WHY CAN LATENT DYNAMICS MODELS ADAPT TO VISUAL SHIFTS FROM A FEW EXAMPLES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent dynamics models built on pretrained foundation representations often adapt to dynamics-preserving visual shifts from only a few target-domain episodes. We explain this sample efficiency through the structure of the required functional correction. Much of this correction is globally shared across inputs and both statistically and optimization-wise accessible through low cross-episode latent variability and low-dimensional parameter updates. The state-dependent remainder is learned more gradually but concentrates in a substantially lower-dimensional subspace than under source-specific representations. Transport analysis link these structures to the interaction between representation geometry and pretrained dynamics. These findings show that foundation representations facilitate adaptation not merely through transferable features, but by structuring the functional change required to reuse source-trained dynamics across domains.

1 INTRODUCTION

Learning predictive dynamics enables control of complex systems by allowing agents to anticipate future states and plan actions (Hafner et al., 2025; Hansen et al., 2024; Maes et al., 2026). However, it is well known that learning robust long-horizon models requires substantial interaction data, as even small transition errors can compound over successive steps. The challenge lies not only in learning the transition function, but also in obtaining a representation that makes the underlying dynamics predictable from available experience. Pretrained on diverse web-scale data, foundation representations offer a promising route by decoupling representation learning from dynamics fitting. Recent world models built on these representations have demonstrated strong data efficiency and zero-shot transfer across diverse control settings (Zhou et al., 2024; Baldassarre et al., 2025), suggesting that foundation representations provide a favorable latent geometry for dynamics learning.

But foundation representations do not eliminate the need for target-domain dynamics fine-tuning. Broad statistical regularities acquired through foundation pretraining support meaningful zero-shot transfer in unfamiliar domains, well beyond trivial baselines, but performance often remains below the source-domain level. Even when source and target domains share the same physical transition law, its latent expression can differ, leaving a prediction mismatch that requires adaptation. What is striking is that, with a frozen foundation representation, continued dynamics training from a well-trained source model can close much of the performance gap using very little target-domain experience (Parisi et al., 2022; Oquab et al., 2023; Lanier et al., 2025; Wang et al., 2026).

This raises the central question of this work: *why can foundation-based latent dynamics models learn a nontrivial target-domain correction from only a few episodes?* The challenge is to incorporate this correction while preserving the predictive structure encoded by the source-trained model. At its core, this is a question on how the representation expresses the visual domain shift. Invariance offers one limiting case, mapping corresponding physical states to identical latents across domains. Under perfect invariance and exact source dynamics, the latent transition law remains unchanged and no adaptation is required. Invariance, however, is stronger than necessary. Source and target latent geometries may differ pointwise yet retain a regular relationship across states.

In this work, we show that foundation representations enable nontrivial latent-dynamics adaptation from few episodes by structuring the required functional correction, i.e., the change in source-model predictions needed to recover target-domain dynamics. Much of this correction is globally shared across inputs and realizable through low-complexity parameter updates, while its few-shot statistical

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

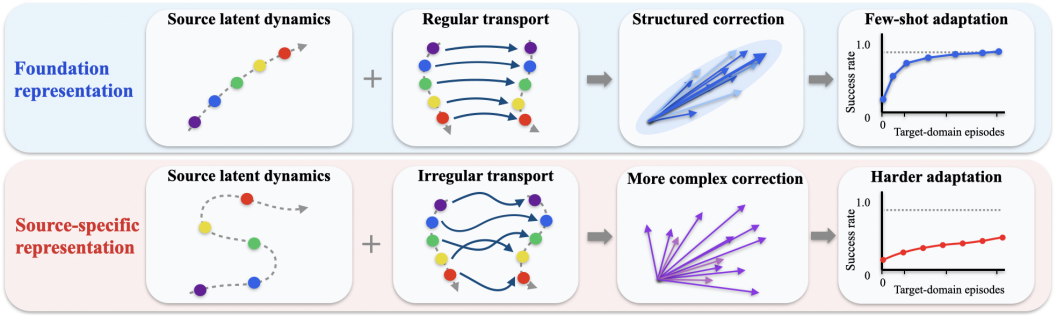


Figure 1: Illustration of how latent structure shapes few-shot target-domain adaptation of latent dynamics models. The correction required to recover target-domain latent dynamics is determined by the interaction between source-trained dynamics and cross-domain transport. Under a foundation representation, both can exhibit regular structure, and their interaction can yield a structured correction learnable from only a few target-domain episodes, enabling the performance to rapidly approach the source-domain level. This work focuses on characterizing that required correction. In contrast, a source-specific representation may fit source-domain dynamics with high accuracy while leaving unseen cross-domain transport unconstrained. Their interaction can then induce a complex correction, making target-domain adaptation substantially harder.

recovery follows from a representation-induced correction geometry that remains consistent across target episodes. The state-dependent remainder is learned more gradually but concentrates in a low-dimensional subspace. Substantial zero-shot prediction error can therefore coexist with useful behavior when that error does not disrupt the predictive distinctions needed for control, and with rapid adaptation when the remaining mismatch is easy to estimate and realize.

We develop this explanation in four steps. First, we characterize the required functional and separate its globally shared and state-dependent components, revealing how recovery of the former can account for substantial early progress (Section 4). Second, we explain few-shot recovery of this dominant component by isolating its statistical and realization difficulties (Section 5). Next, we examine what remains beyond global recovery and show that the state-dependent correction retains strong low-dimensional structure under foundation representations (Section 6). Finally, we trace this functional structure upstream to the regular cross-domain transformations induced in foundation representation space (Section 7). Together, these analyses connect pretrained representation geometry to the effective complexity of latent-dynamics adaptation.

