SUPPORT-ONLY HINDCASTING FOR TEST-TIME NEURAL OPERATOR SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time neural operator selection chooses an operator or short composition from a frozen codebook using a short observed trajectory. Existing Neural Operator Splitting (NOS) search ranks candidates by one-step fitting error inside this observed context, whereas deployment exposes the selected candidate to accumulated closed-loop error. We introduce HINDCAST-NOS, a training-free criterion for reranking a frozen NOS candidate beam. Each candidate is restarted from several earlier observed states, rolled forward autoregressively inside the observed prefix, and ranked by its mean error against the corresponding suffixes. The winning candidate is then deployed into the unseen future. We bound future normalized root-mean-square error (NRMSE) over finite horizons using PDE energy estimates for true-flow stability. Experiments reveal that support-only reranking significantly improves conventional NOS across all three PDEs.

1 INTRODUCTION

Neural operators learn maps between functions and provide efficient surrogate solvers for partial differential equations (Li et al., 2020; Lu et al., 2019; Kovachki et al., 2023). Their central deployment challenge is out-of-distribution generalization: a target system may have unseen coefficients or combine physical processes that appeared only separately during training. Broad multi-physics pretraining can reduce this gap (McCabe et al., 2024; Morel et al., 2025), while target-specific adaptation adjusts model parameters or latent variables (Kirchmeyer et al., 2022; Kassaï Koupaï et al., 2024). We study a complementary *context-conditioned zero-shot* setting in which the observed prefix selects a deployment configuration for a frozen surrogate.

Neural Operator Splitting (NOS) (Serrano et al., 2026) realizes this idea through explicit reuse and composition. It extracts a frozen codebook of evolution operators from a pretrained model and searches over short compositions that can approximate the unknown target dynamics. Given an observed prefix $S = (u^1, \dots, u^L)$, NOS scores each candidate by one-step fitting error, selects the best-fitting composition, and repeatedly applies it to predict the unseen future. Its search-time predictions start from observed states, whereas forecast-time predictions start from the candidate's previous outputs. A candidate with low one-step error can therefore accumulate error rapidly, while a candidate with slightly worse local fit can remain more accurate in closed loop. Conventional NOS can consequently select the wrong candidate even when its beam contains a better rollout candidate, as formalized by Proposition 2.1. This selection–deployment gap motivates closed-loop reranking of the structured NOS beam.

HINDCAST-NOS turns the observed prefix into a test-time reranking set. It restarts each retained candidate from an earlier observed state and rolls it through the rest of the prefix. The known suffix measures closed-loop error before future forecasting. This replay asks how each candidate would have evolved the already observed system had it been deployed earlier.

The resulting per-trajectory selector couples closed-loop beam reranking with a PDE-specific analysis of support-to-future transfer. Figure 1 shows the two selection rules.

Our contributions are:

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

- We introduce a training-free test-time criterion that reranks a frozen NOS candidate beam under PDE distribution shift using multi-origin closed-loop replay on observed support states.
- We establish finite-horizon guarantees for support-to-future transfer, including a sharper ordering result under positive error memory, and use PDE energy estimates to verify the required true-flow stability for all three equations considered.
- We demonstrate effective reranking on three 128-target PDE evaluations: the criterion improves NOS selection, and rank-correlation, compute, beam-width, and horizon analyses characterize the gain.

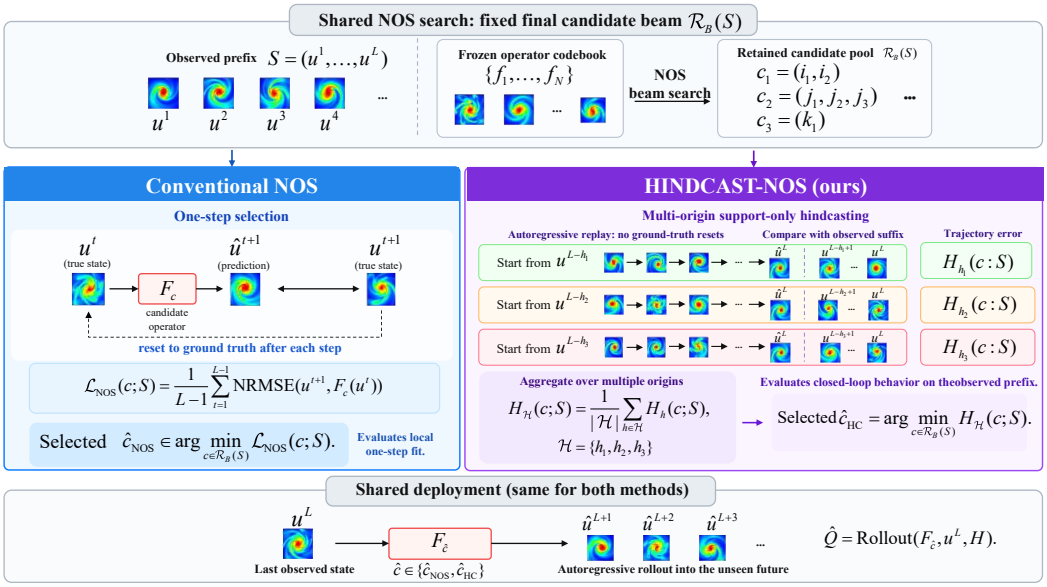


Figure 1: Comparison between conventional NOS and HINDCAST-NOS. Both methods share the frozen backbone, operator codebook, retained beam, and autoregressive deployment rule. Conventional NOS selects by its one-step fitting error, whereas HINDCAST-NOS reranks the same candidates by closed-loop replays from multiple observed origins and averages their scores over a replay set $\mathcal{H}$; Table 1 gives the experimental choice.

2 Background and Problem Setting

Neural operators such as FNO and DeepONet learn surrogate solution maps for PDEs (Li et al., 2020; Lu et al., 2019; Kovachki et al., 2023), while multi-physics models such as MPP and DISCO share representations across heterogeneous systems (McCabe et al., 2024; Morel et al., 2025). Under distribution shift, meta-learning methods adapt parameters or latent environment variables (Finn et al., 2017; Kirchmeyer et al., 2022; Kassaï Koupaï et al., 2024). We focus on short-context OOD forecasting with a frozen pretrained model. NOS keeps the pretrained model fixed and searches a dictionary of source operators at test time (Serrano et al., 2026). We retain this frozen-search setting and replace the final selection criterion applied to its beam.

2.1 Context-conditioned zero-shot operator selection

Consider a parametric evolution equation whose dynamics combine physical components,

$$\partial_t u = \sum_{k=1}^{K} \mu_k \mathcal{F}_k\big(u, \nabla u, \nabla^2 u, \dots\big). \tag{1}$$

Here $u(t,\cdot)\in\mathcal{X}$ is the state field on a spatial domain Ω, while μ_k and $\mathcal{F}_k$ denote physical coefficients and dynamical components. During source training, trajectories contain isolated physical components

or coefficients from prescribed training ranges. A pretrained DISCO model encodes these source trajectories into a dictionary of neural evolution operators. At test time, we consider the two forms of distribution shift studied in NOS (Serrano et al., 2026): *parameter extrapolation*, where a target coefficient lies outside the source-training range, and *operator composition*, where physical mechanisms observed separately in the source environments act jointly in the target PDE. The pretrained backbone and operator dictionary remain fixed; test-time inference only selects and composes dictionary operators using the observed target prefix.

Formally, let $\mathcal{D} = \{f_i\}_{i=1}^N$ denote the frozen dictionary of N neural evolution operators. For each target trajectory, the selector receives the short prefix $S := u^{1:L} = (u^1, \dots, u^L)$ and predicts the future $Q := u^{L+1:L+H} = (u^{L+1}, \dots, u^{L+H})$. A candidate $c = (i_1, \dots, i_m)$, with $i_j \in \{1, \dots, N\}$, indexes a composition of length m (including $m = 1$). The map $F_c : \mathcal{X} \to \mathcal{X}$ denotes its resulting one-step state update under the fixed NOS splitting integrator. For any map F, initial state z, and integer $r \geq 1$, define

$$\mathrm{Rollout}(F, z, r) := (z^1, \dots, z^r), \qquad z^0 = z, \quad z^j = F(z^{j-1}). \tag{2}$$

Thus c predicts $\widehat{Q}_c = \mathrm{Rollout}(F_c, u^L, H)$. Following NOS terminology (Serrano et al., 2026), this is *context-conditioned zero-shot* prediction: the target prefix drives discrete operator selection, model parameters remain frozen, and the future is reserved for final evaluation. Few-shot adaptation instead updates model or latent parameters using target examples. Unless stated otherwise, $\|\cdot\|_2$ is the discrete Euclidean norm over spatial grid points and state channels. Following NOS (Serrano et al., 2026), the error of one predicted frame is $\mathrm{NRMSE}(a, \widehat{a}) := \|a - \widehat{a}\|_2 / \|a\|_2$; errors over several frames are averaged in time. We use this mean framewise NRMSE for fitting, hindcasting, and future evaluation.

2.2 Conventional NOS beam search

Following NOS (Serrano et al., 2026), candidate generation evaluates one-step fitting error on all observed transitions. Its fitting score is the mean framewise NRMSE

$$\mathcal{L}_{\mathrm{NOS}}(c; S) := \frac{1}{L-1} \sum_{t=1}^{L-1} \mathrm{NRMSE}\big(u^{t+1}, F_c(u^t)\big)\,. \tag{3}$$

Beam search constructs compositions progressively. It first scores length-one candidates, keeps the B lowest-scoring candidates under Equation (3), and expands each survivor by appending one dictionary operator. After every expansion, it rescores and prunes the pool back to width B, continuing until the stopping criterion or maximum composition length is reached. Let $\mathcal{R}_B(S)$ denote this final beam of retained candidates. Conventional NOS returns $\widehat{c}_{\mathrm{NOS}}(S) \in \arg\min_{c \in \mathcal{R}_B(S)} \mathcal{L}_{\mathrm{NOS}}(c; S)$.

