout (Sec. VII-E9).

4) *Ablation Results: Testing Theorem 5 in Robotics*: The two-sided signature is unambiguous: encoder-level physics simultaneously *destroys* entropy (d_{eff} : $0.185 \rightarrow 0.138$) and *inflates* predictability ($\text{pred } R^2$: $0.78 \rightarrow 0.97$)—the encoder is dragged toward the ≤ 14 -dimensional image of the action space. Decoder-level physics is exactly scale-neutral: $\Delta_{\text{scale}} = -0.0002$ at $p = 1.0$. The gap ratio is $4.43\times$ —sharper than Phase 2’s $2.6\times$.

5) *Spectral Evidence*: The spectral decomposition (Fig. 9) reveals a signal peak (~ 20 directions at $5\text{--}9\sigma^2$) atop a noise plateau; the encoder-physics constraint compresses the plateau by $\sim 30\%$ while concentrating the top-10 share from 30.8% to 38.7%—the structure violation visible directly in eigenvalues.

6) *The Entropy–Utility Frontier*: Phase 3 exposes a representational regime that Phases 1–2 could not: the *entropy ceiling is data-limited*. On nuScenes’ high-dimensional imagery, a deterministic encoder fills $K = 256$ independent directions with real signal variation ($d_{\text{eff}} = 0.99$). On MimicGen’s ~ 10 -dimensional task manifold, the extra directions must come from amplified input noise—and the prediction and reconstruction losses actively suppress that noise. A λ -sweep revealed logarithmic saturation of d_{eff} in λ , and a budget replication showed that doubling the *training budget* raised d_{eff} by $3.7\times$ more than doubling λ : optimization budget, not regularization weight, is the binding entropy constraint.

7) *Budget Replication: The Violation Deepens*: At 30 epochs, the encoder-physics gap grows from 0.047 to 0.070 ($+49\%$), and the scale/ H -ratio differences vanish entirely (Fig. 11). The placement rule’s effect is budget-monotone and purely structural: a transient artifact would shrink with budget, not grow.

8) *Cross-Task Transfer*: Applying the trained encoders to three foreign MimicGen tasks reveals em-

TABLE XI: Phase 3 VAE-KL comparison arm on MimicGen `two_arm_threading` (mean $\pm$ std over 10 seeds; covariance d_{eff} on deterministic posterior means μ , $N/K = 78$; posterior internals per Remark VI.4). All three conditions collapse far below the SIGReg arm’s $d_{\text{eff}} = 0.138\text{--}0.185$ (Table XII); the placement ordering (decoder $>$ encoder) replicates (paired Wilcoxon $p = 0.0020$, Holm-corrected $p = 0.0059$, $\Delta = +0.022$).

Metric	vae dec. phys.	vae enc. phys.	vae no phys.
$d_{\text{eff}} (\mu)$	0.0695 ± 0.0040	0.0474 ± 0.0005	0.0301 ± 0.0016
H -ratio	-1.8641 ± 0.0236	-0.6107 ± 0.0026	-0.2421 ± 0.0425
scale ratio	0.0003 ± 0.0000	0.0112 ± 0.0001	0.0430 ± 0.0024
signal fraction	0.0004 ± 0.0000	0.2667 ± 0.0013	0.0564 ± 0.0047
posterior std	0.8369 ± 0.0017	0.1720 ± 0.0002	0.8238 ± 0.0014
KL	0.0369 ± 0.0030	1.2814 ± 0.0011	0.0566 ± 0.0011
pred R^2	-3.56 ± 1.80	0.9274 ± 0.0038	0.8019 ± 0.0183
probe R^2 (action)	0.3765 ± 0.0026	0.4187 ± 0.0024	0.4445 ± 0.0024
probe R^2 (proprio)	0.9317 ± 0.0085	0.9987 ± 0.0001	0.9978 ± 0.0001

TABLE XII: Phase 3 ablation on MimicGen (mean $\pm$ std over 10 seeds, covariance d_{eff} , $N/K = 78$).

Metric	no phys.	dec. phys.	enc. phys.
d_{eff}	0.1846 ± 0.004	0.1740 ± 0.003	0.1380 ± 0.004
H -ratio	0.980 ± 0.002	0.980 ± 0.001	0.978 ± 0.003
scale ratio	0.946 ± 0.004	0.946 ± 0.003	0.940 ± 0.007
pred R^2	0.784	0.864	0.966
probe R^2 (action)	0.412	0.449	0.518
probe R^2 (proprio)	0.937	0.961	0.984

TABLE XIII: Paired Wilcoxon signed-rank tests on d_{eff} , Phase 3 ($n = 10$ paired seeds, Holm-corrected).