2 Related Work

Latent dynamics models. Latent dynamics models enable future prediction and control from high-dimensional observations. PlaNet (Hafner et al., 2019) learns stochastic latent dynamics from pixels for online planning, while Dreamer (Hafner et al., 2020; 2025) optimizes behavior through imagined latent trajectories. TD-MPC (Hansen et al., 2022; 2024) develops task-oriented latent models for scalable model-predictive control. Recently, LeWorldModel (Maes et al., 2026) jointly learns an encoder and transition predictor from pixels without pretrained features. These approaches underscore the coupling between representation and dynamics learning: useful latent spaces must preserve control-relevant predictive distinctions while supporting transition modeling.

Foundation representation and foundation-based latent dynamics. Large-scale pretraining yields transferable features for perception or control through video-based or visual–language representation learning, as in R3M (Oquab et al., 2023) and VIP (Ma et al., 2022); self-supervised learning, as in DINO (Caron et al., 2021; Oquab et al., 2023) and MAE (He et al., 2022; Fan et al., 2025); and vision–language alignment, as in SigLIP (Zhai et al., 2023; Tschannen et al., 2025). They also enable dynamics learning without training representation from scratch. DINO-WM (Zhou et al., 2024) predicts future DINO patch features for zero-shot planning, while DINO-world (Baldassarre et al., 2025) scales prediction to broader video distributions before action-conditioned fine-tuning. These approaches establish pretrained representations as a strong, reusable foundation for latent prediction, decoupling dynamics fitting from representation learning.

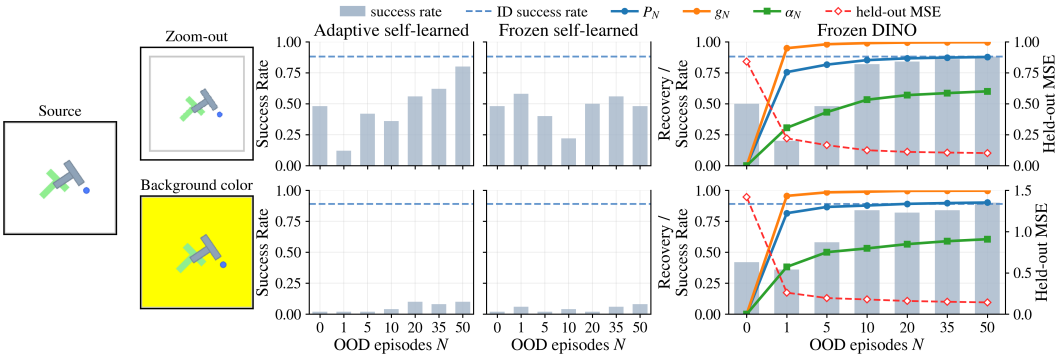


Figure 2: Latent-dynamics fine-tuning results on PushT under zoom-out (top) and background-color (bottom) shifts. **Left:** Source and target observations. **Middle:** Closed-loop success remains low with frozen self-learned encoders and improves slowly under joint encoder–dynamics adaptation. **Right:** Fine-tuning with a frozen DINO encoder rapidly restores performance and reduces held-out prediction error. The decomposition of total progress P_N attributes most early functional progress to global recovery g_N, while the state-dependent remainder α_N proceeds more gradually. Dashed horizontal lines indicate source-domain success rates.

Latent dynamics adaptation. A growing body of work studies how pretrained latent dynamics can be reused or updated under distribution shift. MoVie (Yang et al., 2023) addresses visual generalization through reward-free test-time adaptation of model-based policies. ReDRAW (Lanier et al., 2025) addresses dynamical discrepancies by learning latent-state residual corrections. More recently, AdaJEPA (Wang et al., 2026) uses newly observed transitions as self-supervised targets during deployment, while Sandwich-Residuals (Soni et al., 2026) learns lightweight corrections around the frozen predictor to achieve parameter-efficient adaptation. In this work, our focus is not a new adaptation algorithm, but characterizing the functional structure induced by foundation representations that enables sample-efficient latent-dynamics adaptation.

3 Preliminary

Controlled dynamics and visual shifts. We consider a controlled system with internal state $s_t \in \mathcal{S}$, action $a_t \in \mathcal{A}$, and true dynamics $s_{t+1} = f_*(s_t, a_t)$. A visual domain $d \in \{\mathrm{I}, \mathrm{O}\}$ specifies an observation map $R_d : \mathcal{S} \to \mathcal{O}$, yielding $o_t^d = R_d(s_t)$. We study visual shifts from source domain I to target domain O that change the observation map while preserving physical dynamics.

Foundation-based latent dynamics. Let $\phi : \mathcal{O} \to \mathcal{Z}$ be a frozen pretrained foundation encoder. The latent observation and length-H context in domain d are

$$z_t^d = \phi(R_d(s_t)), \qquad C_t^d = (z_{t-H+1}^d, \ldots, z_t^d). \tag{1}$$

A source-trained latent dynamics model F_{θ_I} predicts the next latent observation from the current latent context and action: $\hat{z}_{t+1}^{\mathrm{I}} = F_{\theta_\mathrm{I}}(C_t^{\mathrm{I}}, a_t)$. Unchanged physical dynamics do not imply unchanged latent dynamics: visual shifts alter the transition relation induced by $\phi \circ R_d$.