Each term $F_c(u^t)$ receives the observed state u^t, and the next term starts again from the observation u^{t+1}. The NOS fitting score therefore measures local transition fit. During deployment, $F_c(\widehat{u}_c^t)$ receives a candidate-generated state, so errors alter subsequent inputs and can reverse the one-step ranking over a rollout.

This is an objective mismatch: it can reverse the selected candidate even for linear autonomous dynamics.

Proposition 2.1 (One-step selection can reverse rollout ranking)**.** *There exist globally linear, smooth, and Lipschitz autonomous dynamics and two candidates F_A, F_B such that the NOS one-step fitting score strictly prefers F_A, while both support hindcasting and future autoregressive rollout strictly prefer F_B.*

Proof. Take $u^0 = 1$ and

$$G(x) = x, \qquad F_A(x) = 1.4x, \qquad F_B(x) = 0.5x, \tag{4}$$

Every target state equals one. The one-step errors are therefore 0.4 and 0.5, so NOS selects A on any observed transition. In closed loop, $\widehat{u}_A^{1:2} = (1.4, 1.96)$ whereas $\widehat{u}_B^{1:2} = (0.5, 0.25)$, and

$$E_2(A) = (0.4 + 0.96)/2 = 0.68 > 0.625 = (0.5 + 0.75)/2 = E_2(B). \tag{5}$$

Here E_2 is mean framewise NRMSE over the two rollout steps. Thus replay prefers B; because the target is constant, the same calculation holds on the adjacent future segment. □

The reversal already occurs for linear autonomous dynamics with identical support and future regimes; each additional observed-state transition repeats the same ordering.

3 SUPPORT-ONLY HINDCASTING

To address the selection–deployment mismatch above, HINDCAST-NOS evaluates each retained candidate under the autoregressive feedback it will face at deployment. It first restarts the candidate from an earlier observed state and rolls it over the remaining observed suffix, which provides a closed-loop error measurement. We then average the same support test over several earlier states to cover multiple rollout origins. Both scores rerank the NOS beam $\mathcal{R}_B(S)$ and measure error amplification under self-generated inputs before forecasting begins.

3.1 HINDCAST-BASED SELECTION

Take a candidate from the conventional NOS beam and pretend that it had been deployed h steps before the observed prefix ends. Starting from the true state u^{L-h}, we run the candidate freely for h steps—feeding each prediction back as the next input—and compare the resulting trajectory with the already observed suffix. Using the rollout notation in Equation (2), this gives the single-origin hindcast score

$$H_h(c;S) := \frac{1}{h}\sum_{j=1}^{h} \mathrm{NRMSE}\big(u^{L-h+j}, F_c^j(u^{L-h})\big)\,. \tag{6}$$

This score directly tests error amplification under candidate-generated inputs: it penalizes amplification and rewards stable or self-correcting rollouts. It is computed on observed states. The horizon h determines how much of the prefix is replayed: $h = L-1$ uses the full prefix, while a smaller h uses a recent suffix.

To reduce dependence on a single rollout origin, let $\mathcal{H}$ be a set of observed lookback lengths and aggregate

$$H_{\mathcal{H}}(c;S) := \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} H_h(c;S). \tag{7}$$

We space replay lengths uniformly over the $L-1$ observed transitions: with J origins, $h_j = \mathrm{round}(j(L-1)/J)$ for $j = 1,\dots,J$. Our main configuration uses $J = 3$ with equal weights, which gives $\mathcal{H} = \{5, 10, 15\}$ when $L = 16$. Setting $J = 1$ recovers full-prefix single-origin hindcasting.

Algorithm 1: Single-/multi-origin HINDCAST-NOS

Require: $S = u^{1:L}$, $\mathcal{D}$, B, $\mathcal{H}$, forecast horizon H
Ensure: Forecast $\widehat{Q}$
$\mathcal{R}_B(S) \leftarrow \textsc{NosBeam}(S, \mathcal{D}, B, \mathcal{L}_{\mathrm{NOS}})$
for all $c \in \mathcal{R}_B(S)$ **do**
  for all $h \in \mathcal{H}$ **do**
    Roll out F_c from u^{L-h} for h steps
    Compute $H_h(c;S)$ by Eq. (6)
  end for
  $H_{\mathcal{H}}(c;S) \leftarrow |\mathcal{H}|^{-1}\sum_{h\in\mathcal{H}} H_h(c;S)$
end for
$\widehat{c}_{\mathrm{HC}} \leftarrow \arg\min_{c\in\mathcal{R}_B(S)} H_{\mathcal{H}}(c;S)$
$\widehat{Q} \leftarrow \textsc{Rollout}(F_{\widehat{c}_{\mathrm{HC}}}, u^L, H)$
return $\widehat{Q}$

We evaluate this score for every candidate retained by NOS, select the lowest one, and roll it forward from the final observed state:

$$\widehat{c}_{\mathrm{HC}}(S) \in \arg\min_{c\in\mathcal{R}_B(S)} H_{\mathcal{H}}(c;S), \qquad \widehat{Q} = \mathrm{Rollout}(F_{\widehat{c}_{\mathrm{HC}}}, u^L, H). \tag{8}$$

Thus HINDCAST-NOS reranks the existing beam and deploys the winner with the same forecasting rule. Section 4 analyzes the transfer of each constituent hindcast score, and Table 1 specifies $\mathcal{H}$.

Algorithm 1 covers both variants. A singleton $\mathcal{H} = \{h\}$ gives single-origin hindcasting; several lookback lengths give multi-origin hindcasting by averaging the constituent scores. Every replay starts and ends inside the observed prefix. Both variants are training-free and gradient-free, and Q is reserved for final evaluation.

4 PDE SUPPORT-TO-FUTURE TRANSFER

Section 3 matches candidate selection to the closed-loop feedback used at deployment. We now formalize why replay error on the observed suffix transfers to the adjacent future for PDE trajectories.

Under autonomous dynamics $u^{t+1} = G(u^t)$, both segments are generated by the same solution map, and PDE energy estimates control how this map propagates a change in rollout origin. The analysis first gives a finite-horizon transfer bound and then verifies its true-flow stability condition for each of the three PDEs studied experimentally.

4.1 From observed hindcast to future error

Condition on the retained pool $\mathcal{R} := \mathcal{R}_B(S)$ and set $x = u^{L-h}$ and $y = u^L = G^h(x)$. Define candidate c's future error by

$$E_c := \frac{1}{H}\sum_{j=1}^{H} \mathrm{NRMSE}\big(G^j(y), F_c^j(y)\big)\,. \tag{9}$$

Let $c_H^\star \in \arg\min_{c\in\mathcal{R}} E_c$ be the future oracle in the retained beam. For any constituent origin h, let $\widehat{c}_h$ minimize the observed score $H_h(c;S)$. The analysis asks when this support ranking agrees with the future ranking. The main multi-origin selector averages these constituent support-only scores as in Equation (7).

Consider the states visited by F_c and G from the two origins x and y up to horizon H. On this finite set, assume

$$\sup_z \|F_c(z) - G(z)\|_2 \le \delta_c, \qquad \mathrm{Lip}(F_c) \le q_c, \qquad \mathrm{Lip}(G) \le q_G. \tag{10}$$

Thus δ_c measures local one-step discrepancy, while q_c and q_G control how the candidate and true flow amplify perturbations. The PDE estimates below control q_G. Let $\varepsilon_c(h,H)$ be the explicit transfer radius in Appendix A.1. It accumulates these local bounds, the distance between the two rollout origins, and the target-state normalizations over the replay and forecast horizons. We assume $1 \le h \le \min\{L-1, H\}$ and nonzero target-frame norms.

Theorem 4.1 (Future NRMSE transfer)**.** *Under the conditions above, every $c \in \mathcal{R}$ satisfies the first bound below. Writing $E_\star = E_{c_H^\star}$, the selected candidate satisfies the second:*

$$\left|E_c - \frac{h}{H}H_h(c;S)\right| \le \varepsilon_c(h,H), \qquad 0 \le E_{\widehat{c}_h} - E_\star \le \varepsilon_{\widehat{c}_h}(h,H) + \varepsilon_{c_H^\star}(h,H). \tag{11}$$

Equation (11) quantifies the gap between future error and the time-calibrated support score $(h/H)H_h$. Since $h/H > 0$ is common to all candidates, this calibration preserves their support ranking. The second inequality bounds the selected candidate's excess future NRMSE over the future oracle in the retained beam. Its proof is in Appendix A.1, and Appendix A.2 gives a sufficient condition for strict improvement over NOS.

Applying Theorem 4.1 to each replay origin and averaging the result gives the corresponding oracle bound for the multi-origin selector. The positive calibration is common to all candidates, so the implemented equal-weight ranking is preserved even when the observed suffixes overlap. The formal statement and proof are in Appendix A.2.

4.2 Sharper guarantees under positive error memory

The preceding bounds apply to general Lipschitz candidates. A two-coordinate model makes the error-memory mechanism explicit and yields a stronger ordering guarantee against the NOS winner. Consider

$$u^t = (a^t, 0), \qquad F_c(s,z) = \big(as,\, b_c s + \rho_c z\big), \tag{12}$$

with $0 < a < 1$ and $b_c, \rho_c \ge 0$. The first coordinate is the observed trajectory; b_c injects error into the transverse coordinate, and ρ_c determines how strongly that error persists. Define

$$\ell_n(c) := \frac{b_c}{a}\frac{1}{n}\sum_{j=1}^{n}\sum_{k=0}^{j-1}\Big(\frac{\rho_c}{a}\Big)^k. \tag{13}$$

Direct iteration gives $H_h(c;S) = \ell_h(c)$ and $E_H(c) = \ell_H(c)$, while the NOS one-step fitting score is b_c/a. Thus NOS minimizes injected error, whereas hindcasting also measures how long that error persists.