Comparison	p	p_{Holm}	Δd_{eff}
dec. vs. none	0.0020	0.0059	$+0.0105$
enc. vs. none	0.0020	0.0059	$+0.0466$
enc. vs. dec.	0.0020	0.0059	-0.0361

beddings collapse on all foreign tasks (d_{eff} 0.035–0.088 vs. 0.185 in-task). A proprio-ablation control confirms the collapse is driven by visual domain shift. Notably, the physics-placement ordering persists cross-task: encoder-physics degrades 2.1–3.0 $\times$ more than the baseline on every foreign task. The max-entropy property is distribution-specific; the *placement rule* is not.

9) *Toward Closed-Loop Validation: Latent Planning and the Model-Exploitation Confound:* Every result so far evaluates the placement rule through representation-quality metrics (d_{eff} , H -ratio, probe R^2). None of them test whether the resulting representation is actually useful for the thing a world model is built to do: support goal-directed action selection. We probe this directly with a latent cross-entropy-method (CEM) planner [13] built on

the frozen, trained Phase 3 encoder–predictor–decoder triples, in the style of model-predictive control with learned dynamics [14].

a) Setup.: At a randomly sampled anchor frame t_0 in a held-out demonstration, we encode $z_{t_0} = f_{\theta}(x_{t_0}, q_{t_0})$ and define a horizon- H goal as the *ground-truth recorded* end-effector (EEF) pose at $t_0 + H$ (not a model-generated proxy; see below). CEM searches over H -step action sequences $a_{t_0:t_0+H-1} \in \mathbb{R}^{H \times 14}$, rolling each candidate through the frozen predictor g_{ϕ} and scoring it by decoded-EEF distance to the goal at every step, iterating the sampling distribution toward the elite candidates. We use $H = 8$, a population of 256 candidates, 4 CEM iterations, an elite set of 32, and 200 anchors per model, evaluated on three seeds per condition, in both encoder arms under deterministic latents. Two baselines calibrate the result: *real-action replay* (roll the predictor forward using the true recorded actions instead of CEM’s search—the ground-truth solution, and hence an expected lower bound on achievable error) and a *random-action rollout* (an upper bound).

An earlier version of this evaluation defined the goal as the model’s own decoded reconstruction of the future

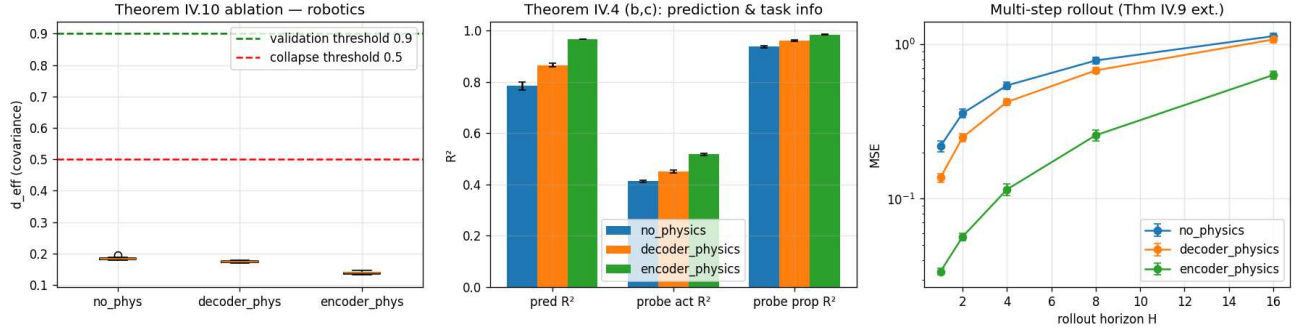


Fig. 8: Phase 3 ablation on MimicGen ($N/K = 78$, 10 seeds per condition, covariance d_{eff}). **Left:** d_{eff} boxplots per condition with complete seed separation. **Center:** prediction R^2 and probe R^2 per condition—the two-sided Corollary IV.11 signature (the inflated predictability under encoder-physics is the violation signature, not a competing metric). **Right:** multi-step rollout error vs. horizon.