Target-domain fine-tuning. Given N target-domain episodes $\{(o_{e,t}^{\mathrm{O}}, a_{e,t})_{t=1}^{T_e}\}_{e=1}^{N}$, the frozen encoder ϕ yields latent observations $z_{e,t}^{\mathrm{O}} = \phi(o_{e,t}^{\mathrm{O}})$ and corresponding context $C_{e,t}^{\mathrm{O}}$. Starting from the source-trained parameters θ_I, we fine-tune the dynamics model by optimizing

$$\theta_N = \arg\min_{\theta \leftarrow \theta_\mathrm{I}} \frac{1}{N} \sum_{e=1}^{N} \frac{1}{T_e - H} \sum_{t=H}^{T_e - 1} L\Big(F_\theta(C_{e,t}^{\mathrm{O}}, a_{e,t}), z_{e,t+1}^{\mathrm{O}}\Big), \tag{2}$$

where L is a generic prediction loss and $T_e > H$. The resulting parameters θ_N define the adapted model $F_N := F_{\theta_N}$, while $F_0 := F_{\theta_\mathrm{I}}$. Let F_* denote the true target-domain latent dynamics induced by f_* and $\phi \circ R_\mathrm{O}$. We seek to explain why, for many visual shifts, foundation representations enable F_N to approximate F_* with only a few target-domain episodes.

4 FUNCTIONAL STRUCTURE OF TARGET-DOMAIN ADAPTATION

A source-trained latent dynamics model F_0 already captures action-conditioned predictive structure from source data. Under dynamics-preserving visual shifts, adaptation must correct how the same physical transition law is expressed through representation. To make this dependence explicit, suppose corresponding source- and target-domain latent observations admit a transport κ_ϕ satisfying $\phi(R_{\mathrm{O}}(s)) = \kappa_\phi(\phi(R_{\mathrm{I}}(s)))$, extended componentwise to contexts. The true dynamics then satisfy

$$F_*^{\mathrm{O}}(\kappa_\phi(C), a) = \kappa_\phi(F_*^{\mathrm{I}}(C, a)). \tag{3}$$

Neglecting transported source-model error, the required target-domain correction is

$$\Delta_*^{\phi}(\kappa_\phi(C), a) \approx \kappa_\phi(F_0(C, a)) - F_0(\kappa_\phi(C), a). \tag{4}$$

This expression isolates the representation-dependent burden: the mismatch between transporting a prediction and predicting a transported input. Its magnitude alone does not determine adaptation difficulty: even a large mismatch can admit a simple correction when it is consistent across inputs. We therefore study adaptation directly through the functional correction required of F_0.

Fine-tuning progress. For a context-action input $x \coloneqq (C^{\mathrm{O}}, a)$, let

$$\Delta_*(x) \coloneqq F_*(x) - F_0(x), \qquad \Delta_N(x) \coloneqq F_N(x) - F_0(x). \tag{5}$$

denote the correction required to recover the true target-domain dynamics and the correction learned from N target-domain episodes, respectively. We measure *fine-tuning progress* as the fraction of the source model's functional error removed after adaptation:

$$P_N \coloneqq 1 - \frac{\mathbb{E}_x\|\Delta_N(x) - \Delta_*(x)\|_2^2}{\mathbb{E}_x\|\Delta_*(x)\|_2^2}, \tag{6}$$

with expectations over a fixed target-domain input distribution. This normalization measures recovery relative to the required correction. Thus, $P_N = 0$ for the unadapted model and $P_N = 1$ upon exact recovery of F_* almost surely. Consequently, examining P_N distinguishes the magnitude of the target-domain mismatch from the efficiency with which fine-tuning corrects it.

Global and state-dependent adaptation. For $j \in \{*, N\}$, each correction decomposes into a globally shared component and an input-dependent remainder:

$$\Delta_j(x) = \mu_j + \delta_j(x), \qquad \mu_j \coloneqq \mathbb{E}_x[\Delta_j(x)], \tag{7}$$

where μ_j is the best constant approximation to $\Delta_j(x)$, and δ_j captures the remaining state-dependent variation. Since the remainder has zero mean, the two components are orthogonal in L^2, and their energies add. This yields the *global adaptation fraction*

$$\eta \coloneqq \frac{\|\mu_*\|_2^2}{\mathbb{E}_x\|\Delta_*(x)\|_2^2} = \frac{\|\mu_*\|_2^2}{\|\mu_*\|_2^2 + \mathbb{E}_x\|\delta_*(x)\|_2^2} \in [0, 1]. \tag{8}$$

The quantity η measures the globally shared fraction of required adaptation energy. A large η indicates that the correction is predominantly shared across inputs, not that little adaptation is required. For a fixed input distribution, η is determined by F_0 and F_*, reflecting how the representation expresses physical transitions across domains, independently of N and subsequent fine-tuning.

Lemma 1 (Exact decomposition of fine-tuning progress)**.** *Under the preceding assumptions with $0 < \eta < 1$, every adapted model F_N satisfies $P_N = \eta g_N + (1 - \eta)\alpha_N$, where the global and state-dependent recovery scores are, respectively,*

$$g_N \coloneqq 1 - \frac{\|\mu_N - \mu_*\|_2^2}{\|\mu_*\|_2^2}, \qquad \alpha_N \coloneqq 1 - \frac{\mathbb{E}_x\|\delta_N(x) - \delta_*(x)\|_2^2}{\mathbb{E}_x\|\delta_*(x)\|_2^2}. \tag{9}$$

The lemma gives η a direct operational interpretation: exact recovery of the global correction, with the state-dependent remainder unlearned, yields $g_N = 1$, $\alpha_N = 0$, and $P_N = \eta$. A large η therefore permits substantial functional recovery before detailed state dependence is learned.

