THE GEOMETRY OF ROLLOUT ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon rollout remains a central challenge in learned physical dynamics. Models with similar one-step accuracy can produce radically different trajectories, so local prediction error alone does not determine behavior under repeated composition. We analyze this separation through the finite-time transport of local residuals: errors introduced at different times and in different directions can have very different consequences depending on the subsequent learned dynamics through which they propagate.

This perspective motivates separating three choices in autoregressive learned dynamics: *temporal representation*, which determines whether the model learns an instantaneous generator or the finite-time transition that is actually composed at deployment; *forward exposure*, which determines how many self-fed predicted states are encountered during training; and *temporal credit*, which determines how far learning signals propagate backward through that rollout. Across ten evaluation conditions spanning nine physical systems from The Well, DISCO-FLO, a direct finite-time-transition formulation of DISCO, achieves lower final-horizon median error than generator-based DISCO in all ten conditions despite lower one-step median error in only four. A matched-credit experiment shows that the representation effect persists when generator and transition models receive the same self-fed exposure and backward credit, while credit-depth ablations show that increasing temporal credit is not uniformly beneficial across systems.

Together, these results suggest that long-horizon learned dynamics should be understood and optimized through the geometry of their deployed composition rather than inferred from local prediction accuracy alone.

1 INTRODUCTION

Learned physical dynamics are commonly optimized using local prediction objectives but deployed through repeated self-fed composition. Yet local accuracy does not, in general, determine accuracy under composition: models with comparable one-step error can produce very different long-horizon trajectories. This distinction can be understood through the transport of local prediction residuals by subsequent learned dynamics (Zhang & Jo, 2026). For a self-fed rollout initialized from the reference state, let

$$\boldsymbol{e}_t = \hat{\boldsymbol{x}}_t - \boldsymbol{x}_t, \qquad \boldsymbol{r}_t = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \Phi_{\Delta}(\boldsymbol{x}_t)$$

denote the rollout error and local prediction residual, respectively. If $\overline{\boldsymbol{J}}_t$ denotes the averaged Jacobian along the segment between $\boldsymbol{x}_t$ and $\hat{\boldsymbol{x}}_t$, then the finite-horizon error admits the exact residual-transport decomposition

$$\boldsymbol{e}_H = \sum_{t=0}^{H-1} \overline{\boldsymbol{J}}_{H-1:t+1} \boldsymbol{r}_t, \tag{1}$$

where $\overline{\boldsymbol{J}}_{H-1:t+1} = \overline{\boldsymbol{J}}_{H-1} \cdots \overline{\boldsymbol{J}}_{t+1}$. Thus the long-horizon consequence of a local error depends not only on its magnitude, but also on its direction, when it occurs, and the finite-time dynamics it subsequently encounters. Equal one-step error therefore need not imply equal finite-horizon error.

This residual-transport view motivates what we call the *geometry of rollout error*. Locally, repeated composition induces a finite-horizon geometry on prediction residuals: errors of equal magnitude can incur very different trajectory costs depending on when they are introduced and how they align with directions that are amplified, suppressed, or redirected by subsequent dynamics. The relevant

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.

object is therefore not local prediction error alone, but the interaction between the residuals a model makes and the dynamics through which those residuals propagate.

This perspective suggests separating three choices in autoregressive learned dynamics. The first is *temporal representation*: whether a model learns an instantaneous generator that is subsequently integrated or directly learns the finite-time transition composed at inference. The second is *forward exposure*: how many states generated from the model's own predictions it encounters during training. The third is *temporal credit*: how far losses from those future states propagate backward through the rollout. These choices are conceptually independent: representation changes the dynamical object being composed, exposure changes the state distribution encountered during training, and credit changes how learning signals propagate through that composition.

Motivated by this geometry, we introduce *DISCO-FLO*, a Finite-time Learned Operator formulation of DISCO that directly learns the deployed finite-time transition and optimizes it through self-fed rollout. Across nine physical systems from The Well, DISCO-FLO substantially improves long-horizon prediction over generator-based DISCO. Importantly, this gain is not simply a consequence of giving the transition model a multi-step training objective. In matched-credit experiments, generator and transition models receive the same forward rollout exposure and the same backward credit, yet their long-horizon behavior still separates substantially. The temporal object being learned therefore matters independently of the amount of rollout credit used to train it.

Temporal credit is itself consequential, but its effect is not universally monotonic. Holding the finite-time transition representation and forward rollout horizon fixed, shallow credit performs consistently worse, while the relative benefit of intermediate and full backpropagation varies across physical systems. Long-horizon performance therefore cannot be reduced either to "use a transition rather than a generator" or to "backpropagate through as many rollout steps as possible." Representation and temporal credit therefore have distinct effects even when forward exposure is controlled.

Our contributions are:

- **A finite-horizon geometry of rollout error.** Building on the error-transport view of autoregressive dynamics, we collect the sequence of local residuals into a finite-horizon transport operator whose pullback defines the local quadratic rollout cost. This makes explicit how residual direction, timing, and cross-time interaction determine trajectory error under composition.
- **DISCO-FLO across heterogeneous physical dynamics.** We introduce DISCO-FLO, a finite-time-transition formulation of DISCO trained under self-fed rollout. Across ten evaluation conditions spanning nine physical systems from The Well, DISCO-FLO has lower final-horizon median NRMSE than generator-based DISCO in all ten conditions despite lower one-step median error in only four.
- **Controlled separation of representation, exposure, and credit.** A matched-exposure, matched-credit experiment isolates a substantial temporal representation effect, while credit-depth sweeps at fixed representation and forward exposure show that shallow credit is consistently worse but additional credit is not monotonically beneficial across systems.

2 Related Work

Autoregressive rollout training. Learned simulators and neural PDE solvers are commonly deployed autoregressively, so prediction errors alter the inputs encountered at later steps. Prior work addresses this train–deployment mismatch through input perturbations (Sanchez-Gonzalez et al., 2020), temporal bundling and pushforward training (Brandstetter et al., 2022), iterative refinement (Lippe et al., 2023), and multi-step rollout objectives. Forward rollout and backward gradient propagation are themselves distinct choices: List et al. (2025) explicitly compare differentiable unrolling with forward unrolling without temporal gradients, while List et al. (2022) truncate gradient propagation within longer differentiable solver rollouts. APEBench further systematizes autoregressive PDE-emulator training and gradient cuts within unrolled objectives (Koehler et al., 2024). We therefore do not treat self-fed exposure or truncated temporal credit as new. Instead, we separate *forward exposure*—how many self-fed states are encountered during training—from *temporal credit*—how far gradients propagate through that rollout—and study both together with the temporal representation being composed.

Generators and finite-time transitions. Both instantaneous generators and finite-time evolution maps are established representations of dynamical systems. Flow-map methods directly approximate finite-time evolution (Churchill & Xiu, 2023; Chen & Wu, 2023), while structured discrete-time models can additionally preserve properties such as symplecticity (Chen & Tao, 2021). Recent work has compared direct next-state prediction with learning an instantaneous derivative followed by numerical integration, finding substantial representation-dependent differences in accuracy and stability (Zhou & Barati Farimani, 2025; Ainslie et al., 2026). Thus finite-time transition learning itself is not our contribution. Our focus is the effect of this representational choice under autoregressive composition: we evaluate generator and finite-time-transition formulations across a heterogeneous collection of physical systems from The Well, and separately isolate representation on periodic Euler under matched self-fed exposure and matched temporal credit. This broad cross-physics comparison complements prior studies conducted in more restricted dynamical settings.

Local error, transport, and long-horizon behavior. The distinction between local defects and accumulated global error is classical in numerical analysis (Hairer et al., 1993; Cao & Petzold, 2004). Closely related mechanisms have recently been made explicit for learned dynamics. Zhang & Jo (2026) derive an exact path-integral error recursion separating local discrepancy injection from its transport through subsequent learned dynamics, together with a linearized multi-step expansion involving ordered Jacobian products. Ainslie et al. (2026) likewise relate autoregressive stability to the Jacobian structure of the learned transition, while other work studies long-horizon neural-operator stability (McCabe et al., 2023) and transient amplification under products of learned Jacobians (Pervez & Locatello, 2026). We therefore do not claim the local-to-global distinction or Jacobian-product error transport as new. Our formulation collects the full residual sequence into a finite-horizon transport operator and pulls the rollout loss back to residual space, making the effects of residual direction, timing, and cross-time interaction explicit. We use this perspective to organize and experimentally separate three design choices—temporal representation, forward exposure, and temporal credit—and test their consequences across diverse physical dynamics.