Intuition. Consider two candidates with the same observed trajectory. One may make a smaller error on the next frame but retain that error in a transverse mode; another may make a larger one-step error but rapidly damp the resulting mode. NOS one-step fitting sees only the initial injection, because every transition starts from its observed input. A closed-loop replay instead feeds the candidate's own output back into itself, so the retained error is propagated and accumulated across frames. The parameter b_c represents the initial injection, while ρ_c represents its memory; hindcasting separates these two effects. In a PDE, the same distinction appears when a local surrogate error is transported or diffused by subsequent operator applications: its size at the next step is not enough to determine its future impact.

Theorem 4.2 (Persistent improvement under positive error memory)**.** *Assume Equation* (12)*, at least one replay length $h \geq 2$, and $H \geq \max \mathcal{H}$. Let $c_0 \in \arg\min_{c\in\mathcal{R}} b_c$ be a NOS winner and let $\widehat{c} \in \arg\min_{c\in\mathcal{R}} H_{\mathcal{H}}(c; S)$. If $b_{c_0} > 0$, then*

$$\frac{E_H(\widehat{c})}{E_H(c_0)} \leq \frac{H_{\mathcal{H}}(\widehat{c}; S)}{H_{\mathcal{H}}(c_0; S)} \leq 1. \tag{14}$$

If $b_{c_0} = 0$, both selectors attain zero future error. Hence every strict hindcast improvement over the NOS winner persists at all deployment horizons beyond the replay lengths.

When transverse memory decays at least as fast as the observed signal, $0 \leq \rho_c \leq a$, define

$$\chi_n := \frac{n+1}{2}, \qquad \Gamma_{\mathcal{H},H} := \frac{|\mathcal{H}|\chi_H}{\sum_{h\in\mathcal{H}} \chi_h}. \tag{15}$$

Theorem 4.3 (Sharp same-beam oracle factors)**.** *Under Equation* (12) *and $0 \leq \rho_c \leq a$, every $c \in \mathcal{R}$ satisfies*

$$H_{\mathcal{H}}(c; S) \leq E_H(c) \leq \Gamma_{\mathcal{H},H} H_{\mathcal{H}}(c; S). \tag{16}$$

If $E^*_H = \min_{c\in\mathcal{R}} E_H(c)$, then

$$E_H(\widehat{c}) \leq \Gamma_{\mathcal{H},H} E^*_H, \qquad E_H(c_0) \leq \chi_H E^*_H. \tag{17}$$

Both factors are sharp over this model class and finite candidate beams.

The model isolates a property absent from one-step fitting: positive memory of errors transverse to the observed trajectory. Appendix B contains the proofs, a strictly contractive separation from one-step reweighting, perturbation bounds, a signed-memory counterexample, and a reaction–diffusion realization of the two-coordinate dynamics.

4.3 PDE FLOW STABILITY

Theorem 4.1 uses continuous dependence of the target flow on its initial state, a verifiable property of the PDE solution maps considered here. Specifically, the q_G^j term in the first sum of Equation (22) measures how much the true flow can amplify the shift from the hindcast origin x to the forecast origin y. Standard energy estimates give explicit q_G for all three target PDEs. On a periodic domain, the advection–diffusion flow is non-expansive, giving the strongest case: its true-flow contribution is uniformly bounded in j.

Proposition 4.4 (Advection–diffusion stability)**.** *For $\partial_t u + v\partial_x u = D\partial_{xx} u$ with $D > 0$, $\|\Phi_t(u_0) - \Phi_t(\bar{u}_0)\|_2 \leq \|u_0 - \bar{u}_0\|_2$; hence $q_G \leq 1$ for $G = \Phi_{\Delta t}$.*

Proposition 4.5 (Viscous Burgers stability)**.** *If two sufficiently regular periodic solutions satisfy $\|\partial_x u(t)\|_\infty + \|\partial_x \bar{u}(t)\|_\infty \leq M_B$ on $[0, T]$, then*

$$\|u(t) - \bar{u}(t)\|_2 \leq e^{M_B t/4} \|u(0) - \bar{u}(0)\|_2, \tag{18}$$

so the one-step map satisfies $q_G \leq e^{M_B \Delta t/4}$.

Proposition 4.6 (Two-dimensional viscous Navier–Stokes stability)**.** *For two sufficiently regular periodic vorticity solutions, if $\|\nabla\bar{\omega}(t)\|_{L^4} \leq M_{\mathrm{NS}}$ on $[0, T]$, then a domain-dependent $C_{\mathrm{BS}} > 0$ satisfies*

$$\|\omega(t) - \bar{\omega}(t)\|_2 \leq e^{C_{\mathrm{BS}} M_{\mathrm{NS}} t} \|\omega(0) - \bar{\omega}(0)\|_2. \tag{19}$$

Thus the finite-time flow is Lipschitz on trajectory classes with uniform M_{NS}, and the one-step map satisfies $q_G \leq e^{C_{\mathrm{BS}} M_{\mathrm{NS}} \Delta t}$. Proofs are in Appendix A.

The benchmark trajectories agree with these PDE stability estimates. Table 5 reports nearest-neighbor one-step amplification ratios. Advection–diffusion remains non-expansive with maximum ratio 0.9991. For Burgers, all 128 target–neighbor pairs satisfy the empirical bound $e^{\widehat{M}_B \Delta t/4}$ over their observed transitions. Navier–Stokes amplification stays close to one, with maximum ratio 1.0283.

PDE stability verification. Propositions 4.4–4.6 verify the true-flow condition in Theorem 4.1 for every benchmark: $q_G \leq 1$ for advection–diffusion and finite-time exponential bounds for Burgers and Navier–Stokes. The governing equations therefore control the target-flow side of support-to-future transfer. Selection remains equation free and operates from observed trajectories, while δ_c and q_c describe learned-candidate regularity and replay measures the resulting closed-loop behavior.

5 Experiments

5.1 Environment Settings

We evaluate three distribution-shifted PDE environments, each with 128 target trajectories. In **2D Navier–Stokes**, the target vorticity dynamics combine Euler transport and diffusion operators observed separately in the source environments. In **1D advection–diffusion**, the target likewise activates advection and diffusion jointly, while the source operator families contain the two mechanisms separately. These two environments test unseen operator composition under the context-conditioned zero-shot NOS protocol. In **PDEBench 1D Burgers** (Takamoto et al., 2022), the target viscosity is $\nu = 0.001$, outside the source viscosities $\nu \in \{0.01, 0.1\}$, and therefore tests parameter extrapolation. The governing equations are given in Appendix D.

For Navier–Stokes and advection–diffusion, the pretrained backbone, codebook construction, temporal discretization, and conventional beam search follow NOS (Serrano et al., 2026). Burgers uses the same frozen codebook-and-search pipeline on the PDEBench environments. HINDCAST-NOS replaces the terminal ranking of the retained beam. The main rows use beam width $B = 4$ and replay all $R = 4$ final-beam candidates. The beam width B follows the NOS baseline configuration, and we replay all $R = B$ surviving candidates. Following the uniform rule in Section 3, we use $J = 3$ equal-weight replay lengths. Since every task has $L = 16$, this gives $\mathcal{H} = \{5, 10, 15\}$, spanning short, intermediate, and full-prefix replay. This prefix-length rule fixes the origins uniformly across tasks and targets. Table 1 records the task-specific forecast horizons; complete implementation details are in Appendix C. We select the replay hyperparameters, including the number of origins J, on a separate validation set and keep them fixed for all test evaluations. This yields $J = 3$; given the uniform-origin rule and $L = 16$, the replay set is therefore $\mathcal{H} = \{5, 10, 15\}$. No test-set future errors are used for hyperparameter selection.

Table 1: Evaluation protocols. B is beam width and R is the number of retained candidates replayed on observed support.

Task	L	H	Targets	B/R	$\mathcal{H}$
2D Navier–Stokes	16	16	128	4/4	$\{5, 10, 15\}$
1D Advection–Diffusion	16	34	128	4/4	$\{5, 10, 15\}$
1D Burgers	16	16	128	4/4	$\{5, 10, 15\}$

Evaluation. Throughout the experiments, *future error* is mean framewise NRMSE over the complete forecast segment,

$$\mathcal{E}(Q, \widehat{Q}) := \frac{1}{H} \sum_{j=1}^{H} \mathrm{NRMSE}\big(u^{L+j}, \widehat{u}^{L+j}\big)\,, \tag{20}$$

which is then averaged over targets. Statistical comparisons are paired over the same trajectories. We report two-sided Pratt–Wilcoxon signed-rank tests.

5.2 MAIN RESULTS

Table 2: Mean framewise NRMSE across PDE forecasting tasks. AD and NS baseline values are published results reported in NOS; Burgers uses our 128-target evaluation. Single- and multi-origin denote HINDCAST-NOS. Their respective reductions over NOS are $2.452/2.476\%$ on AD, $0.696/1.160\%$ on NS, and $0.589/0.611\%$ on Burgers. * denotes a significant paired improvement over NOS ($p < 0.05$). Lower is better.

Method	AD	NS	Burgers
MPP (McCabe et al., 2024)	0.270	0.273	0.62441
Zebra (Serrano et al., 2024)	0.893	0.198	0.46525
GEPS (Kassaï Koupaï et al., 2024)	0.039	0.786	1.60374
DISCO (Morel et al., 2025)	0.170	0.572	0.47031
NOS (Serrano et al., 2026)	0.014708	0.064220	0.220064
Single-origin (ours)	0.014347	0.063773	0.218768
Multi-origin (ours)	**0.014344***	**0.063475***	**0.218720***

Single-origin hindcasting improves conventional NOS on all three tasks; multi-origin aggregation increases the reductions to 2.476% on advection–diffusion, 1.160% on Navier–Stokes, and 0.611% on Burgers. All three multi-origin paired tests are significant. Table 3 examines selection directly: support replay more often chooses the candidate with the lowest future error in the fixed beam.