Global dominance is representation-dependent. Figure 2 reveals a representation-dependent gap in adaptation efficiency. Foundation-based fine-tuning yields rapid recovery from only a few target-domain episodes. In contrast, dynamics-only fine-tuning with a frozen, source-specific encoder provides little improvement, while joint encoder–dynamics adaptation improves performance but remains below the source-domain level under the same data budget. Strong source-domain prediction alone therefore does not ensure few-shot adaptation.

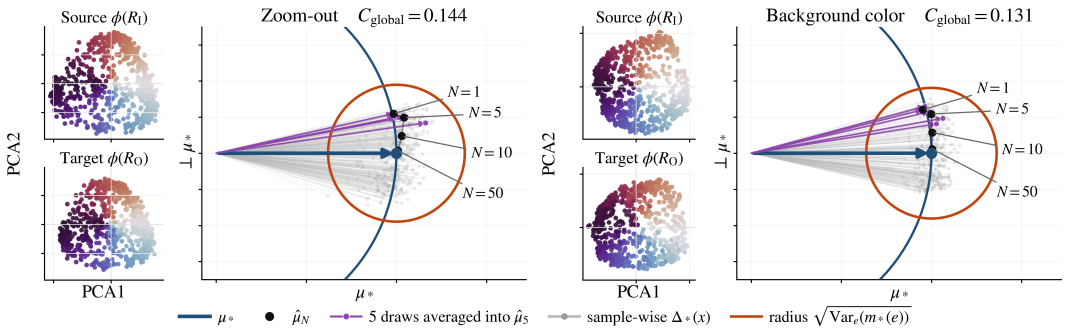


Figure 3: Cross-episode correction geometry on PushT. Insets show source and target latent distributions under frozen DINO features, revealing modest but nonzero changes in representation geometry across domains. The correction plots show pointwise state dependence but low variability of episode means relative to the shared correction's magnitude, yielding small C_{global} and rapid convergence of $\widehat{\mu}_N$ toward μ_*. Purple examples show that five-episode averages closely approximate μ_*.

This contrast reflects a representation-dependent difference in correction structure. Table 1 shows that the globally shared component μ_* accounts for a large fraction of the required adaptation energy under the foundation representation, whereas this global structure is much weaker under the self-learned representation. The observed fine-tuning trajectories follow this structure. Under the foundation representation, rapid global recovery g_N drives early progress, while state-dependent recovery α_N improves more gradually. A large η indicates that most adaptation energy is globally shared, so recovering μ_* alone can remove a substantial fraction of the target-domain error. Faster adaptation need not coincide with lower initial latent prediction error. The foundation-based model can exhibit larger zero-shot held-out error yet recover substantially faster for simple correction structure.

Representation	**Shift**	η
Self-learned	Zoom-out	0.0087
	Bg. color	0.194
DINO	Zoom-out	0.699
	Bg. color	0.756

Table 1: Global fraction across shifts.

5 Why Is the Globally Shared Correction Few-Shot Learnable?

Global dominance does not by itself establish few-shot learnability. Two questions remain. First, how accurately can the global correction μ_* be estimated from a few target-domain episodes? Second, what parameter-space complexity is required to realize this correction while preserving the source model's learned predictive structure? We analyze these questions through statistical estimation and parameter-space realization. For a target-domain training set $D_N = \{e_i\}_{i=1}^N$, define the empirical mean correction and decompose the global recovery error as

$$\hat{\mu}_N := \frac{1}{N}\sum_{i=1}^{N} E_{x|e_i}[\Delta_*(x)], \qquad \mu_N - \mu_* = \underbrace{\mu_N - \hat{\mu}_N}_{\text{realization error}} + \underbrace{\hat{\mu}_N - \mu_*}_{\text{statistical error}}. \tag{10}$$

The statistical term asks whether the sampled episodes identifies μ_*, while the realization term measures how closely fine-tuning reproduces $\hat{\mu}_N$ in population mean. We examine these requirements through relative cross-episode variability and intrinsic realization complexity, respectively.

5.1 Statistical Difficulty: Cross-Episode Consistency

The shared correction is statistically accessible when individual episodes provide consistent estimates of its population mean. For a target episode e, define $m_*(e) := \mathbb{E}_{x|e}[\Delta_*(x)]$. The population correction and its estimate from N target-domain episodes are

$$\mu_* = \mathbb{E}_e[m_*(e)], \qquad \hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} m_*(e_i). \tag{11}$$

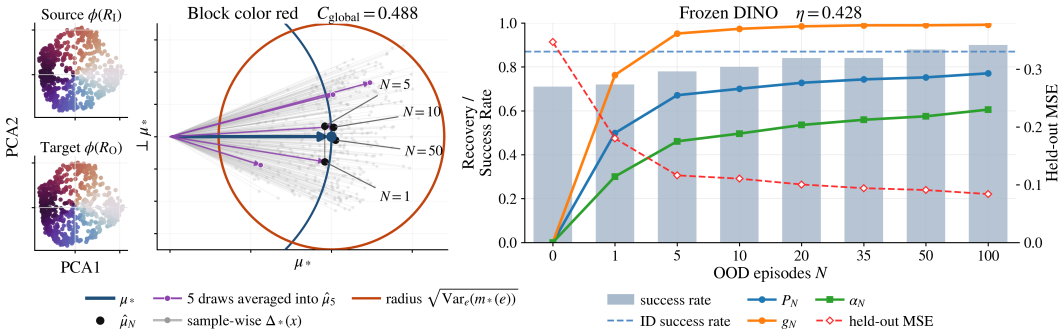


Figure 4: PushT block-color shift under DINO representations. **Left:** Latent distributions and the correction geometry. **Right:** Fine-tuning progress. Relative to the low-C_{global} shifts, greater cross-episode variability delays global recovery, with control recovery observed only after the shared correction is largely learned.