3 The Geometry of Rollout Error

One-step prediction error does not determine rollout error. Two learned transitions can have identical one-step error and nevertheless produce arbitrarily different errors after repeated composition (Proposition 8.1). The missing information is how local residuals are transported and interact under the subsequent learned dynamics.

3.1 Exact residual transport

Let

$$\boldsymbol{x}_{t+1} = \Phi_\Delta(\boldsymbol{x}_t), \qquad \hat{\boldsymbol{x}}_{t+1} = T_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t), \qquad \hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0, \tag{2}$$

and define rollout error and local residual as

$$\boldsymbol{e}_t = \hat{\boldsymbol{x}}_t - \boldsymbol{x}_t, \qquad \boldsymbol{r}_t = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \Phi_\Delta(\boldsymbol{x}_t). \tag{3}$$

Thus $\boldsymbol{r}_t$ is the error the model would make if evaluated on the correct state, whereas rollout error additionally reflects errors inherited from earlier predictions.

Define the path-integrated Jacobian

$$\overline{\boldsymbol{J}}_t = \int_0^1 DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t)\, ds. \tag{4}$$

Then

$$\boldsymbol{e}_{t+1} = \overline{\boldsymbol{J}}_t \boldsymbol{e}_t + \boldsymbol{r}_t, \tag{5}$$

and repeated substitution gives the exact decomposition

$$\boxed{\boldsymbol{e}_H = \sum_{t=0}^{H-1} \overline{\boldsymbol{J}}_{H-1:t+1} \boldsymbol{r}_t}. \tag{6}$$

This residual-transport identity separates the errors a model injects locally from the dynamics through which those errors subsequently propagate. The long-horizon consequence of a local residual therefore depends not only on its magnitude, but also on its direction, when it occurs, and the finite-time dynamics it encounters.

3.2 Local rollout geometry

The exact transport identity is nonlinear because its path-integrated Jacobians depend on the realized rollout error. To expose the local geometry near the reference trajectory, define

$$\boldsymbol{J}_t = DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t) \tag{7}$$

and the residual scale

$$\rho_H = \left(\sum_{t=0}^{H-1} \|\boldsymbol{r}_t\|_2^2 \right)^{1/2}. \tag{8}$$

For fixed H, the rollout error admits the first-order expansion

$$\boldsymbol{e}_H = \sum_{t=0}^{H-1} \boldsymbol{J}_{H-1:t+1} \boldsymbol{r}_t + O(\rho_H^2). \tag{9}$$

Stacking the residuals and the corresponding linearized rollout errors gives

$$\boldsymbol{\xi}_{1:H} = \boldsymbol{T}_H \boldsymbol{r}_{0:H-1}, \tag{10}$$

where $\boldsymbol{T}_H$ is the block lower-triangular residual-transport operator whose (h, t) block contains the Jacobian product transporting residual $\boldsymbol{r}_t$ to horizon h. Its structure captures the temporal asymmetry of rollout error: early residuals are transported through more subsequent dynamics than late ones.

For a weighted squared rollout objective with $\boldsymbol{W} = \mathrm{diag}(w_1 \boldsymbol{I}, \ldots, w_H \boldsymbol{I})$, define

$$\boxed{\boldsymbol{M}_{H,\boldsymbol{W}} = \boldsymbol{T}_H^\top \boldsymbol{W} \boldsymbol{T}_H}. \tag{11}$$

The off-diagonal temporal blocks of $\boldsymbol{M}_{H,\boldsymbol{W}}$ capture constructive or destructive interactions between residuals introduced at different times after transport to common future horizons.

Then, locally,

$$\boxed{\mathcal{L}_H^{\mathrm{sq}} = \boldsymbol{r}_{0:H-1}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{r}_{0:H-1} + O(\rho_H^3)}. \tag{12}$$

We call $\boldsymbol{M}_{H,\boldsymbol{W}}$ the *finite-horizon rollout geometry*. A one-step squared-error objective penalizes residual magnitude in the ambient state-space geometry. Finite-horizon rollout instead induces a dynamics-dependent quadratic geometry on the entire residual sequence. Errors of equal local magnitude can therefore incur different trajectory costs depending on when they occur and how they align with directions transported by the subsequent dynamics. This distinction provides the theoretical basis for studying models that are similar under local prediction metrics but behave differently under repeated composition.

4 Learning the Finite-Time Transition

Section 3 shows that rollout error depends not only on local prediction residuals, but on how those residuals are transported under repeated composition. This motivates separating three choices that are often coupled in autoregressive training: *temporal representation*, the dynamical object being learned; *forward exposure*, how far the model is rolled out on its own predictions during training; and *temporal credit*, how far gradients propagate through that rollout.

4.1 Temporal representation

A generator model learns an instantaneous rate,

$$G_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx \dot{\boldsymbol{x}}, \qquad T_{\boldsymbol{\theta}}^{\mathrm{gen}}(\boldsymbol{x}) = \mathrm{Integrate}(G_{\boldsymbol{\theta}}, \boldsymbol{x}, \Delta), \tag{13}$$

whereas a transition model directly learns the finite-time map,

$$T_{\boldsymbol{\theta}}^{\mathrm{trans}}(\boldsymbol{x}) \approx \Phi_{\Delta}(\boldsymbol{x}). \tag{14}$$

Both ultimately define a transition that is repeatedly composed at deployment, but they optimize different temporal objects: an instantaneous generator whose flow must be numerically integrated, or the finite-time transition that will itself be composed during rollout.

4.2 Forward exposure and temporal credit

Starting from a ground-truth anchor $\boldsymbol{x}_t$, we generate an H-step self-fed rollout,

$$\hat{\boldsymbol{x}}_{t|t} = \boldsymbol{x}_t, \qquad \hat{\boldsymbol{x}}_{t+h|t} = T_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{x}}_{t+h-1|t}\right), \quad h = 1, \dots, H. \tag{15}$$

The quadratic objective in Section 3 is used only to characterize the local rollout geometry analytically. For training, we use channel-normalized RMS error:

$$\mathcal{L}_H^{\mathrm{train}}(\boldsymbol{\theta}) = \ell\left(\hat{\boldsymbol{x}}_{t+1|t}, \boldsymbol{x}_{t+1}\right) + \frac{\lambda_{\mathrm{roll}}}{H} \sum_{h=1}^{H} \ell\left(\hat{\boldsymbol{x}}_{t+h|t}, \boldsymbol{x}_{t+h}\right), \tag{16}$$

with $\lambda_{\mathrm{roll}} = 0.5$.

Crucially, we distinguish the *forward exposure horizon* H from the *credit depth* k. All H states are generated autoregressively and contribute to the loss regardless of k; changing k alters only the backward computational graph. For the $H - 1$ inter-step connections, gradients are retained according to

$$b_j = \mathbf{1}\{j > H - k\}, \qquad j = 1, \dots, H - 1, \tag{17}$$

with the remaining connections detached. Thus $k = 1$ gives stepwise credit, $k = H$ gives full backpropagation through the rollout, and intermediate values retain a contiguous trailing credit window while leaving the forward trajectory unchanged.

Our primary experiments use $H = 8$ and $k = 4$, held fixed across physical systems without per-system tuning. Separate ablations vary k while keeping the forward horizon $H = 8$ fixed, allowing temporal credit to be studied without changing the states encountered during training. These choices are empirical design choices rather than consequences of the rollout geometry in Section 3.

5 Experiments

We test three consequences of the rollout-error geometry developed in Section 3. First, does learning the finite-time transition that is actually composed at deployment improve long-horizon prediction across heterogeneous physical systems? Second, does this advantage remain when generator and transition models receive the same forward exposure and temporal credit? Third, how does the depth of temporal credit affect long-horizon behavior?

5.1 Setup

We evaluate generator-based DISCO and DISCO-FLO on ten evaluation conditions spanning nine physical systems from The Well: active matter, Gray–Scott, MHD, Rayleigh–Bénard, shear flow, Euler with periodic and open boundary conditions, radiative layer, Helmholtz, and gravity cooling. Our primary DISCO-FLO configuration uses an $H = 8$ self-fed training horizon with credit depth $k = 4$, held fixed across systems without per-system tuning.

All models are evaluated autoregressively: after the initial context, each prediction is fed back as input to the next step. We report normalized RMSE (NRMSE). Because several systems develop heavy-tailed rollout failures, the primary rollout curves show median error across evaluation trajectories; we also report mean error, maximum error, and failure-tail statistics where relevant. Helmholtz and gravity cooling are evaluated through $t + 40$ because of trajectory length; the remaining conditions are evaluated through $t + 64$.

We additionally train an autoregressive Fourier Neural Operator (FNO) under the same data split, normalization, rollout protocol, and evaluation metric. Because FNO differs substantially from

Table 1: **Local accuracy does not determine long-horizon performance.** Median NRMSE at one step and the final evaluation horizon. Bold indicates the lower error within each DISCO/DISCO-FLO pair. Final horizon is $t + 64$ except for Helmholtz and gravity cooling ($t + 40$).