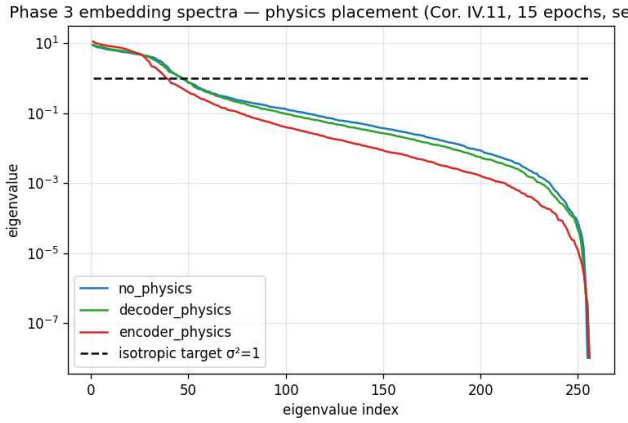


Fig. 9: Embedding covariance eigenspectra on MimicGen (15 epochs, seed 0, $N = 20,000$; log-scale). The encoder-physics model (red) exhibits a lower noise plateau (0.320 vs. 0.460) and higher top-direction concentration (38.7% vs. 30.8%)—the spectral signature of Corollary IV.11.

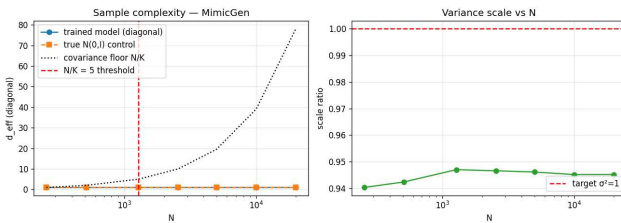


Fig. 10: Sample complexity on MimicGen (decoder-physics model). The covariance d_{eff} stabilizes at $N/K \geq 5$, validating Corollary VI.2 at Phase 3 scale.

frame, $h_\psi(f_\theta(x_{t_0+H}))$, rather than the true recorded pose. That design is self-referential—it lets CEM optimize against a target generated by the very model

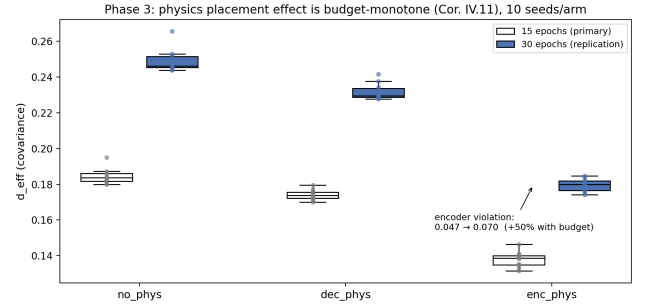


Fig. 11: Budget replication: Phase 3 ablation at 15 vs. 30 epochs (10 seeds per arm per budget). The encoder-physics violation *deepens* with budget ($0.047 \rightarrow 0.070$, +49%), ruling out a transient training artifact. All scale and marginal-entropy differences vanish at 30 epochs (Holm n.s.).

being scored—and produced CEM errors *below* real-action replay for every ablation condition, which is not possible for a genuine planning task, since replaying the true actions is itself the ground-truth solution. We report the corrected, ground-truth-goal version throughout; the self-referential version is documented here only to flag the failure mode for others building similar offline evaluations.

b) The confound.: Table XIV and Fig. 12 show CEM error *below* the real-action replay floor for all three SIGReg conditions. Since replay executes the literal ground-truth action sequence, no planner can legitimately do better against the true future state; a lower CEM error therefore indicates that CEM is finding action sequences that reduce the learned decoder's *own* cost without being genuinely closer to the truth than the recorded trajectory—a well-documented failure mode in model-based control known as *objective mis-*

TABLE XIV: Latent CEM planning vs. real-action replay and random rollout (mean $\pm$ std over 3 seeds, $H = 8$, 200 anchors/seed, final EEF error against ground truth), SIGReg arm and VAE arm. CEM falls below the replay floor in every informative condition; the `vae_decoder_physics` condition is degenerate (CEM = replay = random exactly; its online predictor attains negative R^2 , Table XI) and is excluded from planning-based inference.