Cross-episode statistical difficulty. For i.i.d. episodes, independence and centering yield

$$\mathbb{E}_{D_N}\left[\frac{\|\hat{\mu}_N - \mu_*\|_2^2}{\|\mu_*\|_2^2}\right] = \frac{1}{\|\mu_*\|_2^2}\mathbb{E}_{D_N}\left\|\frac{1}{N}\sum_{i=1}^{N}(m_*(e_i) - \mu_*)\right\|_2^2 = \frac{1}{N}\frac{\text{Var}_e(m_*(e))}{\|\mu_*\|_2^2}. \tag{12}$$

where $\text{Var}_e(Y)$ denotes total variance. This identity identifies the *global statistical difficulty*

$$C_{\text{global}} := \frac{\text{Var}_e(m_*(e))}{\|\mu_*\|_2^2}, \tag{13}$$

which quantifies cross-episode variability relative to the shared correction's squared magnitude.

Lemma 2 (Statistical recovery of the global correction)**.** *Under the preceding assumptions, the expected relative squared estimation error of global recovery equals C_{global}/N, decaying at rate N^{-1}. Consequently, achieving expected error at most ε requires $N \geq C_{\text{global}}/\varepsilon$.*

Whereas η measures the globally shared fraction of required adaptation, C_{global} quantifies how consistently that correction is exposed across episodes. Few-shot recovery depends on dispersion of episode means $m_*(e_i)$, not on pointwise variation of $\Delta_*(x)$.

Cross-episode correction geometry governs global recovery. The correction geometry distinguishes two scales of variation: the spread of $\Delta_*(x)$ within episodes and the dispersion of episode means $m_*(e)$ across episodes. Few-shot estimation depends on the latter relative to the shared correction's magnitude. Thus, substantial pointwise state dependence can coexist with small C_{global} when episode-level centroids concentrate around the common mean μ_*. Figure 3 illustrates this regime: despite representation non-invariance and pointwise variation, mean corrections from independently sampled episodes cluster tightly around μ_*. The dominant component is therefore not merely shared in expectation but consistently exposed by individual episodes.

The block-color shift provides the contrasting geometry. Its episode-level means are substantially more dispersed relative to $\|\mu_*\|$ (Figure 4), yielding $C_{\text{global}} = 0.488$. The Lemma 2 then requires at least 49 episodes for the episode-mean estimator to achieve expected estimation error below 1%, since $N > 0.488/0.01 = 48.8$. Consistently, control performance approaches the source-domain level only around $N = 50$, despite 71% zero-shot success. This contrast distinguishes initial transfer from few-shot adaptability: favorable zero-shot performance does not ensure that the remaining shared correction is geometrically consistent enough across episodes to be estimated from limited data. Thus, C_{global} translates the cross-episode geometry of the correction field induced by the foundation representation into a sample-complexity criterion for adapting the globally shared correction.



Figure 5: PushT under block-color shift.

5.2 Realization Difficulty: Low-Complexity Adaptation

Beyond statistical accessibility, the estimated correction $\widehat{\mu}_N$ must also be expressible without broadly reconfiguring the source-trained dynamics. Since F_0 already encodes source-domain predic-

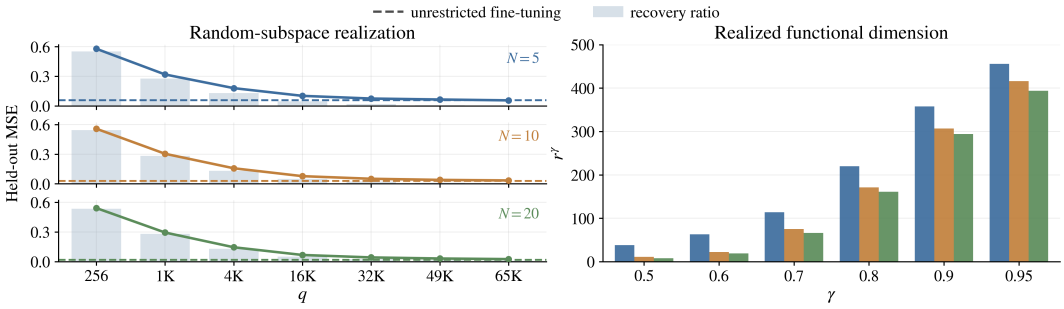


Figure 6: Realization complexity of the global correction under the PushT zoom-out shift. **Left:** Restricted fine-tuning approaches unrestricted global recovery as the random-subspace dimension q increases. **Right:** Only $r^\gamma = O(10^2)$ leading functional modes capture most of the realized correction's energy. At a fixed energy level, r^γ decreases with N, consistent with reduced sampling variability in $\widehat{\mu}_N$.

tive structure, we ask whether this correction can be realized through a low-complexity modification around the pretrained solution.