	$t+1$		Final horizon	
System	DISCO	FLO	DISCO	FLO
Active matter	**0.155**	0.179	1660	**1.19**
Euler (open)	**0.0704**	0.0761	222	**4.68**
Euler (periodic)	**0.0877**	0.0935	217	**0.763**
Gravity cooling	0.158	**0.144**	6.04	**1.49**
Gray–Scott	0.00318	**0.00262**	0.212	**0.163**
Helmholtz	**0.00723**	0.00990	1.07	**0.375**
MHD	0.306	**0.290**	171	**1.16**
Radiative layer	0.114	**0.0809**	37355	**0.841**
Rayleigh–Bénard	**0.0672**	0.102	3.84	**0.771**
Shear flow	**0.0168**	0.0204	0.710	**0.271**

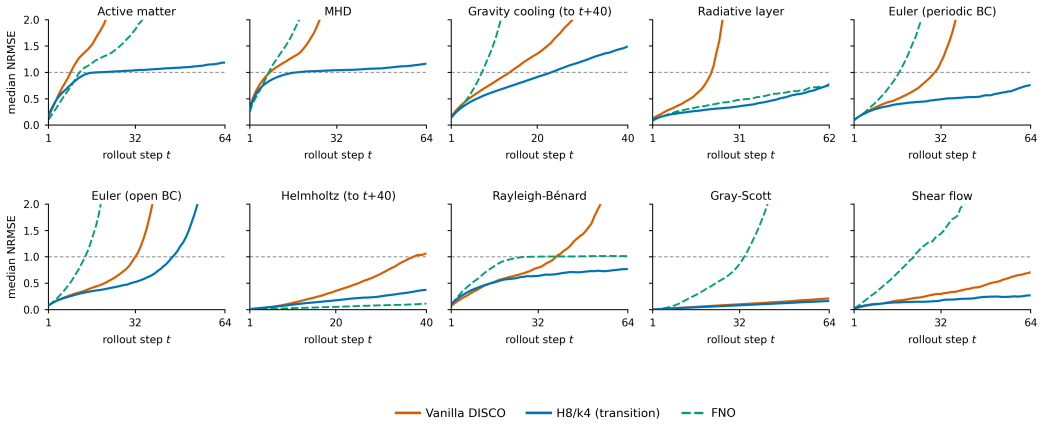

Figure 1: **Long-horizon rollout across heterogeneous physical systems.** Median NRMSE under autoregressive self-fed deployment for vanilla DISCO (instantaneous generator followed by numerical integration), DISCO-FLO with $H = 8, k = 4$, and an autoregressive FNO reference baseline. The dashed horizontal line marks NRMSE $= 1$. Helmholtz and gravity cooling are evaluated through $t + 40$ because of trajectory length.

DISCO in architecture and capacity, we use it only as an architecture-independent reference. Claims about temporal representation are based on the controlled DISCO versus DISCO-FLO comparisons below.

5.2 Finite-time transition learning improves rollout across physics

Figure 1 shows the central cross-physics result. DISCO-FLO has lower final-horizon median NRMSE than generator-based DISCO in all ten evaluation conditions, despite having lower one-step median error in only four of ten. Thus the long-horizon advantage is not simply inherited from uniformly better local prediction. The effect ranges from modest improvements on systems such as Gray–Scott to orders-of-magnitude differences on active matter, MHD, periodic Euler, and the radiative layer.

Table 1 makes this dissociation explicit: DISCO-FLO has lower one-step median error in only four of ten conditions, but lower final-horizon median error in all ten.

The FNO reference provides additional context: it avoids several catastrophic failures of generator DISCO but also develops substantial autoregressive error on multiple systems. DISCO-FLO has

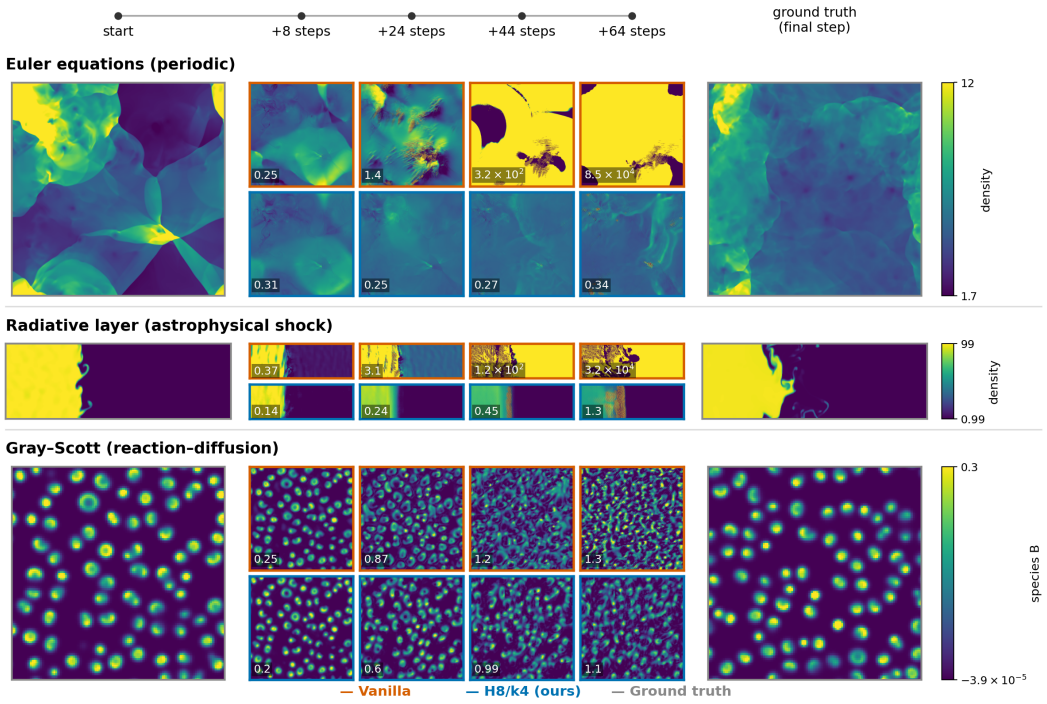

Figure 2: **Small local differences can produce qualitatively different trajectories under composition.** Representative self-fed rollouts for periodic Euler, radiative-layer dynamics, and Gray–Scott. Orange-bordered panels show generator DISCO and blue-bordered panels show DISCO-FLO with $H = 8, k = 4$. Insets report relative L_2 error for the displayed field, computed independently of the aggregate NRMSE in Figure 1. Ground truth and predictions share a color scale within each physical system. Euler and the radiative layer illustrate catastrophic divergence, while Gray–Scott shows the same long-horizon ordering without numerical blow-up.

the lowest final-horizon median error on eight of ten conditions, with FNO lower on Helmholtz and slightly lower on the radiative layer. Because this comparison is not architecture-controlled, we draw no conclusions about temporal representation from it.

The state-space examples in Figure 2 make these different rollout regimes concrete. In periodic Euler and the radiative layer, generator DISCO eventually enters a catastrophic regime while DISCO-FLO remains substantially closer to the target trajectory. Gray–Scott provides a bounded contrast: both models remain numerically stable, but DISCO-FLO accumulates error more slowly.

5.3 The representation effect survives matched temporal credit

The cross-physics comparison changes both temporal representation and training: the primary DISCO-FLO model is trained through self-fed rollout, whereas vanilla DISCO uses a local objective. To isolate temporal representation, we therefore repeat the periodic-Euler experiment with matched $H = 8$ forward exposure and matched temporal credit for generator DISCO and DISCO-FLO.

At $k = 4$, the two models are almost indistinguishable over the horizon on which they are trained: mean NRMSE is 0.107 versus 0.108 at $t+1$ and 0.323 versus 0.320 at $t+8$ for generator DISCO and DISCO-FLO, respectively. Beyond the training horizon, however, their behavior separates sharply. At $t + 64$, median NRMSE is 1.16 for the generator and 0.76 for DISCO-FLO; mean NRMSE is 5.68 versus 0.94, and the maximum observed error falls from 799 to 24.