Table 3: Future-oracle hit rate in the fixed conventional beam. Each entry is the number and percentage of 128 targets for which the selector chooses the candidate with lowest future mean framewise NRMSE. Bold marks the highest rate within each task.

Task	Conventional NOS	Single-origin	Multi-origin
AD	$84/128$ (65.6%)	$109/128$ (85.2%)	$111/128$ (86.7%)
NS2D	$83/128$ (64.8%)	$98/128$ (76.6%)	$102/128$ (79.7%)
Burgers	$70/128$ (54.7%)	$82/128$ (59.4%)	$76/128$ (64.1%)

The analyses below further examine rank correlation, compute cost, beam width, and local horizon sensitivity.

Rank-correlation analysis on 2D Navier–Stokes. On 128 Navier–Stokes targets, rank correlation with future NRMSE increases from $\rho = 0.833$ for one-step error to $\rho = 0.892$ for hindcast error. Hindcast improvement also correlates with future improvement ($\rho = 0.431$, $p = 3.71 \times 10^{-7}$), while reducing the mean from 0.064220 to 0.063773 in this constituent-score analysis.

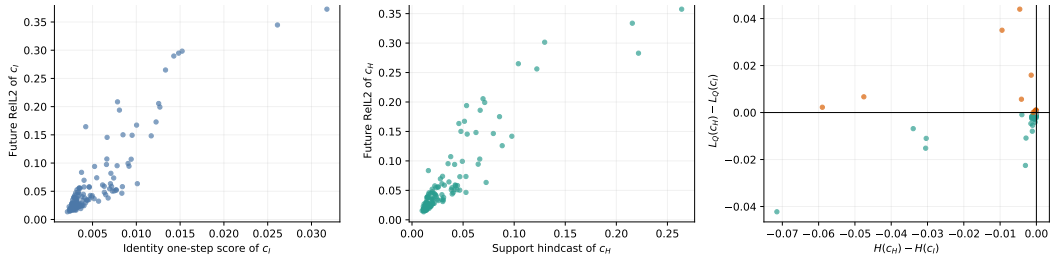


Figure 2: Single-origin Navier–Stokes B4/R4 rank-correlation analysis. Future mean framewise NRMSE aligns better with hindcast than one-step score ($\rho = 0.892$ vs. 0.833), and support-error change predicts future-error change ($\rho = 0.431$).

5.3 COMPARISONS AND ABLATIONS

The baseline and HINDCAST-NOS share the frozen network, codebook, beam, context, forecast horizon, candidate pool, and error definition within each row. Their terminal rankings use the NOS fitting score and support replay, respectively. Every main result uses the hard argmin in Equation (8) with the origins in Table 1; the single-origin rows rerank the same B4/R4 pool from one observed origin.

Selection cost. On Navier–Stokes, normalized selection cost is 1.00 for conventional NOS, 1.03 for single-origin replay, and 1.06 for multi-origin replay. All three share the same post-selection future rollout. Thus the multi-origin selector obtains its improved ranking with a 6% increase in selection-time operator-step evaluations (Figure 3).

Beam width and retained pool. With beam width $B = 8$, replaying the top $R = 2$ and $R = 4$ candidates reduces future NRMSE relative to the corresponding NOS winner by 2.92% and 5.87%, respectively. The stronger B8/R4 result shows that hindcasting benefits from access to more surviving alternatives (Table 6).

Replay horizons and origins. The Navier–Stokes gain remains positive for every tested single-origin horizon $h \in \{11, \dots, 15\}$ (Figure 4). Likewise, all seven nonempty subsets of $\{5, 10, 15\}$ improve conventional NOS, and the full three-origin set gives the largest reduction, 1.160% (Table 7).

Observed context length. Table 4 varies the number of observed frames while keeping the B4/R4 setting and forecast horizon fixed. Single-origin replay uses $h = L - 1$, while multi-origin replay applies the uniform three-origin rule from Section 3. Both variants improve their matched conventional NOS baseline at every context length. Single-origin gives its largest reduction at $L = 12$, whereas multi-origin gives its largest reduction at $L = 16$.

Table 4: Navier–Stokes context-length ablation over 128 targets. Each hindcast result is compared with conventional NOS at the same L.

L	Conventional NOS	Single-origin	Reduction	Multi-origin	Reduction
8	0.064334	0.063770	0.877%	0.063985	0.542%
12	0.064163	0.063070	$\mathbf{1.704}\%$	0.063757	0.632%
16	0.064220	0.063773	0.696%	0.063475	$\mathbf{1.160}\%$

Single- versus multi-origin replay. Multi-origin replay attains the lowest mean NRMSE on all three tasks. Its largest advantage over single-origin replay occurs on Navier–Stokes, where the error decreases from 0.063773 to 0.063475 (Table 8).

Per-target behavior. The paired results in Figure 5 show how the aggregate gains are distributed across targets. The largest-reduction cases change the selected operator composition and reduce future NRMSE by 10.02–21.64% (Table 9 and Figures 6, 7, and 8).

6 CONCLUSION

We introduced HINDCAST-NOS, a support-only selector for frozen NOS candidate beams under PDE distribution shift. Rather than accepting the one-step NOS winner, the method replays each retained operator composition autoregressively from multiple observed origins and deploys the candidate with the lowest mean suffix error. Our analysis connects replay on the observed prefix to the adjacent future through finite-horizon transfer bounds, and the positive-error-memory model explains why a hindcast advantage can persist over longer rollouts. Experimental results show that an observed PDE trajectory can serve both as context for operator search and as a closed-loop validation set for test-time selection.

AI USE STATEMENT

Generative AI tools assisted with mathematical proof checking, code verification, and manuscript polish. They were not used to generate experimental data or results. The authors reviewed and verified all AI-assisted outputs and take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The selection rule and implementation details are specified in Section 3 and Algorithm 1; datasets, evaluation protocols, and hyperparameters are documented in Section 5 and the appendix. Complete proofs are provided in the appendix, and the supplementary material includes the code and per-target experimental outputs.

REFERENCES

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Armand Kassaï Koupaï, Jorge Mifsut Benet, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Boosting generalization in parametric pde neural solvers through adaptive conditioning. *Advances in Neural Information Processing Systems*, 37:70659–70692, 2024.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *International conference on machine learning*, pp. 11283–11301. PMLR, 2022.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for spatiotemporal surrogate models. *Advances in Neural Information Processing Systems*, 37:119301–119335, 2024.

Rudy Morel, Jiequn Han, and Edouard Oyallon. Disco: learning to discover an evolution operator for multi-physics-agnostic prediction. *arXiv preprint arXiv:2504.19496*, 2025.

Louis Serrano, Armand Kassaï Koupaï, Thomas X Wang, Pierre Erbacher, and Patrick Gallinari. Zebra: In-context generative pretraining for solving parametric pdes. *arXiv preprint arXiv:2410.03437*, 2024.

Louis Serrano, Jiequn Han, Edouard Oyallon, Shirley Ho, and Rudy Morel. Test-time generalization for physics through neural operator splitting. *arXiv preprint arXiv:2602.00884*, 2026.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in neural information processing systems*, 35:1596–1611, 2022.

A Proofs for the Support-to-Future Analysis

A.1 Explicit radius and proof of Theorem 4.1

Set $d_h := \|x - y\|_2$. For $1 \le j \le H$, define

$$B_{c,j} := \delta_c \sum_{k=0}^{j-1} \min(q_c, q_G)^k. \tag{21}$$

The transfer radius used in Theorem 4.1 is

$$\begin{aligned} \varepsilon_c(h, H) := \frac{1}{H} \Bigg\{ \sum_{j=1}^{h} \left[\frac{\min\{d_h(q_c^j + q_G^j), 2B_{c,j}\}}{\|G^j(y)\|_2} + \frac{B_{c,j} d_h q_G^j}{\|G^j(x)\|_2 \|G^j(y)\|_2} \right] \\ + \sum_{j=h+1}^{H} \frac{B_{c,j}}{\|G^j(y)\|_2} \Bigg\}. \end{aligned} \tag{22}$$

The first sum accounts for moving the rollout origin from x to y; the second covers forecast steps beyond the observed replay. Empty sums are zero.

Proof. Fix c, abbreviate $B_j = B_{c,j}$, and put $e_{c,j}(z) = F_c^j(z) - G^j(z)$ and $r_j(z) = \|e_{c,j}(z)\|_2 / \|G^j(z)\|_2$ for $z \in \{x, y\}$.