Intrinsic adaptation complexity. Ambient parameter count can overstate adaptation complexity. The intrinsic-dimension perspective characterizes this complexity by the dimension of a randomly oriented parameter subspace sufficient to recover near-full performance (Li et al., 2018). For p adaptable parameters, we restrict fine-tuning to a q-dimensional subspace:

$$\theta = \theta_0 + P\beta, \qquad P \in \mathbb{R}^{p\times q}, \qquad \beta \in \mathbb{R}^q, \tag{14}$$

where P is randomly sampled and held fixed, and only β is optimized. Matching unrestricted recovery with $q \ll p$ demonstrates that the required correction is accessible through a low-dimensional update subspace anchored at the pretrained solution, rather than requiring full $\mathbb{R}^p$ optimization.

Realized functional complexity. Restricted fitting, however, still gives only a conservative parameter-space notion of complexity. Random orientation may require a large search subspace to expose useful update directions, even when the corresponding prediction change concentrates in few functional modes. For an N-episode dataset D_N containing M_N transitions with latent prediction dimension d_z, let $J_{D_N}(\theta)$ denote the stacked prediction Jacobian with respect to β. For a successful update β_N, we characterize the realized update through the path-averaged Jacobian

$$\overline{\mathcal{J}}_{D_N} := \int_0^1 \mathcal{J}_{D_N}(\theta_0 + tP\beta_N)\, dt \in \mathbb{R}^{M_N d_z \times q}. \tag{15}$$

By calculus, $\overline{\mathcal{J}}_{D_N}\beta_N$ equals the exact stacked prediction change on D_N. Writing $\overline{\mathcal{J}}_{D_N} = U\Sigma V^\top$, with singular values σ_i and right singular vectors v_i, the realized correction contributes energy $e_i := \sigma_i^2(v_i^\top \beta_N)^2$ to mode i, and

$$r^\gamma := \min\left\{ r : \frac{\sum_{i=1}^r e_i}{\sum_i e_i} \geq \gamma \right\}, \tag{16}$$

counts the leading singular modes needed to capture a fraction γ of the realized correction's energy. Thus, q measures the parameter-space width available to restricted optimization, whereas r^γ the complexity of the correction actually used.

Low-complexity global realization around pretrained dynamics. We restrict fine-tuning to random parameter subspaces of increasing dimension q, supervised by the empirical correction $\widehat{\mu}_N$. Figure 6 demonstrates that the shared correction is accessible through only a thin portion of the model's parameter space. Across sample sizes, increasing q systematically closes the realization gap. A 65,536-dimensional random subspace, only 0.32% of the model's 20.2-million-dimensional parameter space, matches unrestricted global recovery.

The realized functional change is simpler still. Although the random search subspace contains tens of thousands of parameter directions, more than 90% of the realized correction energy concentrates in only $O(10^2)$ leading functional modes. Thus, the parameter-space width used to access a successful update substantially exceeds the functional dimension carrying most of the correction. Together,

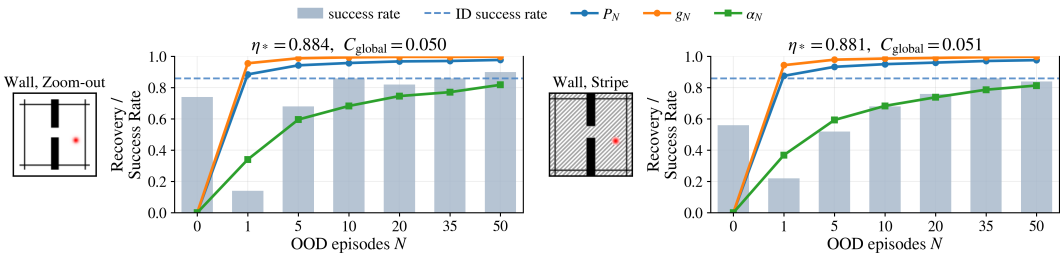


Figure 7: Fine-tuning progress on Wall under zoom-out and stripe shifts under DINO representations. Despite nearly identical component recovery, closed-loop success improves at different rates. Thus, closely matched functional recovery does not uniquely determine behavioral recovery.

these results demonstrate that the correction identified from few target episodes is accessible through low-dimensional updates anchored at the pretrained solution and concentrated in an even narrower functional spectrum, without requiring unrestricted optimization of the dynamics model.

Moreover, for fixed γ, r^γ decreases with additional target episodes. This trend is consistent with episode averaging suppressing sampling variability in $\widehat{\mu}_N$, making it more representative of μ_*. Improved statistical estimation is thus accompanied by greater concentration of the realized correction in fewer leading functional modes, rather than increased functional complexity.

6 Beyond Global Recovery: Structured State-Dependency

Section 5 establishes that the dominant shared correction is statistically accessible from few target episodes and realizable through low-complexity updates. A complete account of adaptation must also characterize the state-dependent remainder. We examine its relevance to downstream recovery and show that foundation representations retain a second level of structure beyond global dominance.

Limits of global recovery. We examine Wall, whose largely static scene geometry contrasts with PushT. This setting tests the persistence of globally shared adaptation while highlighting an important distinction: global recovery is a necessary but not sufficient condition for closed-loop recovery. Figure 7 shows high η and low C_{global} across visual shifts, with rapid global recovery g_N accounting for most early functional progress P_N. Yet closed-loop performance improves at different rates after g_N has largely saturated, despite nearly identical global recovery. These results separate recovery of the dominant correction from behavioral recovery, suggesting that lower-energy state-dependent directions remain consequential after the shared component is learned.