The same ordering appears with shallow credit: DISCO-FLO with $k = 1$ outperforms generator DISCO with $k = 1$ at long horizon and also outperforms the $k = 4$ generator. The transition advantage therefore cannot be explained simply by providing DISCO-FLO with deeper multi-step

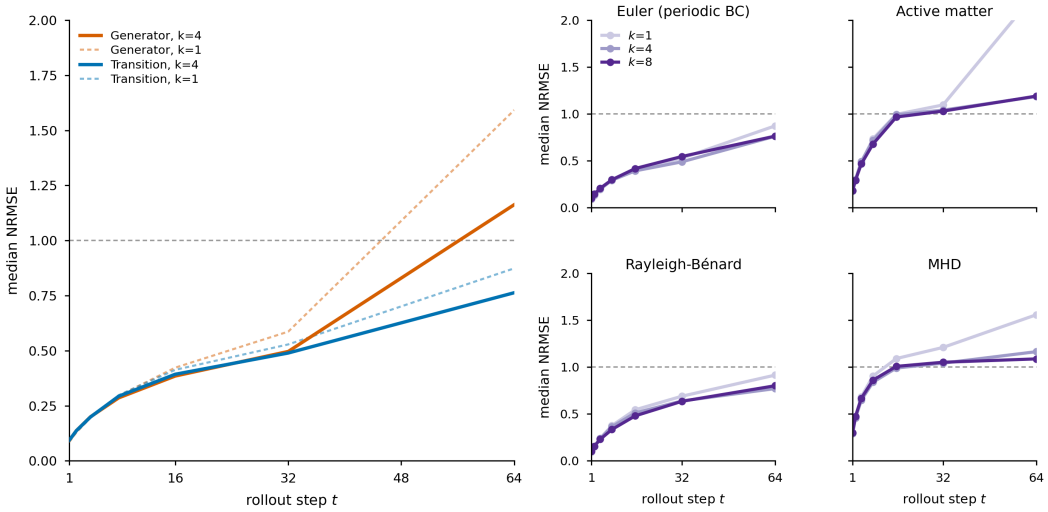

Figure 3: **Temporal representation and rollout credit are distinct axes. Left:** Generator versus finite-time-transition learning on periodic Euler with matched $H = 8$ rollout training and matched credit depth. Transition learning substantially improves long-horizon behavior even when the generator receives the same temporal credit. **Right:** Credit-depth sweeps within the transition architecture across four physical systems. Increasing credit from $k = 1$ to $k = 4$ consistently improves long-horizon rollout, while the ordering of $k = 4$ and full $k = 8$ credit depends on the physical system.

gradients. When forward exposure and temporal credit are matched, changing the temporal object being learned still substantially changes behavior under repeated composition.

5.4 Temporal credit is a distinct, system-dependent axis

We next hold the transition representation and forward horizon $H = 8$ fixed and vary only the credit depth k. Figure 3B shows the resulting $k \in \{1, 4, 8\}$ sweep on periodic Euler, active matter, Rayleigh–Bénard, and MHD.

Increasing credit from $k = 1$ to $k = 4$ improves the final-horizon result on all four systems. Increasing further to full $k = 8$ credit, however, has a system-dependent effect: $k = 4$ and $k = 8$ are nearly indistinguishable on periodic Euler and Rayleigh–Bénard, $k = 4$ performs better on active matter, and $k = 8$ performs best on MHD.

Forward exposure and temporal credit are therefore distinct design axes, and there is no single credit depth that dominates across dynamics. In particular, rollout training cannot be summarized as either "more BPTT is always better" or "truncated BPTT is always better." The useful depth of temporal credit depends on the dynamics being learned.

6 Limitations

Our experiments show a consistent long-horizon advantage for finite-time transition learning across the systems studied here, but do not imply that generator models are universally inferior. The magnitude of the gain varies substantially across physics, and the best temporal credit depth among those tested is not universal. Moreover, the matched-credit representation experiment is performed on periodic Euler, while the credit-depth sweep spans four physical systems.

More fundamentally, improved rollout error should not be conflated with physically faithful simulation. In several systems, DISCO-FLO replaces catastrophic divergence with substantially more stable trajectories, but the resulting predictions can still accumulate considerable error over long horizons. Our experiments evaluate predictive trajectory error rather than whether learned dynam-

ics preserve physical invariants, remain on the correct attractor, or reproduce the long-run statistics of the underlying system. Avoiding numerical or autoregressive failure is therefore only one step toward learning physically faithful dynamics.

7 Conclusion

The central difficulty of long-horizon prediction is not simply that local errors accumulate. Local residuals are transported, redirected, reinforced, or suppressed by the subsequent learned dynamics. The resulting trajectory error therefore depends jointly on the errors a model makes and the finite-time geometry through which those errors propagate. This explains why models with similar one-step accuracy can behave very differently under repeated self-fed composition.

Our experiments vary two of these axes while controlling the third. First, temporal representation changes the dynamical object being composed. Across ten evaluation conditions spanning nine physical systems, DISCO-FLO has lower final-horizon median error than generator-based DISCO in every condition, despite having lower one-step median error in only four of ten. A matched-credit experiment shows that this separation persists when generator and transition models receive the same forward exposure and temporal credit. Second, temporal credit changes how the deployed composition is optimized: shallow credit is consistently worse in our sweep, while the relative value of intermediate and full backpropagation depends on the dynamics.

These results support a view of rollout learning in which local prediction accuracy is only one component of the problem. The residuals produced by a model, the finite-time dynamics that transport them, and the training signals used to shape that composition together determine long-horizon behavior. Temporal representation, forward exposure, and temporal credit should therefore be treated as distinct design choices for controlling the geometry of rollout error.

DISCO-FLO does not solve long-horizon physical simulation: stable rollout is not the same as physically correct rollout. Rather, it provides evidence for a broader principle. When a learned dynamical model will be deployed through repeated composition, the object and objective used for learning should be chosen with that finite-time composition—and the geometry of error it induces—explicitly in view.

AI Assistance

Generative AI tools were used during the development of this work to assist with research ideation and experimental design, literature retrieval and discovery, drafting and editing portions of the manuscript, and the presentation and writing of mathematical arguments and proofs. AI-assisted suggestions were used as inputs to the authors' research and writing process; experimental results, mathematical claims, citations, and manuscript content were reviewed and verified by the authors. Generative AI was not used to generate synthetic datasets. The authors take full responsibility for the final content of the paper.

References

Conrad Ainslie, Pedram Hassanzadeh, Michael W. Mahoney, and Ashesh Chattopadhyay. Eigen-analysis framework for autoregressive neural emulators of multi-scale chaotic dynamics, 2026.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural pde solvers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vSix3HPYKSU.

Yang Cao and Linda Petzold. A posteriori error estimation and global error control for ordinary differential equations by the adjoint method. *SIAM Journal on Scientific Computing*, 26(2):359–374, 2004. doi: 10.1137/S1064827503420969.

Junfeng Chen and Kailiang Wu. Deep-OSG: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493:112498, 2023. doi: 10.1016/j.jcp.2023.112498. arXiv:2302.03358.

Renyi Chen and Molei Tao. Data-driven prediction of general hamiltonian dynamics via learning exactly-symplectic maps. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1717–1727. PMLR, 2021. URL https://proceedings.mlr.press/v139/chen21r.html.

Victor Churchill and Dongbin Xiu. Flow map learning for unknown dynamical systems: Overview, implementation, and benchmarks. *Journal of Machine Learning for Modeling and Computing*, 2023. arXiv:2307.11013.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, Berlin, Heidelberg, 2 edition, 1993. doi: 10.1007/978-3-540-78862-1.

George Haller. Lagrangian coherent structures. *Annual Review of Fluid Mechanics*, 47:137–162, 2015. doi: 10.1146/annurev-fluid-010313-141322.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120252–120310, 2024. doi: 10.52202/079017-3822.

John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, 2 edition, 2018. doi: 10.1007/978-3-319-91755-9.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural pde solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023. doi: 10.52202/075280-2946. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d529b943af3dba734f8a7d49efcb6d09-Abstract-Conference.html.

Bjoern List, Li-Wei Chen, Kartik Bali, and Nils Thuerey. Differentiability in unrolled training of neural physics simulators on transient dynamics. *Computer Methods in Applied Mechanics and Engineering*, 433:117441, 2025. doi: 10.1016/j.cma.2024.117441. arXiv:2402.12971.

Björn List, Li-Wei Chen, and Nils Thuerey. Learned turbulence modelling with differentiable fluid solvers: Physics-based loss functions and optimisation horizons. *Journal of Fluid Mechanics*, 949:A25, 2022. doi: 10.1017/jfm.2022.738.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023.

Adeel Pervez and Francesco Locatello. Controlling transient amplification improves long-horizon rollouts, 2026.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468. PMLR, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Shawn C. Shadden, Francois Lekien, and Jerrold E. Marsden. Definition and properties of lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. *Physica D: Nonlinear Phenomena*, 212(3–4):271–304, 2005. doi: 10.1016/j.physd.2005.10.007.