Adding and subtracting either $F_c(G^j z)$ or $G(F_c^j z)$ and iterating from $e_{c,0}(z) = 0$ gives

$$\|e_{c,j+1}(z)\|_2 \le \delta_c + \min(q_c, q_G)\|e_{c,j}(z)\|_2 \quad \Longrightarrow \quad \|e_{c,j}(z)\|_2 \le B_j, \quad z \in \{x, y\}. \tag{23}$$

Write $R_{c,j}(z) = \|e_{c,j}(z)\|_2$. Lipschitz continuity at the two origins and the reverse triangle inequality give

$$\begin{aligned} \|e_{c,j}(x) - e_{c,j}(y)\|_2 &\le d_h(q_c^j + q_G^j), \\ |r_j(y) - r_j(x)| &\le \frac{|R_{c,j}(y) - R_{c,j}(x)|}{\|G^j(y)\|_2} \\ &\quad + \frac{R_{c,j}(x) | \|G^j(x)\|_2 - \|G^j(y)\|_2 |}{\|G^j(x)\|_2 \|G^j(y)\|_2} \\ &\le \frac{\min\{d_h(q_c^j + q_G^j), 2B_j\}}{\|G^j(y)\|_2} + \frac{B_j d_h q_G^j}{\|G^j(x)\|_2 \|G^j(y)\|_2}, \end{aligned} \tag{24}$$

For $j > h$, Equation (23) gives $r_j(y) \le B_j / \|G^j(y)\|_2$, and hence

$$\left| E_c - \frac{h}{H} H_h(c; S) \right| \le \frac{1}{H} \left[\sum_{j=1}^{h} |r_j(y) - r_j(x)| + \sum_{j=h+1}^{H} r_j(y) \right] \le \varepsilon_c(h, H).$$

Because $\widehat{c}_h$ minimizes H_h and $h/H > 0$, applying Equation (11) twice gives

$$\begin{aligned} E_{\widehat{c}_h} &\le \frac{h}{H} H_h(\widehat{c}_h; S) + \varepsilon_{\widehat{c}_h}(h, H) \\ &\le \frac{h}{H} H_h(c_H^\star; S) + \varepsilon_{\widehat{c}_h}(h, H) \\ &\le E_\star + \varepsilon_{c_H^\star}(h, H) + \varepsilon_{\widehat{c}_h}(h, H). \end{aligned}$$

The oracle property gives $E_{\widehat{c}_h} \ge E_\star$, completing the second inequality in Equation (11). □

A.2 Multi-origin extension

For each $h \in \mathcal{H}$, evaluate Equation (22) with $x_h = u^{L-h}$ and define

$$\alpha_{\mathcal{H}} := \frac{|\mathcal{H}|}{H \sum_{\ell \in \mathcal{H}} \ell^{-1}}, qquad \bar{\varepsilon}_c := \frac{\sum_{h \in \mathcal{H}} h^{-1} \varepsilon_c(h, H)}{\sum_{\ell \in \mathcal{H}} \ell^{-1}}. \tag{25}$$

Corollary A.1 (Multi-origin transfer)**.** *Under the conditions of Theorem 4.1 uniformly over $h \in \mathcal{H}$, every $c \in \mathcal{R}$ satisfies*

$$|E_c - \alpha_{\mathcal{H}} H_{\mathcal{H}}(c;S)| \le \bar{\varepsilon}_c, \qquad 0 \le E_{\widehat{c}_{\mathrm{HC}}} - E_\star \le \bar{\varepsilon}_{\widehat{c}_{\mathrm{HC}}} + \bar{\varepsilon}_{c_H^\star}. \tag{26}$$

Proof. For each $h \in \mathcal{H}$, put $s_{c,h} := (h/H)H_h(c;S)$ and $\varepsilon_{c,h} := \varepsilon_c(h,H)$. Theorem 4.1 gives $|E_c - s_{c,h}| \le \varepsilon_{c,h}$. Let $\beta_h = h^{-1}/\sum_{\ell\in\mathcal{H}} \ell^{-1}$. These weights are nonnegative and sum to one, and

$$\begin{aligned} \sum_{h\in\mathcal{H}} \beta_h s_{c,h} &= \frac{\sum_{h\in\mathcal{H}} H_h(c;S)}{H\sum_{\ell\in\mathcal{H}}\ell^{-1}} = \alpha_{\mathcal{H}} H_{\mathcal{H}}(c;S), \\ |E_c - \alpha_{\mathcal{H}} H_{\mathcal{H}}(c;S)| &\le \sum_{h\in\mathcal{H}} \beta_h |E_c - s_{c,h}| \le \bar{\varepsilon}_c. \end{aligned} \tag{27}$$

Since $\widehat{c}_{\mathrm{HC}}$ minimizes $H_{\mathcal{H}}$ and $\alpha_{\mathcal{H}} > 0$,

$$\begin{aligned} E_{\widehat{c}_{\mathrm{HC}}} &\le \alpha_{\mathcal{H}} H_{\mathcal{H}}(\widehat{c}_{\mathrm{HC}};S) + \bar{\varepsilon}_{\widehat{c}_{\mathrm{HC}}} \\ &\le \alpha_{\mathcal{H}} H_{\mathcal{H}}(c_H^\star;S) + \bar{\varepsilon}_{\widehat{c}_{\mathrm{HC}}} \\ &\le E_\star + \bar{\varepsilon}_{c_H^\star} + \bar{\varepsilon}_{\widehat{c}_{\mathrm{HC}}}. \end{aligned} \tag{28}$$

The lower bound follows from the definition of $c_H^\star$. □

A sufficient condition for improving NOS. Let $c_{\mathrm{NOS}} \in \mathcal{R}$. A sufficient condition for $E_{\widehat{c}_h} < E_{c_{\mathrm{NOS}}}$ is

$$hH_h(\widehat{c}_h;S) + H\varepsilon_{\widehat{c}_h}(h,H) < hH_h(c_{\mathrm{NOS}};S) - H\varepsilon_{c_{\mathrm{NOS}}}(h,H). \tag{29}$$

This follows by applying Equation (11) to the two candidates and strictly separating the resulting upper and lower bounds.

Proof of Proposition 4.4. Let $w = u - \bar{u}$. Periodicity and the L^2 energy identity give

$$\frac{1}{2}\frac{d}{dt}\|w\|_2^2 = -\frac{v}{2}\int \partial_x(w^2)\,dx - D\|\partial_x w\|_2^2 = -D\|\partial_x w\|_2^2 \le 0,$$

hence $\|w(t)\|_2 \le \|w(0)\|_2$. □

Proof of Proposition 4.5. For $w = u - \bar{u}$, $u\partial_x u - \bar{u}\partial_x\bar{u} = \frac{1}{2}\partial_x[(u+\bar{u})w]$. The L^2 estimate is

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|w\|_2^2 + \nu\|\partial_x w\|_2^2 &= -\frac{1}{4}\int (\partial_x u + \partial_x\bar{u})w^2\,dx \le \frac{M_B}{4}\|w\|_2^2, \\ \frac{d}{dt}\|w\|_2^2 &\le \frac{M_B}{2}\|w\|_2^2. \end{aligned}$$

Gronwall's inequality gives $\|w(t)\|_2^2 \le e^{M_B t/2}\|w(0)\|_2^2$; taking square roots proves the stated bound. □

Proof of Proposition 4.6. Let $\xi = \omega - \bar{\omega}$ and $z = v - \bar{v}$. Since $\nabla\cdot v = 0$, subtraction and the L^2 energy identity give

$$\begin{aligned} &\partial_t\xi + v\cdot\nabla\xi + z\cdot\nabla\bar{\omega} = \nu\Delta\xi, \\ &\frac{1}{2}\frac{d}{dt}\|\xi\|_2^2 + \nu\|\nabla\xi\|_2^2 = -\int (z\cdot\nabla\bar{\omega})\xi\,dx \le C_{\mathrm{BS}}M_{\mathrm{NS}}\|\xi\|_2^2. \end{aligned}$$

Here we used $\|z\|_{L^4} \le C_{\mathrm{BS}}\|\xi\|_2$ and Hölder's inequality. Dropping dissipation and applying Gronwall proves the result. □

B Positive Error Memory

For $t \ge 0$, abbreviate

$$s_j(t) := \sum_{k=0}^{j-1} t^k, \qquad V_n(t) := \frac{1}{n}\sum_{j=1}^{n} s_j(t), \qquad \ell_n(c) = \frac{b_c}{a} V_n\Big(\frac{\rho_c}{a}\Big). \tag{30}$$

B.1 Ordered cumulative responses

Lemma B.1 (Ordered cumulative responses)**.** *For $0 \le t_1 \le t_2$, the sequence $n \mapsto V_n(t_2)/V_n(t_1)$ is nondecreasing.*

Proof. For $t \ge 0$,

$$\frac{s_{j+1}(t)}{s_j(t)} = 1 + \frac{t^j}{s_j(t)}, \qquad \frac{d}{dt}\frac{t^j}{s_j(t)} = \frac{t^{j-1}\sum_{k=0}^{j-1}(j-k)t^k}{s_j(t)^2} \ge 0.$$

Hence $s_j(t_2)/s_j(t_1)$ is nondecreasing in j. The ratio $V_n(t_2)/V_n(t_1)$ is its weighted average with weights $s_j(t_1)$. Appending the next, no-smaller ratio cannot decrease that average. □

B.2 Persistence

Iterating (12) from $(s, 0)$ gives

$$F_c^j(s,0) - (a^j s, 0) = \left(0, s b_c \sum_{k=0}^{j-1} a^{j-1-k}\rho_c^k\right). \tag{31}$$

Dividing the error at frame j by the target norm $|s|a^j$ and averaging over frames gives $H_h(c; S) = \ell_h(c)$ and $E_H(c) = \ell_H(c)$.

Proof of Theorem 4.2. If $b_{c_0} = 0$, its replay score is zero. Since $V_h(t) > 0$, every replay minimizer also has $b_c = 0$, and the result follows. Assume $b_{c_0} > 0$.

For $h \ge 2$, $V_h(t)$ is strictly increasing in t. Since $b_{\widehat{c}} \ge b_{c_0}$ and $H_{\mathcal{H}}(\widehat{c}; S) \le H_{\mathcal{H}}(c_0; S)$, it follows that $t_{\widehat{c}} \le t_{c_0}$, where $t_c = \rho_c/a$.