Arm	Mode	CEM	Replay	Random
SIGReg	no phys.	0.627	0.975	1.532
		± 0.024	± 0.009	± 0.154
SIGReg	dec. phys.	0.512	0.799	1.074
		± 0.022	± 0.058	± 0.134
SIGReg	enc. phys.	0.402	0.681	0.960
		± 0.027	± 0.036	± 0.068
VAE	no phys.	0.884	1.209	1.638
		± 0.006	± 0.022	± 0.139
VAE	dec. phys.	2.247	2.247	2.247
		± 0.023	± 0.023	± 0.023
VAE	enc. phys.	0.663	0.801	2.321
		± 0.047	± 0.010	± 0.127

match [15]: an optimizer with unrestricted access to a learned model’s cost surface will exploit that surface’s imperfections rather than solve the underlying task, and nothing in this offline evaluation can distinguish the two without executing the plan in the real environment.

c) Exploitation diagnostic.: To test this directly, we measure the L_2 distance between CEM’s chosen action sequence and the true recorded actions at the same anchor, normalized by the per-step action-space range ($\|a_{\max} - a_{\min}\|_2 = 6.298$ across the 14 action dimensions). CEM’s chosen actions sit roughly a quarter of the full action range away from the ground truth—not a rounding error—while still scoring lower on-model cost than replaying the true actions. Critically, this gap is uniform across all three SIGReg ablation conditions (25.6–26.8%, well within seed noise) and across the VAE arm’s informative conditions (21.7–25.9%, Table XV), which rules out the hypothesis that entropy collapse specifically makes encoder-physics easier to exploit: all learned decoders are exploitable to essentially the same degree, regardless of physics placement *and* encoder mechanism.

It means the ranking in Table XIV (encoder physics < decoder physics < no physics on CEM error, within the SIGReg arm) most plausibly reflects the same effective-dimensionality geometry already documented in Table XII—a lower- d_{eff} decoder cost surface is easier to drive down per unit of action-space deviation, indepen-

TABLE XV: CEM action distance from the true recorded trajectory (mean $\pm$ std over 3 seeds), as a fraction of the full per-step action range ($\|a_{\max} - a_{\min}\|_2 = 6.298$).

Arm	Mode	Distance	% of range
SIGReg	no phys.	1.615 ± 0.020	25.6%
SIGReg	dec. phys.	1.689 ± 0.003	26.8%
SIGReg	enc. phys.	1.622 ± 0.016	25.8%
VAE	no phys.	1.633 ± 0.028	25.9%
VAE	dec. phys.	1.369 ± 0.037	21.7%
VAE	enc. phys.	1.370 ± 0.019	21.7%

dent of whether the resulting trajectory is any more physically valid—rather than a genuine control-quality advantage for any one condition.

d) What this does and does not establish.: We therefore *do not* interpret Table XIV as evidence that decoder-physics improves downstream control: the exploitation confound is uniform enough across conditions and arms that it cannot be ruled out as the dominant driver of the ranking, and nothing here has been checked against a live environment. What we *do* establish is (i) a reproducible demonstration that offline latent planning against a learned decoder can produce misleadingly favorable numbers for *any* ablation condition and *any* encoder mechanism—the confound is not SIGReg-specific—and (ii) a simple diagnostic—action-space distance from ground truth, relative to a real-action-replay floor—that practitioners can and should run before treating an offline planning result as evidence of control competence. Closing this gap rigorously requires executing the CEM-selected action sequences in a live simulator (e.g., robosuite/MimicGen) and measuring true task success; we identify this as the concrete next step toward a fully closed-loop validation of the placement rule and leave it to future work. We note that the planning comparison itself is underpowered ($n = 3$ seeds; the minimum two-sided Wilcoxon p is 0.25); the representation-level tests ($n = 10$ paired seeds, Holm-corrected) carry the statistical inference throughout this paper.