Lloyd N. Trefethen, Anne E. Trefethen, Satish C. Reddy, and Tobin A. Driscoll. Hydrodynamic stability without eigenvalues. *Science*, 261(5121):578–584, 1993. doi: 10.1126/science.261.5121.578.

Jingyi Zhang and Gwanghyun Jo. Analysis of error propagation in autoencoder-based reduced-order neural ordinary differential equations, 2026.

Anthony Zhou and Amir Barati Farimani. Predicting change, not states: An alternate framework for neural PDE surrogates. *Computer Methods in Applied Mechanics and Engineering*, 441:117990, 2025. doi: 10.1016/j.cma.2025.117990. arXiv:2412.13074.

8 PROOFS AND ADDITIONAL THEORY

The results below formalize the local-to-global error mechanism introduced in section 3.

8.1 NOTATION AND ASSUMPTIONS

Let

$$\Phi_\Delta : \mathbb{X} \to \mathbb{X}, \qquad T_{\boldsymbol{\theta}} : \mathbb{X} \to \mathbb{X}, \qquad \mathbb{X} \subseteq \mathbb{R}^d,$$

denote the exact evolution over one observation interval and the learned transition.

Starting from the same initial condition,

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0,$$

we define

$$\boldsymbol{x}_{t+1} = \Phi_\Delta(\boldsymbol{x}_t), \qquad \hat{\boldsymbol{x}}_{t+1} = T_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t). \tag{18}$$

The rollout error and local residual are defined by

$$\boldsymbol{e}_t = \hat{\boldsymbol{x}}_t - \boldsymbol{x}_t, \qquad \boldsymbol{r}_t = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \Phi_\Delta(\boldsymbol{x}_t). \tag{19}$$

For any time-indexed matrix sequence $\{\boldsymbol{A}_t\}$, we use

$$\boldsymbol{A}_{b:a} = \boldsymbol{A}_b \boldsymbol{A}_{b-1} \cdots \boldsymbol{A}_a, \qquad b \geq a,$$

with the convention that an empty product is $\boldsymbol{I}$. We use the same notation for the Jacobian sequences $\boldsymbol{J}_t$, $\overline{\boldsymbol{J}}_t$, and $\widehat{\boldsymbol{J}}_t$.

Unless otherwise stated, $\|\cdot\|_2$ denotes the Euclidean norm for vectors and its induced spectral norm for matrices.

For the higher-order results, assume that $T_{\boldsymbol{\theta}}$ is twice continuously differentiable on a neighborhood $\mathcal{U}$ containing all line segments joining $\boldsymbol{x}_t$ and $\hat{\boldsymbol{x}}_t$ through the horizon of interest. We assume there exist finite constants L, β such that

$$\sup_{\boldsymbol{x} \in \mathcal{U}} \|DT_{\boldsymbol{\theta}}(\boldsymbol{x})\|_2 \leq L \tag{20}$$

and

$$\sup_{\boldsymbol{x} \in \mathcal{U}} \left\|D^2 T_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|_{\mathrm{op}} \leq \beta, \tag{21}$$

where

$$\left\|D^2 T_{\boldsymbol{\theta}}(\boldsymbol{x})\right\|_{\mathrm{op}} = \sup_{\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2 = 1} \left\|D^2 T_{\boldsymbol{\theta}}(\boldsymbol{x})[\boldsymbol{a}, \boldsymbol{b}]\right\|_2 .$$

8.2 LOCAL ERROR DOES NOT DETERMINE FINITE-HORIZON ERROR

For a probability distribution μ over initial states, define

$$\mathcal{E}_1(T) = \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\|T(\boldsymbol{x}) - \Phi_\Delta(\boldsymbol{x})\|_2^2\right] \tag{22}$$

and

$$\mathcal{E}_H(T) = \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\|T^H(\boldsymbol{x}) - \Phi_{H\Delta}(\boldsymbol{x})\|_2^2\right]. \tag{23}$$

Proposition 8.1 (Separation of local and finite-horizon error)**.** *For every fixed $H \geq 2$ and every $R > 0$, there exist two linear transitions $T_1, T_2 : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying*

$$\mathcal{E}_1(T_1) = \mathcal{E}_1(T_2),$$

but

$$\frac{\mathcal{E}_H(T_1)}{\mathcal{E}_H(T_2)} > R.$$

Proof. Let the true dynamics be the identity,

$$\Phi_\Delta(\boldsymbol{x}) = \boldsymbol{x},$$

and let

$$\mu = \delta_{\boldsymbol{x}_0}, \qquad \boldsymbol{x}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Fix $\varepsilon > 0$ and, for $K > 0$, define

$$\boldsymbol{A}_K = \begin{pmatrix} 1 & 0 \\ \varepsilon & K \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ \varepsilon & 0 \end{pmatrix},$$

with

$$T_1(\boldsymbol{x}) = \boldsymbol{A}_K \boldsymbol{x}, \qquad T_2(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{x}.$$

At $\boldsymbol{x}_0$,

$$T_1(\boldsymbol{x}_0) = T_2(\boldsymbol{x}_0) = \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix},$$

and hence

$$\mathcal{E}_1(T_1) = \mathcal{E}_1(T_2) = \varepsilon^2.$$

For T_1, write

$$T_1^h(\boldsymbol{x}_0) = \begin{pmatrix} 1 \\ y_h \end{pmatrix}.$$

The second coordinate obeys

$$y_{h+1} = \varepsilon + K y_h, \qquad y_0 = 0,$$

so

$$y_H = \varepsilon \sum_{j=0}^{H-1} K^j.$$

Therefore

$$\mathcal{E}_H(T_1) = \varepsilon^2 \left(\sum_{j=0}^{H-1} K^j \right)^2. \tag{24}$$

For T_2,

$$T_2^h(\boldsymbol{x}_0) = \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} \qquad \text{for all } h \geq 1,$$

and thus

$$\mathcal{E}_H(T_2) = \varepsilon^2. \tag{25}$$

Combining Equation 24 and Equation 25,

$$\frac{\mathcal{E}_H(T_1)}{\mathcal{E}_H(T_2)} = \left(\sum_{j=0}^{H-1} K^j \right)^2.$$

For every fixed $H \geq 2$, this ratio diverges as $K \to \infty$. Thus K can be chosen so that the ratio exceeds any prescribed R.

□

Proposition 8.1 is deliberately elementary: equal local prediction error alone imposes no ordering on finite-horizon error. This isolates the distinction tested empirically in the main paper; namely, matching local prediction error does not control the behavior of repeatedly composed transitions.

8.3 One-step sensitivity does not determine finite-time sensitivity

The empirical diagnostic in the main paper concerns the spectral norm of a composed tangent map. Finite-time sensitivity is classically characterized through the derivative of the finite-time flow map and its largest singular value (Shadden et al., 2005; Haller, 2015); by the chain rule, this derivative is the composition of the intermediate tangent maps. Transient amplification can likewise be invisible to eigenvalue-based asymptotic stability analysis (Trefethen et al., 1993).

Proposition 8.2 (Equal one-step sensitivity can yield different finite-time sensitivity)**.** *For every $a > 0$ and $H \geq 2$, there exist two linear transitions $T_A, T_B : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying*

$$\|DT_A\|_2 = \|DT_B\|_2 = a,$$

but

$$\|D(T_A^H)\|_2 = a^H, \qquad \|D(T_B^H)\|_2 = 0.$$

Proof. Let

$$\boldsymbol{A} = a \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \boldsymbol{B} = a \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

and define

$$T_A(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \qquad T_B(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{x}.$$

Both matrices have largest singular value a, so

$$\|\boldsymbol{A}\|_2 = \|\boldsymbol{B}\|_2 = a.$$

However,

$$\boldsymbol{A}^H = a^H \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

whereas

$$\boldsymbol{B}^2 = \boldsymbol{0}_{2\times 2},$$

and therefore

$$\boldsymbol{B}^H = \boldsymbol{0}_{2\times 2} \qquad \text{for every } H \geq 2.$$

Hence

$$\|D(T_A^H)\|_2 = \|\boldsymbol{A}^H\|_2 = a^H$$

and

$$\|D(T_B^H)\|_2 = \|\boldsymbol{B}^H\|_2 = 0.$$

□

From this one can see that one-step sensitivity does not determine finite-time sensitivity, even for autonomous linear transitions. The composed-tangent diagnostic A_H is thus not able to be reduced to a one-step Jacobian norm.