Lemma B.1 shows that

$$\begin{aligned} R_n := \frac{\ell_n(\widehat{c})}{\ell_n(c_0)} &= \frac{b_{\widehat{c}}}{b_{c_0}}\frac{V_n(t_{\widehat{c}})}{V_n(t_{c_0})} \quad \text{is nonincreasing in } n, \\ \frac{H_{\mathcal{H}}(\widehat{c}; S)}{H_{\mathcal{H}}(c_0; S)} &= \frac{\sum_h \ell_h(c_0) R_h}{\sum_h \ell_h(c_0)} \ge R_H = \frac{E_H(\widehat{c})}{E_H(c_0)} \qquad (H \ge \max \mathcal{H}). \end{aligned}$$

□

B.3 Oracle factors and sharpness

Proof of Theorem 4.3. For $t \ge 0$, $V_n(t)$ is nondecreasing in n. For $0 \le t \le 1$, $H \ge h$, and $t_c = \rho_c/a \in [0, 1]$, Lemma B.1 gives

$$\begin{aligned} 1 \le \frac{V_H(t)}{V_h(t)} \le \frac{V_H(1)}{V_h(1)} = \frac{\chi_H}{\chi_h}, \\ \ell_h(c) \le \ell_H(c) \le \frac{\chi_H}{\chi_h}\ell_h(c), \\ H_{\mathcal{H}}(c; S) \le E_H(c) \le \Gamma_{\mathcal{H},H} H_{\mathcal{H}}(c; S). \end{aligned} \tag{32}$$

The last line follows by averaging the left inequality and the rearranged right inequality over $h \in \mathcal{H}$. Writing $\Gamma = \Gamma_{\mathcal{H},H}$, for a future oracle c^*,

$$E_H(\widehat{c}) \le \Gamma H_{\mathcal{H}}(\widehat{c}; S) \le \Gamma H_{\mathcal{H}}(c^*; S) \le \Gamma E_H(c^*).$$

NOS minimizes $\ell_1(c) = b_c/a$. Since $\chi_1 = 1$, the same comparison at $h = 1$ gives $E_H(c_0) \le \chi_H E_H(c^*)$.

For sharpness, take two candidates with memory rates a and zero. Their errors are $b_A\chi_n/a$ and b_B/a. Let $\bar{\chi} = |\mathcal{H}|^{-1}\sum_h \chi_h$. Choose $b_A = r$ and $b_B = r\bar{\chi}(1+\varepsilon)$ with $r, \varepsilon > 0$. Hindcasting selects A, while B is the future oracle, and the error ratio approaches $\Gamma_{\mathcal{H},H}$ as $\varepsilon \downarrow 0$. Taking r small makes both candidates strictly contractive. For one-step sharpness, choose $b_B = r(1+\varepsilon)$; the ratio approaches χ_H. □

For equal weights on $\mathcal{H} = \{5, 10, 15\}$, these factors are

H	One-step factor	Multi-origin factor
16	8.50000	1.54545
34	17.50000	3.18182

In the exact normal form, the longest replay has the smallest single-origin factor because χ_h increases with h. Multi-origin replay addresses origin dependence and model departures that are absent from this normal form.

B.4 Strictly contractive separation from one-step reweighting

Theorem B.2 (Order-H separation)**.** *For every integer $H \geq 2$, set*

$$\begin{aligned} a = 1 - \frac{1}{4H}, \qquad \epsilon = \frac{1}{8H}, \qquad G = aI_2, \\ F_A = \begin{pmatrix} a & 0 \\ \epsilon & a \end{pmatrix}, \qquad F_B = \begin{pmatrix} a & 0 \\ \frac{4}{3}\epsilon & 0 \end{pmatrix}, \qquad u^0 = (1, 0). \end{aligned} \tag{33}$$

All three maps are strictly contractive. Every nonzero nonnegative reweighting of observed one-step squared errors ranks A first, whereas every replay length $h \geq 2$ ranks B first. Candidate B is also the future oracle, and

$$\frac{E_H(A)}{E_H(B)} = \frac{3}{4}\chi_H \geq \frac{3H}{8}, \qquad E_H(A) \geq \frac{1}{16}, \qquad E_H(B) = \frac{1}{6Ha}. \tag{34}$$

Proof. The triangle inequality gives $\|F_A\| \leq a + \epsilon = 1 - 1/(8H) < 1$ and $\|F_B\| \leq a + 4\epsilon/3 = 1 - 1/(12H) < 1$. At $u^t = (a^t, 0)$, the one-step errors are ϵa^t and $4\epsilon a^t/3$, so every common nonnegative weighting ranks A first.

The rollout errors are $\ell_n(A) = (\epsilon/a)\chi_n$ and $\ell_n(B) = 4\epsilon/(3a)$. Since $\chi_h = (h+1)/2 > 4/3$ for every $h \geq 2$, every such replay selects B. Moreover, $E_H(A)/E_H(B) = 3(H+1)/8 \geq 3H/8$. Substituting $\epsilon = 1/(8H)$ into the two rollout errors gives the remaining bounds in (34). □

Candidate A injects less error but retains it, with transverse error $\epsilon j a^{j-1}$ at step j. Candidate B injects more error and erases the previous transverse component, giving $4\epsilon a^{j-1}/3$. Since A has lower error at every observed one-step transition, time reweighting cannot reproduce the hindcast ordering in this family.

B.5 Perturbation robustness

Theorem B.3 (Robust positive-memory guarantee)**.** *Let S_c^0, E_c^0 be reference-model scores satisfying Theorem 4.3, and let S_c, E_c be the actual scores on the same beam. Suppose uniformly in c that $|S_c - S_c^0| \leq \eta_S$ and $|E_c - E_c^0| \leq \eta_F$. Then the selector $\widehat{c} \in \arg\min_c S_c$ satisfies*

$$E_{\widehat{c}} \leq \Gamma E^* + 2\Gamma\eta_S + (\Gamma + 1)\eta_F, \tag{35}$$

where $E^* = \min_c E_c$ and $\Gamma = \Gamma_{\mathcal{H},H}$.

Let c_0 also be a reference-model one-step winner. If $S_{c_0} - S_{\widehat{c}} > 2\eta_S$ and

$$\begin{aligned} r_+ := \frac{S_{\widehat{c}} + \eta_S}{S_{c_0} - \eta_S} < 1, \\ E_{\widehat{c}} \leq r_+ E_{c_0} + (1 + r_+)\eta_F, \end{aligned} \tag{36}$$

Proof. Let c^* minimize the actual future error. The perturbation bounds and the reference-model sandwich yield

$$\begin{aligned} E_{\widehat{c}} &\leq E_{\widehat{c}}^0 + \eta_F \leq \Gamma(S_{\widehat{c}} + \eta_S) + \eta_F \\ &\leq \Gamma(S_{c^*}^0 + 2\eta_S) + \eta_F \leq \Gamma E_{c^*} + 2\Gamma\eta_S + (\Gamma + 1)\eta_F. \end{aligned}$$

For the second statement, the support margin implies $S^0_{\widehat{c}} < S^0_{c_0}$. Theorem 4.2 and the same two-sided perturbation bounds give

$$E_{\widehat{c}} \le E^0_{\widehat{c}} + \eta_F \le r_+ E^0_{c_0} + \eta_F \le r_+ E_{c_0} + (1 + r_+)\eta_F. \tag{37}$$

□

Lemma B.4 (Map perturbation). *Suppose F^0 is q-Lipschitz and $\|F(z) - F^0(z)\|_2 \le \epsilon$ on the visited states. For rollouts starting from the same state and common nonzero target frames $(t^1, \dots, t^n)$,*

$$\begin{aligned} \|F^j z - (F^0)^j z\|_2 &\le \epsilon \sum_{k=0}^{j-1} q^k, \\ \frac{1}{n}\sum_{j=1}^{n} \left|\mathrm{NRMSE}(t^j, F^j z) - \mathrm{NRMSE}(t^j, (F^0)^j z)\right| &\le \frac{\epsilon}{n}\sum_{j=1}^{n} \frac{\sum_{k=0}^{j-1} q^k}{\|t^j\|_2}. \end{aligned} \tag{38}$$

Proof. Let $d_j = \|F^j z - (F^0)^j z\|_2$. Adding and subtracting $F^0(F^j z)$ gives $d_{j+1} \le \epsilon + q d_j$ with $d_0 = 0$. Induction gives the first line of (38); applying it framewise and using the reverse triangle inequality gives the second. □

B.6 Signed memory can reverse the ordering

Proposition B.5 (Contractivity alone does not imply persistence). *Let $a = 1/2$, $\epsilon = 0.05$, $G = aI$, $u^0 = (1, 0)$, and*

$$F_N = \begin{pmatrix} a & 0 \\ \epsilon & 0 \end{pmatrix}, \qquad F_C = \begin{pmatrix} a & 0 \\ 1.6\epsilon & -a \end{pmatrix}. \tag{39}$$

Both candidates are strictly contractive. One-step fitting selects N, two-step hindcasting selects C, and the future oracle at $H = 3$ is N.

Proof. At one step the injection magnitudes are ϵ and 1.6ϵ, so N wins. In units of ϵ/a, the framewise relative error sequences are $(1, 1, 1)$ for N and $(1.6, 0, 1.6)$ for C. Their two-step means are 1 and 0.8, so hindcasting selects C; their three-step means are 1 and $1.6(2/3) > 1$, so the future oracle is N. Finally, $\|F_N\| \le a + \epsilon < 1$ and $\|F_C\| \le a + 1.6\epsilon < 1$. □

The negative transverse multiplier in F_C produces alternating error memory. Positive-semidefinite memory rules out this cancellation and is the structural condition behind Theorem 4.2.

B.7 Operator composition and a PDE realization

For scalar transverse memory, lower-triangular primitive maps compose as

$$\begin{aligned} M_i &= \begin{pmatrix} a_i & 0 \\ b_i & r_i \end{pmatrix}, \\ M_{i_m} \cdots M_{i_1} &= \begin{pmatrix} A_c & 0 \\ B_c & R_c \end{pmatrix}, \qquad A_c = \prod_{k=1}^{m} a_{i_k}, \quad R_c = \prod_{k=1}^{m} r_{i_k}, \\ B_c &= \sum_{k=1}^{m} b_{i_k} \left(\prod_{\ell<k} a_{i_\ell}\right)\left(\prod_{\ell>k} r_{i_\ell}\right). \end{aligned} \tag{40}$$

Thus compositions sharing $A_c = a$ remain in the scalar normal form, while their injection and memory parameters vary separately.