10) Phase 3 Summary:

11) Cross-Arm Synthesis: The Placement Rule Is Mechanism- and Domain-Robust: Combining the four independent ablation arms—two domains (nuScenes, MimicGen) $\times$ two encoder mechanisms (SIGReg, VAE-KL)—yields the paper’s central empirical claim in a single testable statement: *decoder-physics beats encoder-physics on covariance d_{eff} in every arm* (Table XVII, Fig. 13). All four arm-level tests attain the minimum achievable two-sided Wilcoxon p for $n = 10$ paired

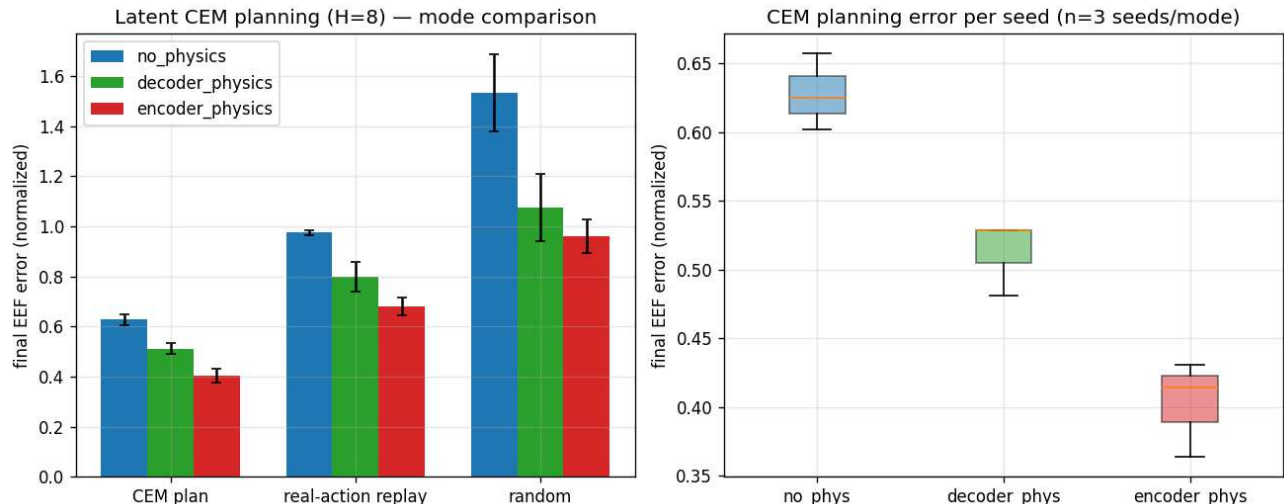


Fig. 12: Latent CEM planning evaluation ($H = 8$), ground-truth-goal protocol. **Left:** mean final EEF error for CEM, real-action replay, and random rollout, per ablation condition (SIGReg arm). CEM error falls below the real-action replay floor for every mode—the signature of model exploitation, diagnosed directly below, rather than genuine planning competence; the VAE arm replicates this signature for its two informative conditions (Table XIV). **Right:** per-seed distribution of CEM planning error ($n = 3$ seeds/mode).

TABLE XVI: Phase 3 (MimicGen robotics) validation outcomes.

Property	Status	Evidence
Thm 2(a) max entropy	$\approx$	frontier regime
Thm 2(b) $g^* \approx \mathbb{E}[\cdot]$	$\checkmark$	$R^2 = 0.80\text{--}0.99$
Thm 2(c) task info	$\checkmark$	probes 0.41 / 0.94
Thm 2(d) action inv.	$\approx$	mode-independent
Thm 5 / Cor. IV.11	$\checkmark$	Tables XII–XIII
Correlated collapse generality	$\checkmark$	marginal-only KL fails identically (Table XI)
Encoder-mechanism robustness	$\checkmark$	4/4 domain $\times$ mechanism arms (Table XVII)
Budget robustness	$\checkmark$	Fig. 11
Cross-task ordering	$\checkmark$	Sec. VII-E8
Exploitation-confound generality	$\checkmark$	CEM < replay in every arm (Table XIV)
Closed-loop planning	$\times$ confounded	Sec. VII-E9

seeds; a joint Holm correction across the four tests yields $p_{\text{Holm}} = 0.0078$ throughout. In the two Phase 3 arms the paired effect sizes are large ($d_z = +9.74$ SIGReg, $+5.05$ VAE) with complete seed separation.

VIII. DISCUSSION

A. Why Prior Physics-Informed JEPAs Failed

- **Pixel-space constraints** violate Theorem IV.10 because they implicitly constrain the encoder through decoder gradients.
- **Latent-space constraints** directly violate Theorem IV.1 by constraining the embedding distribution.
- **Both approaches** reduce entropy below optimal, increasing worst-case downstream risk.