8.4 Exact finite-horizon residual transport

Theorem 8.3 (Exact residual transport)**.** *Assume $T_{\boldsymbol{\theta}} \in C^1(\mathcal{U})$ and that, for each $t = 0, \ldots, H-1$, the line segment*

$$\{\boldsymbol{x}_t + s\boldsymbol{e}_t : s \in [0,1]\} \subset \mathcal{U}.$$

Define

$$\overline{\boldsymbol{J}}_t = \int_0^1 DT_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t + s\boldsymbol{e}_t\right) ds. \tag{26}$$

Then

$$\boldsymbol{e}_{t+1} = \overline{\boldsymbol{J}}_t \boldsymbol{e}_t + \boldsymbol{r}_t. \tag{27}$$

With the convention that

$$\overline{\boldsymbol{J}}_{a-1:a} = \boldsymbol{I}$$

for an empty product, and since $\boldsymbol{e}_0 = \mathbf{0}$*,*

$$\boldsymbol{e}_H = \sum_{t=0}^{H-1} \overline{\boldsymbol{J}}_{H-1:t+1} \boldsymbol{r}_t. \tag{28}$$

Proof. From the definitions of the exact and learned trajectories and the rollout error,

$$\boldsymbol{e}_{t+1} = T_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t) - \Phi_\Delta(\boldsymbol{x}_t).$$

Since

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t + \boldsymbol{e}_t$$

and

$$\boldsymbol{r}_t = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) - \Phi_\Delta(\boldsymbol{x}_t),$$

we obtain

$$\boldsymbol{e}_{t+1} = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t + \boldsymbol{e}_t) - T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) + \boldsymbol{r}_t. \tag{29}$$

Define

$$\boldsymbol{g}_t(s) = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t), \qquad s \in [0, 1].$$

By the assumed inclusion of the line segment in $\mathcal{U}$ and $T_{\boldsymbol{\theta}} \in C^1(\mathcal{U})$, the map $\boldsymbol{g}_t$ is continuously differentiable. By the chain rule,

$$\frac{d}{ds}\boldsymbol{g}_t(s) = DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t)\boldsymbol{e}_t.$$

Therefore, by the fundamental theorem of calculus,

$$T_{\boldsymbol{\theta}}(\boldsymbol{x}_t + \boldsymbol{e}_t) - T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) = \int_0^1 DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t)\boldsymbol{e}_t \, ds \tag{30}$$

$$= \left(\int_0^1 DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t) \, ds \right) \boldsymbol{e}_t \tag{31}$$

$$= \overline{\boldsymbol{J}}_t \boldsymbol{e}_t. \tag{32}$$

Substituting this into Equation 29 gives

$$\boldsymbol{e}_{t+1} = \overline{\boldsymbol{J}}_t \boldsymbol{e}_t + \boldsymbol{r}_t,$$

which proves Equation 27.

Repeated substitution yields

$$\boldsymbol{e}_H = \overline{\boldsymbol{J}}_{H-1}\boldsymbol{e}_{H-1} + \boldsymbol{r}_{H-1} \tag{33}$$

$$= \overline{\boldsymbol{J}}_{H-1}\left(\overline{\boldsymbol{J}}_{H-2}\boldsymbol{e}_{H-2} + \boldsymbol{r}_{H-2}\right) + \boldsymbol{r}_{H-1} \tag{34}$$

$$= \cdots \tag{35}$$

$$= \overline{\boldsymbol{J}}_{H-1:0}\boldsymbol{e}_0 + \sum_{t=0}^{H-1} \overline{\boldsymbol{J}}_{H-1:t+1}\boldsymbol{r}_t. \tag{36}$$

The term corresponding to $t = H - 1$ uses the empty-product convention $\overline{\boldsymbol{J}}_{H-1:H} = \boldsymbol{I}$. Since the learned and exact rollouts start from the same initial condition,

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0.$$

we have $\boldsymbol{e}_0 = \mathbf{0}$. Hence

$$\boldsymbol{e}_H = \sum_{t=0}^{H-1} \overline{\boldsymbol{J}}_{H-1:t+1}\boldsymbol{r}_t,$$

which proves Equation 28. □

Equation 28 is exact and does not rely on a small-error linearization. It separates rollout error into local residuals and their subsequent transport through the learned dynamics. This can be thought of as being analogous to the classical propagation of local defects into global numerical error (Hairer et al., 1993).

Corollary 8.4 (Amplification-weighted rollout bound)**.** *Under the assumptions of Theorem 8.3,*

$$\|\boldsymbol{e}_H\|_2 \le \sum_{t=0}^{H-1} \left\|\overline{\boldsymbol{J}}_{H-1:t+1}\right\|_2 \|\boldsymbol{r}_t\|_2. \tag{37}$$

Proof. By Equation 28, the triangle inequality, and compatibility of the induced matrix norm,

$$\|\boldsymbol{e}_H\|_2 \le \sum_{t=0}^{H-1} \left\|\overline{\boldsymbol{J}}_{H-1:t+1}\boldsymbol{r}_t\right\|_2 \le \sum_{t=0}^{H-1} \left\|\overline{\boldsymbol{J}}_{H-1:t+1}\right\|_2 \|\boldsymbol{r}_t\|_2.$$

□

Under Equation 20,

$$\|\overline{\boldsymbol{J}}_t\|_2 \le \int_0^1 \|DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t + s\boldsymbol{e}_t)\|_2 \, ds \le L,$$

so submultiplicativity gives

$$\left\|\overline{\boldsymbol{J}}_{H-1:t+1}\right\|_2 \le L^{H-1-t}.$$

Thus

$$\|\boldsymbol{e}_H\|_2 \le \sum_{t=0}^{H-1} L^{H-1-t}\|\boldsymbol{r}_t\|_2. \tag{38}$$

Equation 37 makes the central distinction explicit: rollout error depends on both local residuals $\boldsymbol{r}_t$ and their finite-time transport through the products $\overline{\boldsymbol{J}}_{H-1:t+1}$. The uniform Lipschitz bound in Equation 38 replaces this trajectory-dependent mechanism by the worst-case factor L^{H-1-t}.

8.5 First-order expansion and higher-order terms

Define the reference-trajectory Jacobian

$$\boldsymbol{J}_t = DT_{\boldsymbol{\theta}}(\boldsymbol{x}_t), \tag{39}$$

and residual scale

$$\rho_H = \left(\sum_{t=0}^{H-1} \|\boldsymbol{r}_t\|_2^2\right)^{1/2}. \tag{40}$$

Proposition 8.5 (First-order finite-horizon expansion)**.** *Under the assumptions of Theorem 8.3, Equation 20, and Equation 21, for fixed H,*

$$\boldsymbol{e}_H = \sum_{t=0}^{H-1} \boldsymbol{J}_{H-1:t+1}\boldsymbol{r}_t + \boldsymbol{\zeta}_H, \tag{41}$$

where

$$\|\boldsymbol{\zeta}_H\|_2 = O(\rho_H^2) \qquad \text{as } \rho_H \to 0. \tag{42}$$

Proof. Taylor's theorem gives

$$T_{\boldsymbol{\theta}}(\boldsymbol{x}_t + \boldsymbol{e}_t) = T_{\boldsymbol{\theta}}(\boldsymbol{x}_t) + \boldsymbol{J}_t\boldsymbol{e}_t + \boldsymbol{\eta}_t, \qquad \|\boldsymbol{\eta}_t\|_2 \le \frac{\beta}{2}\|\boldsymbol{e}_t\|_2^2. \tag{43}$$

Hence

$$\boldsymbol{e}_{t+1} = \boldsymbol{J}_t\boldsymbol{e}_t + \boldsymbol{r}_t + \boldsymbol{\eta}_t. \tag{44}$$

By Equation 38 and Cauchy–Schwarz, uniformly for $t \le H$,

$$\|\boldsymbol{e}_t\|_2 \le \sum_{j=0}^{t-1} L^{t-1-j}\|\boldsymbol{r}_j\|_2 \le \left(\sum_{k=0}^{t-1} L^{2k}\right)^{1/2} \rho_H \le C_H \rho_H, \qquad C_H := \left(\sum_{j=0}^{H-1} L^{2j}\right)^{1/2}. \tag{45}$$

Thus

$$\|\boldsymbol{\eta}_t\|_2 \le \frac{\beta}{2} C_H^2 \rho_H^2. \tag{46}$$

Define

$$\boldsymbol{\xi}_0 = \mathbf{0}, \qquad \boldsymbol{\xi}_{t+1} = \boldsymbol{J}_t \boldsymbol{\xi}_t + \boldsymbol{r}_t, \tag{47}$$

so that

$$\boldsymbol{\xi}_H = \sum_{t=0}^{H-1} \boldsymbol{J}_{H-1:t+1} \boldsymbol{r}_t. \tag{48}$$

With

$$\boldsymbol{d}_t = \boldsymbol{e}_t - \boldsymbol{\xi}_t,$$

we have

$$\boldsymbol{d}_{t+1} = \boldsymbol{J}_t \boldsymbol{d}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{d}_0 = \mathbf{0}.$$

Therefore

$$\|\boldsymbol{d}_H\|_2 \le \sum_{t=0}^{H-1} L^{H-1-t}\|\boldsymbol{\eta}_t\|_2 \tag{49}$$

$$\le \frac{\beta}{2} C_H^2 \left(\sum_{t=0}^{H-1} L^{H-1-t}\right) \rho_H^2 = O(\rho_H^2), \tag{50}$$

for fixed H, L, and β.