Diffusion gives a concrete instance of nonnegative scalar memory. Let $-\Delta\phi = \lambda\phi$ be a periodic Laplacian mode and consider

$$\partial_t s = \nu \Delta s - \alpha s, \qquad \partial_t z = \nu \Delta z - \mu_c z + \gamma_c s. \tag{41}$$

Its time-τ map on the span of $(\phi, 0)$ and $(0, \phi)$ is lower triangular with

$$a = e^{-(\nu\lambda+\alpha)\tau}, \qquad \rho_c = e^{-(\nu\lambda+\mu_c)\tau},$$
$$b_c = \begin{cases} \gamma_c \dfrac{a-\rho_c}{\mu_c-\alpha}, & \mu_c \neq \alpha, \\ \gamma_c \tau a, & \mu_c = \alpha. \end{cases} \tag{42}$$

For $\mu_c \geq \alpha$ and $\gamma_c \geq 0$, the transverse memory is nonnegative and satisfies $\rho_c \leq a$. Sufficiently small coupling makes the map strictly contractive.

C Experimental Details

Backbone and candidate construction. For Navier–Stokes and advection–diffusion, we use the released NOS implementation and its pretrained DISCO backbone, source codebook construction, time stepping, and conventional beam search (Serrano et al., 2026). The Navier–Stokes model uses the released 17-operator codebook; its search has maximum composition length three, a 1% stopping threshold, and refinement factor four. The advection–diffusion codebook contains 256 source encodings generated with seed 42; its search has maximum composition length five and a 5% stopping threshold. For Burgers, we apply the same frozen NOS pipeline to source environments with $\nu \in \{0.01, 0.1\}$ and allow compositions of up to three operators. Every main experiment uses beam width $B = 4$ and reranks all $R = 4$ surviving candidates with the support-replay score.

Target sets and scoring. The Navier–Stokes evaluation contains eight trajectories at each of 16 log-spaced target viscosities in $[10^{-4}, 10^{-2}]$, for 128 targets in total. Advection–diffusion uses a fixed 128-trajectory subset of the combined advection–diffusion target environment. Burgers uses 128 PDEBench trajectories at the unseen target viscosity $\nu = 0.001$. Each task provides $L = 16$ observed frames. The future horizons are 16, 34, and 16 for Navier–Stokes, advection–diffusion, and Burgers, respectively. Main results use the equal-weight origin set $\mathcal{H} = \{5, 10, 15\}$ and mean framewise NRMSE for candidate fitting, support replay, and future evaluation. Future frames are accessed only after selection. Reported significance uses a two-sided paired Pratt–Wilcoxon signed-rank test over the 128 targets.

Hardware. All candidate search, support replay, and evaluation runs were performed on a single NVIDIA H200 GPU. Model parameters remain frozen throughout; the method requires no training, gradient updates, or target-side optimization.

D Benchmark Equations

Two-dimensional Navier–Stokes. The state is vorticity $\omega(t, x, y)$ on a periodic square. With streamfunction ψ, velocity $\mathbf{v}$, viscosity $\nu > 0$, and $\Delta = \partial_{xx} + \partial_{yy}$, the dynamics are

$$\partial_t \omega + \mathbf{v} \cdot \nabla \omega = \nu \Delta \omega, \qquad \mathbf{v} = (-\partial_y \psi, \partial_x \psi), \qquad -\Delta \psi = \omega. \tag{43}$$

The benchmark composes frozen Euler and diffusion operators following the NOS protocol (Serrano et al., 2026).

One-dimensional advection–diffusion. For a periodic scalar field $u(x, t)$, advection speed v, and diffusivity $D > 0$,

$$\partial_t u + v\, \partial_x u = D\, \partial_{xx} u. \tag{44}$$

The source operator families isolate advection and diffusion, while the combined target activates both mechanisms (Serrano et al., 2026).

One-dimensional Burgers. PDEBench includes the viscous Burgers equation (Takamoto et al., 2022), where $\nu > 0$ is the viscosity:

$$\partial_t u + u\, \partial_x u = \nu\, \partial_{xx} u. \tag{45}$$

E Additional Analyses

Empirical PDE-flow stability. For each target trajectory, we pair its observed states with those of the nearest trajectory in initial-state L_2 and compute the one-step amplification

$$\widehat{a}_t := \frac{\|u^{t+1} - \bar{u}^{t+1}\|_2}{\|u^t - \bar{u}^t\|_2}. \tag{46}$$

Table 5 summarizes these ratios over the observed transitions. For each Burgers pair, we estimate $\widehat{M}_B = \max_t(\|u_x^t\|_\infty + \|\bar{u}_x^t\|_\infty)$; the resulting median, 95th-percentile, and maximum values of $e^{\widehat{M}_B \Delta t/4}$ are 2.43, 5.53, and 8.09. Every one of the 128 target–neighbor pairs satisfies this empirical bound. For Navier–Stokes, $\widehat{M}_{\mathrm{NS}} = \max_t \|\nabla\omega^t\|_{L^4}$ on $[0, 2\pi]^2$. Table 5 reports $\widehat{M}_{\mathrm{NS}}$ and the observed amplification; C_{BS} is domain- and norm-dependent.

Table 5: Empirical stability analysis on the observed benchmark trajectories. Entries report median / 95th percentile / maximum.

PDE	$\widehat{M}$	Observed one-step amplification
Advection–Diffusion	—	0.9786/0.9954/0.9991
Burgers	355.10/684.04/836.43	0.9942/1.0403/1.1774
Navier–Stokes	107.07/252.14/311.26	0.9972/1.0062/1.0283

Test-time cost accounting. The cost axis counts selection-only operator-step equivalents per target. Each candidate score is weighted by its composition length and the 15 observed transitions; the support-hindcast term adds the same weighted count for each of the R retained candidates. The final multi-origin point sums replay lengths $5 + 10 + 15$; its normalized selection cost is 1.06 compared with 1.03 for single-origin $h = 15$. All points share the same post-selection future rollout.

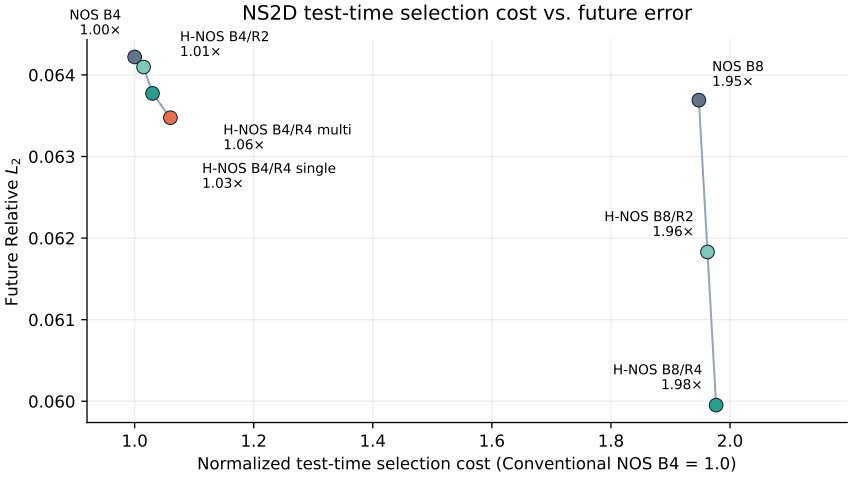


Figure 3: Navier–Stokes selection-cost ablation. The horizontal axis is normalized operator-step cost with Conventional NOS B4 set to 1.0. Candidate evaluations are weighted by composition length and replay length; R2 and R4 replay two and four retained candidates, respectively. The B4/R4 points include both single-origin $h = 15$ and the final multi-origin $\mathcal{H} = \{5, 10, 15\}$ selector. The common post-selection future rollout is excluded from all points. B8 rows are single-origin beam-width ablations.

Beam width and retained candidate pool. For this controlled beam ablation, all rows use single-origin $h = 15$ replay. The B4/R4 row compares against the same B4 Identity candidate set. Table 6 then varies the number R of retained candidates under the wider B8 search. Increasing R from two to

four raises the reduction from 2.92% to 5.87%, indicating that the wider beam is most useful when hindcasting can inspect more of its surviving alternatives.

Table 6: **Single-origin Navier–Stokes beam/pool ablation over 128 targets.** Each row reranks the top R candidates from the indicated NOS beam and is compared with the one-step NOS winner at the same B.

B	R	NOS	Hindcast	Reduction	Pratt p
4	**4**	**0.0642205**	**0.0637733**	**0.696%**	$\mathbf{5.57\times10^{-4}}$
8	2	0.0636905	0.0618300	2.92%	1.41×10^{-5}
8	4	0.0636905	**0.0599526**	**5.87%**	$\mathbf{1.49\times10^{-5}}$

Replay and deployment horizons. Figure 4 holds the primary Identity B4/R4 candidate pool fixed and examines replay horizons $h \in \{11, \dots, 15\}$ at the full deployment horizon $F = 16$ on all 128 targets. This analysis uses mean framewise NRMSE for both reranking and future evaluation. The improvement remains positive throughout this range and increases to 0.602% at $h = 15$, showing stronger selection gains with longer support replay.