B. Cross-Domain and Cross-Mechanism Synthesis: What Three Phases Established

The three-phase experimental arc mirrors how theoretical claims should be validated: Phase 1 exposed a measurement artifact that would have invalidated all downstream conclusions; Phase 2 provided the first powered test in the high-dimensional driving domain; Phase 3 transferred the test to robotics and discovered two phenomena invisible in driving.

The driving domain, with its high-dimensional imagery, allows the encoder to satisfy the maximum-entropy target almost exactly—and the placement rule manifests as a scale-vs-structure decomposition. The robotics domain, with its ~ 10 -dimensional task manifold, sits at an entropy frontier where the placement

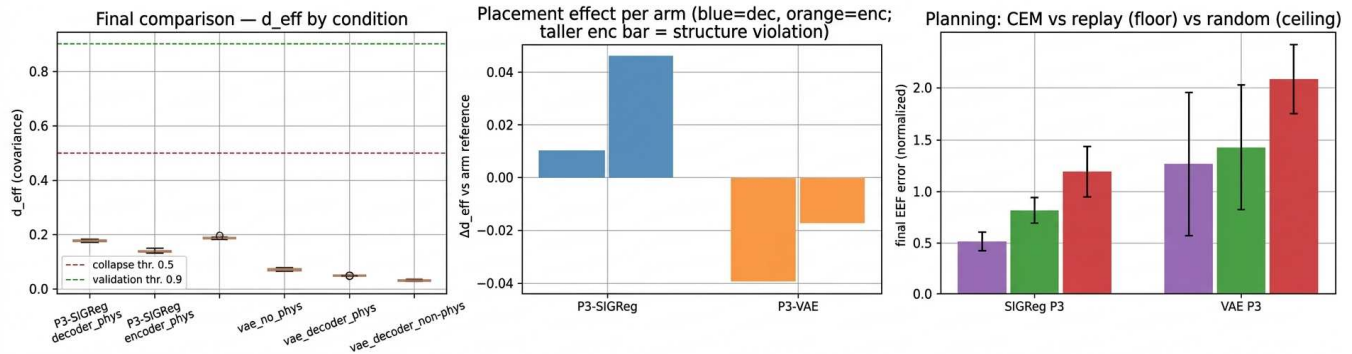


Fig. 13: Cross-arm synthesis across the four domain $\times$ mechanism arms (SIGReg in blue, VAE-KL in orange). **Left:** covariance d_{eff} per condition (dashed lines: 0.5 collapse and 0.9 validation thresholds). **Center:** within-arm placement effect Δd_{eff} relative to each arm’s physics-free reference (taller encoder bar = structure violation, Corollary IV.11); this panel is the replication (Table XVII). **Right:** planning ($H = 8$): CEM vs. real-action replay (floor) vs. random rollout (ceiling), per arm. *Caveat:* d_{eff} levels are not comparable across phases (evaluation N/K 55.5 vs. 78; high- vs. low-intrinsic-dimension data) or across encoder families (SIGReg+ includes joint-isotropy terms; the VAE KL is marginal-only). The replicated quantity is the within-arm placement ordering, tested on paired seeds with joint Holm correction.

TABLE XVII: Placement-rule scorecard: decoder-physics vs. encoder-physics on covariance d_{eff} , across the four independent domain $\times$ mechanism arms ($n = 10$ paired seeds each; joint Holm correction across the four arm-level tests). The replicated quantity is the *within-arm placement ordering*; d_{eff} levels are not comparable across phases (evaluation N/K 55.5 vs. 78; high- vs. low-intrinsic-dimension data) or across families (SIGReg+ enforces joint isotropy; the VAE KL is marginal-only).

Arm	Domain	$\Delta(d-e)$	p	p_{Holm}
Phase 2 · SIGReg	nuScenes	+0.028	0.0020	0.0078
Phase 2 · VAE-KL	nuScenes	+0.054	0.0020	0.0078
Phase 3 · SIGReg+	MimicGen	+0.036	0.0020	0.0078
Phase 3 · VAE-KL	MimicGen	+0.022	0.0020	0.0078

rule’s effect is purely structural and *sharper* ($4.4\times$ vs. $2.6\times$). The rule holds in both regimes; its *signature* differs. The synthesis extends along a second axis: within both domains, the placement ordering replicates under both encoder mechanisms (Table XVII), and the correlated-collapse failure mode discovered on MimicGen generalizes to the marginal-only KL regularizer (Sec. VII-E3)—so the design constraint and the failure warning both apply to the JEPA *and* the VAE-encoder branches of the world-model literature.