Setting

$$\boldsymbol{\zeta}_H = \boldsymbol{d}_H$$

and using Equation 48 proves Equation 41–Equation 42.

□

Proposition 8.5 shows that near the reference trajectory, rollout error is governed to leading order by local residuals transported through products of reference-trajectory Jacobians, with nonlinear corrections entering at second order in the residual scale.

8.6 Finite-horizon residual transport

For terminal horizon $h \le H$, define

$$\boldsymbol{c}_{t\to h} := \boldsymbol{J}_{h-1:t+1} \boldsymbol{r}_t, \qquad 0 \le t \le h-1, \tag{51}$$

where the empty product for $t = h-1$ is $\boldsymbol{I}$. Then

$$\boldsymbol{\xi}_h = \sum_{t=0}^{h-1} \boldsymbol{c}_{t\to h}, \tag{52}$$

so $\boldsymbol{c}_{t\to h}$ isolates the first-order contribution of $\boldsymbol{r}_t$ after transport through subsequent reference-trajectory Jacobians.

Stack the residuals and linearized rollout errors as

$$\boldsymbol{r}_{0:H-1} = \begin{pmatrix} \boldsymbol{r}_0 \\ \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_{H-1} \end{pmatrix}, \qquad \boldsymbol{\xi}_{1:H} = \begin{pmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_H \end{pmatrix}, \qquad \boldsymbol{r}_{0:H-1}, \boldsymbol{\xi}_{1:H} \in \mathbb{R}^{Hd}.$$

Then Equation 47 becomes

$$\boldsymbol{\xi}_{1:H} = \boldsymbol{T}_H \boldsymbol{r}_{0:H-1}, \tag{53}$$

where $\boldsymbol{T}_H \in \mathbb{R}^{Hd \times Hd}$ is

$$\boldsymbol{T}_H = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{J}_1 & \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{J}_{2:1} & \boldsymbol{J}_2 & \boldsymbol{I} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{J}_{H-1:1} & \boldsymbol{J}_{H-1:2} & \cdots & \boldsymbol{J}_{H-1} & \boldsymbol{I} \end{pmatrix}, \tag{54}$$

with $d \times d$ identity and zero blocks.

At terminal horizon H,

$$\|\boldsymbol{\xi}_H\|_2^2 = \sum_{t=0}^{H-1} \|\boldsymbol{c}_{t\to H}\|_2^2 + 2 \sum_{0 \le s < t \le H-1} \boldsymbol{c}_{s\to H}^\top \boldsymbol{c}_{t\to H}. \tag{55}$$

The first term contains individual transported-residual energies; the second captures constructive or cancelling cross-time interactions. Thus $\boldsymbol{T}_H$ maps local residuals to their first-order effects across the rollout window. Its block structure exposes the temporal asymmetry: early residuals pass through more future Jacobians, while late residuals pass through fewer, and Equation 55 shows that finite-horizon error depends on both their magnitudes and relative directions.

8.7 Finite-horizon pullback geometry

Let

$$\boldsymbol{W} = \operatorname{diag}(w_1 \boldsymbol{I}, \ldots, w_H \boldsymbol{I}) \in \mathbb{R}^{Hd \times Hd}, \qquad w_h \ge 0, \tag{56}$$

and define

$$\mathcal{L}_H^{\text{lin}} = \sum_{h=1}^{H} w_h \|\boldsymbol{\xi}_h\|_2^2. \tag{57}$$

Theorem 8.6 (Finite-horizon pullback geometry)**.** *The objective in Equation 57 satisfies*

$$\mathcal{L}_H^{\text{lin}} = \boldsymbol{r}_{0:H-1}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{r}_{0:H-1}, \tag{58}$$

where

$$\boldsymbol{M}_{H,\boldsymbol{W}} = \boldsymbol{T}_H^\top \boldsymbol{W} \boldsymbol{T}_H \in \mathbb{R}^{Hd \times Hd}. \tag{59}$$

Moreover,

$$\boldsymbol{M}_{H,\boldsymbol{W}} \succeq 0,$$

and if $w_h > 0$ for every h, then

$$\boldsymbol{M}_{H,\boldsymbol{W}} \succ 0.$$

Proof. Using Equation 53,

$$\mathcal{L}_H^{\text{lin}} = \boldsymbol{\xi}_{1:H}^\top \boldsymbol{W} \boldsymbol{\xi}_{1:H} = \boldsymbol{r}_{0:H-1}^\top \boldsymbol{T}_H^\top \boldsymbol{W} \boldsymbol{T}_H \boldsymbol{r}_{0:H-1},$$

which proves Equation 58.

For any $\boldsymbol{z} \in \mathbb{R}^{Hd}$,

$$\boldsymbol{z}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{z} = (\boldsymbol{T}_H \boldsymbol{z})^\top \boldsymbol{W} (\boldsymbol{T}_H \boldsymbol{z}) \ge 0,$$

so

$$\boldsymbol{M}_{H,\boldsymbol{W}} \succeq 0.$$

If every $w_h > 0$, then $\boldsymbol{W} \succ 0$. Since $\boldsymbol{T}_H$ is block lower triangular with $\boldsymbol{I}$ on its diagonal, it is invertible. Thus for $\boldsymbol{z} \neq \boldsymbol{0}$,

$$\boldsymbol{T}_H \boldsymbol{z} \neq \boldsymbol{0},$$

and therefore

$$\boldsymbol{z}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{z} = (\boldsymbol{T}_H \boldsymbol{z})^\top \boldsymbol{W} (\boldsymbol{T}_H \boldsymbol{z}) > 0.$$

Hence

$$\boldsymbol{M}_{H,\boldsymbol{W}} \succ 0.$$

□

Here we use "pullback" in its standard geometric sense: the quadratic form defined by $\boldsymbol{W}$ on rollout-error space is pulled back through $\boldsymbol{T}_H$ to residual space, giving

$$\boldsymbol{T}_H^\top \boldsymbol{W} \boldsymbol{T}_H$$

(Lee, 2018). For a single perturbation transported to a terminal time with Euclidean weighting, this reduces to the familiar Cauchy–Green form

$$\boldsymbol{J}_{H-1:0}^\top \boldsymbol{J}_{H-1:0},$$

whose quadratic form quantifies finite-time directional stretching (Shadden et al., 2005; Haller, 2015). Equal-magnitude residual directions can therefore incur different rollout costs when transported differently by future dynamics.

8.8 Nonlinear rollout loss

For squared error, define

$$\mathcal{L}_H^{\mathrm{sq}} = \sum_{h=1}^{H} w_h \|\boldsymbol{e}_h\|_2^2. \tag{60}$$

Corollary 8.7 (Local quadratic rollout geometry)**.** *Under the assumptions of Proposition 8.5, for fixed H,*

$$\mathcal{L}_H^{\mathrm{sq}} = \boldsymbol{r}_{0:H-1}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{r}_{0:H-1} + O(\rho_H^3). \tag{61}$$

Proof. Applying Proposition 8.5 at each $h \le H$ gives

$$\boldsymbol{e}_h = \boldsymbol{\xi}_h + \boldsymbol{\zeta}_h, \qquad \|\boldsymbol{\xi}_h\|_2 = O(\rho_H), \qquad \|\boldsymbol{\zeta}_h\|_2 = O(\rho_H^2),$$

uniformly over fixed $h \le H$. Therefore

$$\|\boldsymbol{e}_h\|_2^2 = \|\boldsymbol{\xi}_h\|_2^2 + 2\boldsymbol{\xi}_h^\top \boldsymbol{\zeta}_h + \|\boldsymbol{\zeta}_h\|_2^2 = \|\boldsymbol{\xi}_h\|_2^2 + O(\rho_H^3).$$

Since H and the weights are fixed,

$$\mathcal{L}_H^{\mathrm{sq}} = \mathcal{L}_H^{\mathrm{lin}} + O(\rho_H^3),$$

and Theorem 8.6 gives

$$\mathcal{L}_H^{\mathrm{sq}} = \boldsymbol{r}_{0:H-1}^\top \boldsymbol{M}_{H,\boldsymbol{W}} \boldsymbol{r}_{0:H-1} + O(\rho_H^3),$$

proving Equation 61.