Low-dimensional structure beyond the shared correction. The state-dependent remainder need not introduce unstructured high-dimensional variation. To characterize it directly, we stack $\delta_*(x)$ across held-out inputs and compute r^δ_{95}, the minimum number of leading singular modes capturing 95% of its energy. Table 2 reports the normalized rank r^δ_{95}/d_z, showing that most residual variation occupies less than 1% of the latent dimension under the foundation representation, whereas source-specific representations require a substantial fraction of that space. The persistence of this concentration therefore supports a representation-dependent reduction in residual functional dimensionality. This residual structure does not make the remainder negligible. Identifying the control-critical directions within this subspace and determining how efficiently they can be learned remain important directions for future work.

Task	Shift	r^δ_{95}/d_z	
		Self-learned	DINO
PushT	Zoom-out	0.453	**0.0066**
	Blk. color	0.286	**0.0073**
	Bg. color	0.271	**0.0068**
Wall	Zoom-out	0.130	**0.0033**
	Stripe	0.182	**0.0039**

Table 2: Normalized 95%-energy rank of the state-dependent correction.

7 Transport Structure Underlying Few-Shot Adaptation

The preceding analysis shows that the structure of functional corrections is shaped by the representation and source dynamics before target-domain fine-tuning even begins. We finally discuss where

this structure comes from. The transport relation in (4) provides a representation-level explanation how the correction is governed by the interaction between source-trained F_0 and cross-domain transport κ_ϕ. If the transport changes arbitrarily from state to state, this interaction can induce correspondingly complex adaptation. In contrast, when the transport follows a structured and consistent rule across the state space, the family of possible corrections is substantially constrained.

The structure of κ_ϕ reflects how foundation pretraining organizes appearance and semantic variation, allowing visual shifts to be expressed through regular rather than arbitrary transformations. We assess this structure using paired source- and target-domain observations of the same physical states. Table 3 shows that even a simple affine parameterization captures much of the cross-domain latent transport for DINO and Web-SSL under the considered shifts, whereas SigLIP exhibits weaker fits. The same visual shifts thus produce substantially different transport regularity depending on the representation. Although the true transport need not be affine, combining affine transport with linear latent dynamics makes explicit how their interaction shapes the required correction. For a fixed action, assume exact source dynamics $F_0(z) = Az + b$ and transport $\kappa_\phi(z) = Mz + c$. The required correction is

Shift	Representation	Affine MSE ↓	Gap closed ↑
Zoom-out	DINO	0.02762	**91.9%**
	Web-SSL	0.01672	91.3%
	SigLIP	0.25273	63.7%
Block color	DINO	0.01586	71.3%
	Web-SSL	0.00206	**92.6%**
	SigLIP	0.18239	51.6%

Table 3: Affine transport fits across foundation representations. Gap closure is measured relative to the true cross-domain transport.

$$\Delta_*(\kappa(z)) = (MA - AM)z + (M - I)b + (I - A)c. \tag{17}$$

Writing $\bar{z} = \mathbb{E}[z]$ under the corresponding input distribution gives

$$\mu_* = (MA - AM)\bar{z} + (M - I)b + (I - A)c, \tag{18}$$

$$\delta_*(\kappa(z)) = (MA - AM)(z - \bar{z}). \tag{19}$$

The global correction combines the constant terms with the mean interaction between transport and dynamics, while all remaining state dependence is governed by the commutator $MA - AM$. When the linear parts commute, $MA = AM$, the correction is purely global despite a nonidentity transport. More generally, the residual lies in the image of $MA - AM$, confining it to a subspace of dimension at most $\mathrm{rank}(MA - AM)$. This example makes the mechanism explicit: foundation-based few-shot adaptability depends on how representation-induced transport interacts with pretrained dynamics to structure the correction that fine-tuning must estimate and realize.

8 Conclusion

We identify a structural explanation for foundation-based few-shot dynamics adaptation. Across the representations and visual shifts studied, much of the required functional correction is globally shared, consistently exposed across episodes, and realizable through low-dimensional parameter updates. These properties explain how finetuning can recover substantial functional accuracy by modifying source-trained dynamics rather than learning the target dynamics from scratch. The state-dependent remainder can remain consequential but concentrates in a low-dimensional subspace. Transport analysis connects these two levels of simplification to the interaction between cross-domain representation geometry and pretrained dynamics.

Taken together, these findings suggest that representations for adaptable dynamics should make cross-domain learning a process of structured refinement rather than reconstruction. Their design should therefore promote commutation between pretrained dynamics and a broad family of cross-domain transports. Since latent dynamics express the underlying physical transition law, this motivates dynamics-aware representations whose cross-domain geometry preserves physical structure. Characterizing which residual directions matter and how training data should expose them remains an important direction. Because behavior policies shape both episode-level correction variability and the residual directions encountered, their design should support reliable global estimation while providing informative coverage of control-relevant state dependence. Jointly designing representations and data-collection policies toward these goals offers a path to dynamics models that acquire new capabilities by extending, rather than overwriting, existing predictive knowledge.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing and sentence-level polishing, code debugging, and implementation support for porting multiple pretrained large models into the downstream training or evaluation pipeline. We also used these tools to assist with figure revision; no figure reported in the paper was generated by AI. We did not use generative AI to generate experimental datasets, or replace the execution of experiments. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

No specific ethic concerns are identified for this work.