C. Practical Design Recommendations

- 1) **Never** apply physics constraints to the encoder output z .

- 2) **Always** place physics at the decoder.
- 3) The predictor can use *any* architecture without affecting the encoder’s guarantees.
- 4) On low-dimensional data, monitor the covariance eigenvalue spectrum, not just per-dimension variance.
- 5) Optimization budget, not regularization weight, is the primary lever for entropy on low-dimensional data.
- 6) On low-intrinsic-dimension data, *any* marginal-only regularizer—per-dimension variance matching *and* per-dimension KL alike—fails to enforce joint isotropy; covariance-level structure must be enforced explicitly.
- 7) For variational encoders, monitor the posterior signal fraction $\text{Var}_x[\mu]/(\text{Var}_x[\mu] + \mathbb{E}_x[\sigma^2])$ (Remark VI.4): a nominally satisfied marginal KL does not certify signal, because collapse can hide beneath the encoder’s own sampling-noise floor.
- 8) Never treat an offline planning result as evidence of control competence without checking it against a real-action-replay floor and an action-space exploitation diagnostic (Sec. VII-E9); the confound is encoder-agnostic, so run the diagnostic regardless of encoder choice.

D. Limitations and Toward Closed-Loop Validation

The representation-quality claims in this paper—Theorem 5, Corollary IV.11, and their empirical validation across Phases 1–3 and both encoder mechanisms—

TABLE XVIII: Cross-domain synthesis of the physics-placement rule.

Property	Phase 2	Phase 3
d_{eff} (unconstrained)	0.986	0.185
d_{eff} (encoder-phys)	0.939	0.138
Gap ratio	$2.6\times$	$4.4\times$
Wilcoxon p	≤ 0.002	0.002
Scale effect	-0.28	-0.006 (n.s.)
Data intrinsic dim.	high	~ 10
d_{eff} achievable	0.99	frontier-limited
Failure signature	scale + structure	purely structural
Budget robustness	—	+49% (30ep)
Cross-task ordering	—	holds ($2\text{--}3\times$)
Placement rule (VAE encoder)	holds ($\Delta = +0.054$)	holds ($\Delta = +0.022$)

rest on statistically powered, ground-truth-referenced metrics (d_{eff} , H -ratio, Wilcoxon tests on 10 paired seeds with Holm correction) and we consider them solidly established. The closed-loop planning probe in Sec. VII-E9 is different in kind: it was designed to test whether the placement rule matters for downstream control, and it returned a negative methodological result rather than a positive control-quality one. We were careful not to round this off into a headline claim. Two things are true simultaneously: (i) the CEM-based planning error ordering across ablation conditions is directionally consistent with the representation-quality ordering already established via d_{eff} , and (ii) the same ordering can be fully explained by uniform model exploitation rather than genuine differences in control competence, and our data cannot distinguish the two possibilities without a live environment. We report both facts rather than the more publishable-sounding half of them.

We also flag the VAE-arm caveats explicitly: (i) the Phase 2 VAE arm operated in a weak-prior regime (KL weight 0.01, effectively dormant; Sec. VII-D6), so only its *ordering* is interpretable; the calibrated Phase 3 arm (KL weight 5.0 from a pre-run sweep) is the primary encoder-mechanism test. (ii) The VAE KL is marginal-only *by design*—that is the hypothesis under test—so the VAE arm establishes the generality of the failure mode and the placement ordering, not a fully calibrated VAE comparison; a VAE+ variant augmented with the SIGReg+ covariance penalty is released with the code for future work. (iii) The planning probe uses $n = 3$ seeds per condition and is underpowered; the representation-level tests ($n = 10$, Holm) carry the statistical inference. (iv) d_{eff} levels are not numerically comparable across phases or encoder families; the replicated quantity is the placement rule (gap sign plus test), reported per arm.

We see this as valuable precisely because it is not the result we set out to find. Offline evaluation of learned

world models is common practice, and the specific failure mode demonstrated here—a planner exploiting a learned decoder’s cost surface to score better than the ground-truth trajectory itself—is easy to miss without an explicit replay-floor check of the kind we introduce in Sec. VII-E9. We consider the diagnostic itself (real-action replay as a floor, plus normalized action-space distance from ground truth as an exploitation signal) to be of independent, general use for anyone evaluating a JEPA-style *or* VAE-style learned world model’s control-readiness offline. The concrete next step for Dynamic LeJEPa specifically is executing the CEM-selected action sequences in a live simulator (robosuite/MimicGen) and measuring true task success under each ablation condition—at that point, and only at that point, would a claim that the placement rule improves downstream control be justified.