□

Corollary 8.7 is a statement about locality. Near the reference trajectory, the nonlinear rollout objective is quadratic in the residual sequence to leading order. This does not imply that it is globally quadratic: as the predicted trajectory departs from the reference trajectory, the relevant Jacobians and higher-order terms change.

8.9 Experimental Details and Reproducibility

Datasets and preprocessing. We evaluate on ten systems from The Well: `active_matter`, periodic and open-boundary `euler_multi_quadrants`, `turbulence_gravity_cooling`, `gray_scott_reaction_diffusion`, `helmholtz_staircase`, `MHD_64`, `active_matter_radiative_layer_2D`, `rayleigh_benard`, and `shear_flow`. Each model receives $n_{\mathrm{past}} = 6$ context frames and predicts the next frame at the dataset's native temporal stride ($\Delta = 1$), with no additional temporal sub- or super-sampling. Input channels are selected per dataset through an explicit `state_labels` index. Each context window is standardized per sample and channel, with statistics pooled over its temporal and spatial axes; predictions are mapped back using the corresponding input statistics. Periodic systems use circular padding in the operator network, while non-periodic systems use reflect padding together with a one-channel frontier mask.

All reported rollouts are autoregressive. The evaluation horizon is $t{+}64$ for eight systems and $t{+}40$ for Helmholtz and gravity cooling, whose native trajectories are too short to support a 64-step rollout after the six-frame context. The current submission reports rollout metrics on the validation partition. Evaluation uses a fixed seeded DataLoader order with batch size 8 and up to 20 batches per arm (at most 160 trajectories); batches that cannot support the requested horizon or produce non-finite states are skipped and the skip count is logged.

DISCO architectures and temporal parameterizations. Generator-DISCO and DISCO-FLO use the same model backbone: a patch-embedding convolutional encoder followed by 12 spatiotemporal self-attention blocks, pooled into a context latent, and a four-level convolutional U-Net operator. The hypernetwork maps the context latent to the operator parameter vector θ, giving each example its own generated operator parameters through vectorized functional calls. Matched generator/transition pairs have identical learned parameter counts: approximately 122M parameters for the 2D systems and 220M for the volumetric systems.

The two formulations differ in the temporal object represented by this operator. Generator-DISCO interprets the generated operator as an instantaneous vector field and obtains the one-step transition by numerical integration. The primary generator baselines use adaptive-step Bogacki–Shampine integration (`odeint_adjoint`, Bosh3, $\mathrm{rtol} = 5 \times 10^{-6}$, minimum step $1/32$). DISCO-FLO instead applies the generated operator once as a direct finite-time residual map,

$$u_{t+1} = u_t + \Delta_\theta(u_t), \tag{62}$$

with no numerical ODE solver. Although the implementation supports variable-horizon conditioning, all experiments reported here disable it, so DISCO-FLO learns a fixed $\Delta = 1$ transition. In both DISCO formulations the context encoding is computed once at the start of a rollout and held fixed while the predicted physical state is fed back autoregressively.

For the periodic-Euler matched-credit experiment only, the rollout-credit-trained generator uses fixed-step forward Euler with four substeps of size $1/4$ for both training and evaluation. This avoids the substantially larger memory cost of differentiating through the adaptive solver and ensures that each checkpoint is evaluated with the same integration rule used during its training.

FNO reference baseline. The FNO reference is an independent autoregressive Fourier Neural Operator rather than a capacity-matched DISCO variant. The 2D models use four Fourier layers, width 32, and 12×12 retained modes; the volumetric MHD and gravity-cooling models use width 20 and $12 \times 12 \times 8$ retained modes. FNO contains 1.19M parameters on the 2D systems and 3.28M on the volumetric systems, compared with roughly 122–220M for DISCO. We therefore use FNO only as an architecture-independent reference. FNO is trained with Adam at learning rate 10^{-3} and a one-step normalized MSE objective.

Training objective and temporal credit. Generator-DISCO and DISCO-FLO use the same channel-scaled NRMSE loss. Channel reference scales are estimated once from 150 pooled training batches and then held fixed during optimization. For rollout-trained models we use the objective defined in Sec. 4.2: a one-step term plus an H-step self-fed rollout term with $H = 8$ and $\lambda_{\mathrm{roll}} = 0.5$.

Forward exposure and temporal credit are controlled separately. All $H = 8$ predicted states are generated and contribute to the loss for every credit-depth setting. Credit depth k changes only the backward graph: the first $H - k$ inter-step connections are detached and the final $k - 1$ connections remain graph-connected. Thus $k = 1$ gives stepwise credit, while $k = 8$ gives full backpropagation through the eight-step rollout. The primary DISCO-FLO model uses $H = 8, k = 4$ on every physical system without per-system tuning. The controlled periodic-Euler comparison trains generator and transition models with the same H and k, and the credit-depth ablation evaluates $k \in \{1, 4, 8\}$ at fixed $H = 8$.

Optimization and checkpoint selection. DISCO models are optimized with AdamW at learning rate 5×10^{-5}, gradient-norm clipping at 1.0, batch size 8, and gradient accumulation over five batches (effective batch size 40), on a single NVIDIA B200 GPU. The cross-physics runs train for 15 epochs except periodic Euler, whose generator, transition, and matched-credit runs train for 30 epochs. Checkpoints are selected by the lowest mean one-step validation NRMSE for every DISCO arm; FNO checkpoints are selected by validation MSE. The primary results use one trained

Table 2: Primary cross-physics runs. Entries show selected checkpoint epoch / maximum training epoch. DISCO-FLO uses $H = 8, k = 4$ throughout.

System	Generator	DISCO-FLO	FNO	Eval. horizon
`active_matter`	14/15	14/15	15/15	64
`euler_obc`	14/15	14/15	11/15	64
`euler_pbc`	29/30	29/30	30/30	64
`gravity_cooling`	15/15	14/15	7/15	40
`gray_scott`	15/15	15/15	12/15	64
`helmholtz`	15/15	15/15	15/15	40
`mhd`	14/15	14/15	14/15	64
`radiative_layer`	15/15	15/15	14/15	64
`rayleigh_benard`	15/15	15/15	9/15	64
`shear_flow`	14/15	15/15	9/15	64

checkpoint per arm and are not seed-averaged. A separate periodic-Euler DISCO-FLO robustness check was additionally run at seeds 1, 2, and 3.

The surviving experiment-specific configurations do not independently recover every shared default inherited from the original base configuration (in particular the exact global weight-decay and warmup settings), so we omit those values here rather than reconstruct them from an incomplete base file.

Evaluation protocol. All paper rollouts are evaluated autoregressively with the same frozen-context convention described above. The evaluation NRMSE is

$$\mathrm{NRMSE}(y, \hat{y}) = \sqrt{\frac{\mathrm{mean}_{\mathrm{spatial}}[(y - \hat{y})^2]}{10^{-7} + \mathrm{mean}_{\mathrm{spatial}}[y^2]}}. \tag{63}$$

The reduction is over spatial dimensions only, leaving sample- and channel-level values before aggregation. At each rollout step we pool per-sample values across the evaluated trajectories and report the median as the primary curve; we additionally record the mean, maximum, and the fraction of samples with NRMSE > 1. Any axis clipping used for visualization is applied only after metric computation.

Implementation provenance. The DISCO experiments use `train.py`, `train_flow_disco.py`, `train_flow_disco_bptt.py`, `train_disco_bptt.py`, and `eval_horizon_frozen.py`; FNO uses a separate `train_fno.py`/`eval_fno_horizon.py` implementation while sharing the data and evaluation conventions. All checkpoints used in the reported experiments were trained after the decoder checkerboard fix at commit `fffbbdc`, which replaced strided transposed convolution with upsample-plus-convolution in the decoder's upsampling path. Pre-fix and post-fix decoder weights are not checkpoint-compatible.

Table 3: Controlled rollout-credit experiments. All rows use forward exposure $H = 8$. Periodic-Euler generator rows use four-step fixed Euler integration; transition rows use a single learned residual step.

System	Arm	k	Params	Epoch/Max
euler_pbc	Generator	1	122.0M	29/30
	Generator	4	122.0M	29/30
	Transition	1	122.0M	29/30
	Transition	4	122.0M	29/30
	Transition	8	122.0M	29/30
active_matter	Transition	1	122.3M	14/15
	Transition	4	122.3M	14/15
	Transition	8	122.3M	14/15
rayleigh_benard	Transition	1	122.0M	15/15
	Transition	4	122.0M	15/15
	Transition	8	122.0M	15/15
mhd	Transition	1	220.1M	13/15
	Transition	4	220.1M	14/15
	Transition	8	220.1M	13/15