REPRODUCIBILITY STATEMENT

All pretrained models and training procedures are either publicly released or follow standard published protocols. We will release the codebase upon publication.

REFERENCES

Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–13. IEEE, 2025.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

JB Lanier, Kyungmin Kim, Armin Karamzade, Yifei Liu, Ankita Sinha, Kat He, Davide Corsi, and Roy Fox. Adapting world models with latent-state dynamics residuals. *arXiv preprint arXiv:2504.02252*, 2025.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. In *international conference on machine learning*, pp. 17359–17371. PMLR, 2022.

Krishnam Soni, Aditya Sehgal, Vedant Dave, and Elmar Rueckert. Sandwich-residuals: Parameter-efficient test-time adaptation of world models. *arXiv preprint arXiv:2609.21740*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ying Wang, Oumayma Bounou, Yann LeCun, and Mengye Ren. Adajepa: An adaptive latent world model. *arXiv preprint arXiv:2606.32026*, 2026.

Sizhe Yang, Yanjie Ze, and Huazhe Xu. Movie: Visual model-based policy adaptation for view generalization. *Advances in Neural Information Processing Systems*, 36:21507–21523, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Appendix

We provide experimental settings in this appendix.

Foundation model usage and latent-dynamics learning. Unless otherwise stated, all experiments use a DINOv2 encoder (Oquab et al., 2023) frozen throughout source dynamics learning and fine-tuning. In Section 7, we experiment on Web-SSL(MAE) (Fan et al., 2025) and the vision–language foundation model SigLIP2 (Tschannen et al., 2025). These foundation models produce high-dimensional representation grids. In our configurations, DINOv2 yields 256 patch tokens of dimension 384, Web-SSL (MAE) yields 196 tokens of dimension 1024, and SigLIP2 yields 196 tokens of dimension 768 per frame. Therefore, for the backbone of the latent dynamics model, we preserve the patchwise representations and model their temporal evolution with a block-casual Transformer following the standard DINO-WM architecture (Zhou et al., 2024).

We train the latent dynamics models using the LeWorldModel training pipeline (Maes et al., 2026) on the expert PushT and Wall datasets released with DINO-WM. The foundation encoder remains frozen at all stages. For source-domain training, we use a large optimization budget of 100 epochs to ensure that subsequent adaptation behavior is not limited by an undertrained source model. The resulting source models achieve 87% closed-loop success on PushT and 86% on Wall, evaluated over 100 trials, matching the publicly reported performance on these tasks.

Target-domain fine-tuning and oracle quantities. Target-domain datasets are constructed from the same expert trajectories as the source data, with observations re-rendered under the considered visual shifts so that the underlying states, actions, and physical dynamics remain unchanged. For each target-domain dataset, we partition the available trajectories into 80% for training and 20% for validation, with the validation split used for model selection and convergence assessment. Consequently, in the low-data regime, the effective training data is smaller than the nominal episode count; for example, $N = 1$ uses only 80% of the transitions from a single episode for optimization. We report results over three random seeds for all settings and increase this to ten for $N \leq 5$.

For quantities defined with respect to the true target dynamics, including η_*, C_{global}, and the recovery terms involving $(\cdot)_*$, we use the ground-truth next observations from the re-rendered trajectories and encode them directly with the foundation encoder. Since the representation is fixed, this provides an oracle target latent transition for each sampled input, without introducing error from an auxiliary learned dynamics model. The reported quantities are therefore estimated directly from ground-truth latent transitions over the evaluation distribution.

Intrinsic-difficulty experiments. For restricted fine-tuning, we construct the fixed random projection $P \in \mathbb{R}^{p\times q}$ using a sparse random-subspace parameterization following the intrinsic-dimension methodology of Li et al. (2018). Each column of P contains exactly $s = 64$ nonzero entries, whose locations are sampled uniformly without replacement from the p ambient parameters and whose signs are sampled independently from $\{-1, +1\}$. All entries are scaled by $1/\sqrt{s}$ so that each column has unit norm. This sparsity makes projections computationally tractable while retaining a random orientation of the update subspace.

For the realized functional-complexity analysis, we estimate the path-averaged Jacobian $\overline{\mathcal{J}}_{D_N}$ using three-point Gauss–Legendre quadrature along the straight-line parameter path from initial source-trained θ_0 to the final unrestricted update $\theta_0 + P\beta_N$. This accounts for finite prediction changes beyond a first-order linerization at the source model. We compute the realized spectrum, using β_N from the largest restricted subspace considered, $q = 65{,}536$, whose held-out recovery matches unrestricted full-parameter fine-tuning. Therefore, the reported functional ranks characterize a successful realization rather than an under-capacity update.

Affine transport fit. We fit cross-domain transports using paired source- and target-domain representations of the same underlying states under each visual shift. Because the latent representation z is high-dimensional, we restrict the affine family to the structured form $PZR + B$, optimizing the parameters P, R and B, instead of fitting an unconstrained full affine map. This is therefore a strict subclass of all affine transformations and should be interpreted as a conservative probe of transport regularity. Nevertheless, this restricted parameterization already explains a large fraction of the paired cross-domain variation for several foundation representations. To isolate this representation-dependent geometry from data scarcity, we fit the transport using the same amount of paired data as used for source-domain dynamics learning.