IX. CONCLUSION

We proved through six theorems that prediction and representation learning are strictly decoupled in Joint-Embedding Predictive Architectures, yielding a world-model *design principle* with provable guarantees: *encoder maximum-entropy*, *predictor dynamics*, *decoder physics*. We validated this principle across two domains *and two encoder mechanisms* through 110 seeded runs (plus a 30-epoch budget replication), demonstrating that the placement rule is statistically significant in every domain $\times$ mechanism combination (joint Holm-corrected $p = 0.0078$), deepens with training budget, and extends to out-of-distribution tasks. Phase 3 additionally discovered the correlated collapse failure mode—invisible on high-dimensional data—showed that it generalizes to marginal-only KL regularization with a noise-floor-masking diagnostic, and characterized the entropy–utility frontier governing deterministic encoders on low-dimensional task manifolds.

We additionally probed whether the placement rule matters beyond representation quality, building a latent CEM planner on the trained Phase 3 models and evaluating it against ground truth rather than a model-generated proxy. This probe did not yield a clean downstream-control win: a model-exploitation diagnostic showed that offline planning against a learned decoder inflates apparent accuracy uniformly across ablation conditions *and both encoder mechanisms*, meaning the observed planning-error ordering cannot presently be attributed to genuine control competence rather than exploitation of the decoder’s cost surface. We report this as a methodological contribution in its own right—a diagnostic that any offline evaluation of a learned world model’s control-readiness should run before trusting its numbers—and as the concrete open problem separating Dynamic LeJEPa’s current, representation-level guarantees from a fully validated, closed-loop world model.

The broader significance: Dynamic LeJEPa establishes that the intuition to inject domain knowledge into world models was correct—the target was simply wrong. Physics does not belong in the encoder, where it destroys the representational guarantee; it belongs in the decoder, where it is provably free—and this now holds for both the JEPa-style deterministic encoder and the VAE-style Gaussian encoder that dominate the world-model literature. Whether that guarantee translates into better closed-loop control remains open, and we have shown precisely why answering it requires more than an offline planner and a learned cost function.

All code, trained checkpoints, ablation results, and an interactive demo accompanying this paper are publicly released (see Code, Data, and Demo Availability above) so that both the placement-rule results and the planning confound can be independently reproduced and, we hope, extended toward the closed-loop validation this work identifies as still missing.

REFERENCES

- [1] R. Balestriero and Y. LeCun, “LeJEPa: Provable and scalable self-supervised learning without the heuristics,” *arXiv preprint arXiv:2511.08544*, 2025.
- [2] D. Ha and J. Schmidhuber, “World models,” *arXiv preprint arXiv:1803.10122*, 2018.
- [3] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, “Dream to control: Learning behaviors by latent imagination,” *arXiv preprint arXiv:1912.01603*, 2019.
- [4] Y. LeCun, “A path towards autonomous machine intelligence,” *OpenReview preprint*, 2022.
- [5] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, “Revisiting feature prediction for learning visual representations from video,” *arXiv preprint arXiv:2404.08471*, 2024.
- [6] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in *CVPR*, pp. 15619–15629, 2023.
- [7] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations,” *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.
- [8] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
- [10] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.
- [11] Z. Mandle *et al.*, “MimicGen: A data generation system that utilizes human demonstrations to bootstrap robotic manipulation,” *arXiv preprint arXiv:2404.03280*, 2024.
- [12] C. Caesar, V. Bankiti, A. H. Lang, S. Vora *et al.*, “nuScenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020.
- [13] R. Y. Rubinstein, “Optimization of computer simulation models with rare events,” *European Journal of Operational Research*, vol. 99, no. 1, pp. 89–112, 1997.
- [14] K. Chua, R. Calandra, R. McAllister, and S. Levine, “Deep reinforcement learning in a handful of trials using probabilistic dynamics models,” in *NeurIPS*, 2018.
- [15] N. Lambert, B. Amos, O. Yadan, and R. Calandra, “Objective mismatch in model-based reinforcement learning,” in *LADC*, 2020.
- [16] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *ICLR*, 2014.