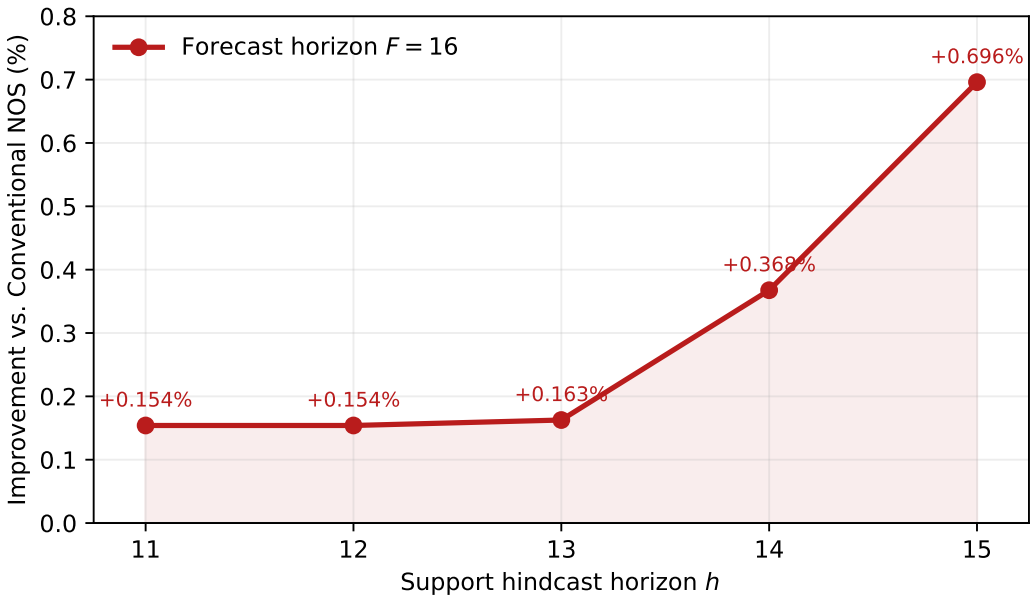


Figure 4: Navier–Stokes B4/R4 horizon sensitivity with a fixed candidate pool and deployment horizon $F = 16$. Values are paired mean improvements over Conventional NOS across 128 targets as the support hindcast horizon varies.

Origin-set sensitivity. We further rerank the same Navier–Stokes B4/R4 candidate pool using every nonempty subset of the three replay lengths $\{5, 10, 15\}$, with equal weights for multi-origin sets. At the full forecast horizon, all seven origin sets improve over conventional NOS, demonstrating robustness to replay-origin choice.

Table 7: Navier–Stokes origin-set ablation on the fixed B4/R4 candidate pool over 128 targets at forecast horizon $F = 16$. Reduction is relative to conventional NOS (0.064220); higher is better.

Origin set $\mathcal{H}$	Mean framewise NRMSE	Reduction
$\{5\}$	0.063790	0.670%
$\{10\}$	0.063894	0.509%
$\{15\}$	0.063773	0.696%
$\{5, 10\}$	0.063791	0.668%
$\{10, 15\}$	0.063866	0.552%
$\{5, 15\}$	0.063668	0.861%
$\{5, 10, 15\}$	**0.063475**	1.160%

Single- versus multi-origin replay. For a cross-PDE comparison, we align single- and multi-origin HINDCAST-NOS on the same final conventional beam. All three tasks compare single-origin $h = 15$ with multi-origin $\mathcal{H} = \{5, 10, 15\}$. Multi-origin replay increases the reduction from 0.696% to 1.160% on Navier–Stokes, from 2.452% to 2.476% on advection–diffusion, and from 0.589% to 0.611% on Burgers.

Table 8: Single- versus multi-origin HINDCAST-NOS over the three PDE tasks. All rows use the same conventional final beam within a task. Multi-origin selection averages equal-weight closed-loop errors on observed support. Parentheses report relative reduction from the fixed conventional beam.

PDE	Fixed conventional beam	Single-origin	Multi-origin
Navier–Stokes	0.064220	0.063773 (+0.696%)	**0.063475** (+1.160%)
Advection–Diffusion	0.014708	0.01434720 (+2.452%)	**0.01434371** (+2.476%)
Burgers	0.220064	0.218768 (+0.589%)	**0.218720** (+0.611%)

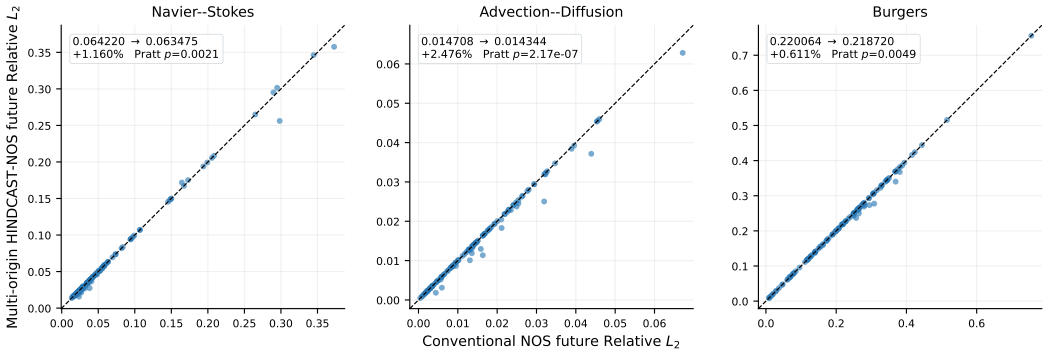


Figure 5: Per-target future mean framewise NRMSE for Conventional NOS and the final multi-origin HINDCAST-NOS selectors. The dashed diagonal denotes equal error. Aggregate means, relative improvements, and two-sided Pratt–Wilcoxon p-values are shown in each panel.

Table 9: Largest-reduction examples from the final multi-origin B4/R4 evaluation. Lower mean framewise NRMSE is better.

Task	Case	NOS beam	Hindcast	Reduction
Advection–diffusion	Maximum	0.03195	0.02503	21.64%
Navier–Stokes	Maximum	0.29834	0.25611	14.16%
Burgers	Maximum	0.30833	0.27744	10.02%

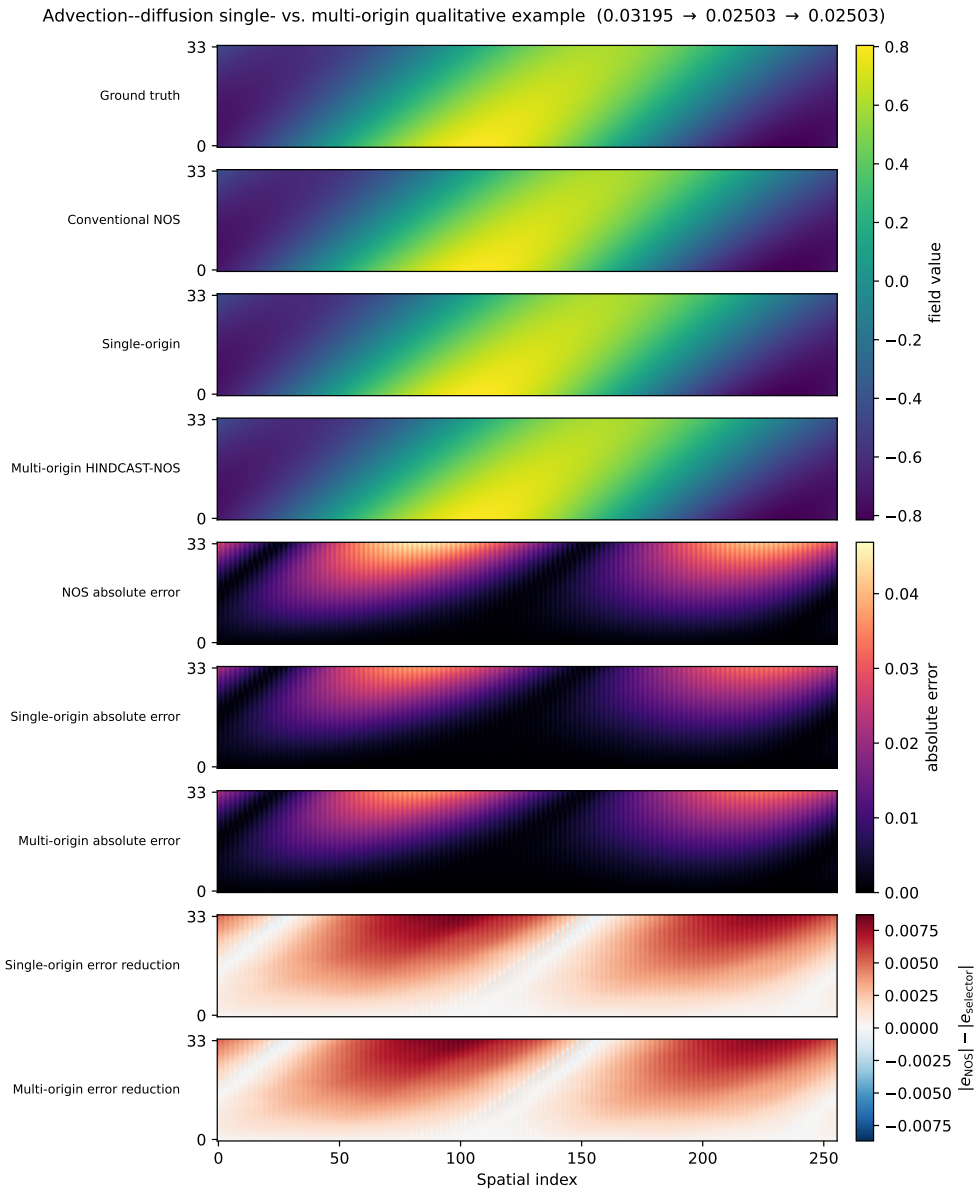


Figure 6: Maximum-improvement advection–diffusion case under multi-origin, with single-origin shown on the same target. Conventional NOS uses $[46, 38]$; both selectors choose $[7, 38]$ and reduce future NRMSE from 0.03195 to 0.02503. Field, absolute-error, and signed-reduction rows use shared scales within their respective groups.

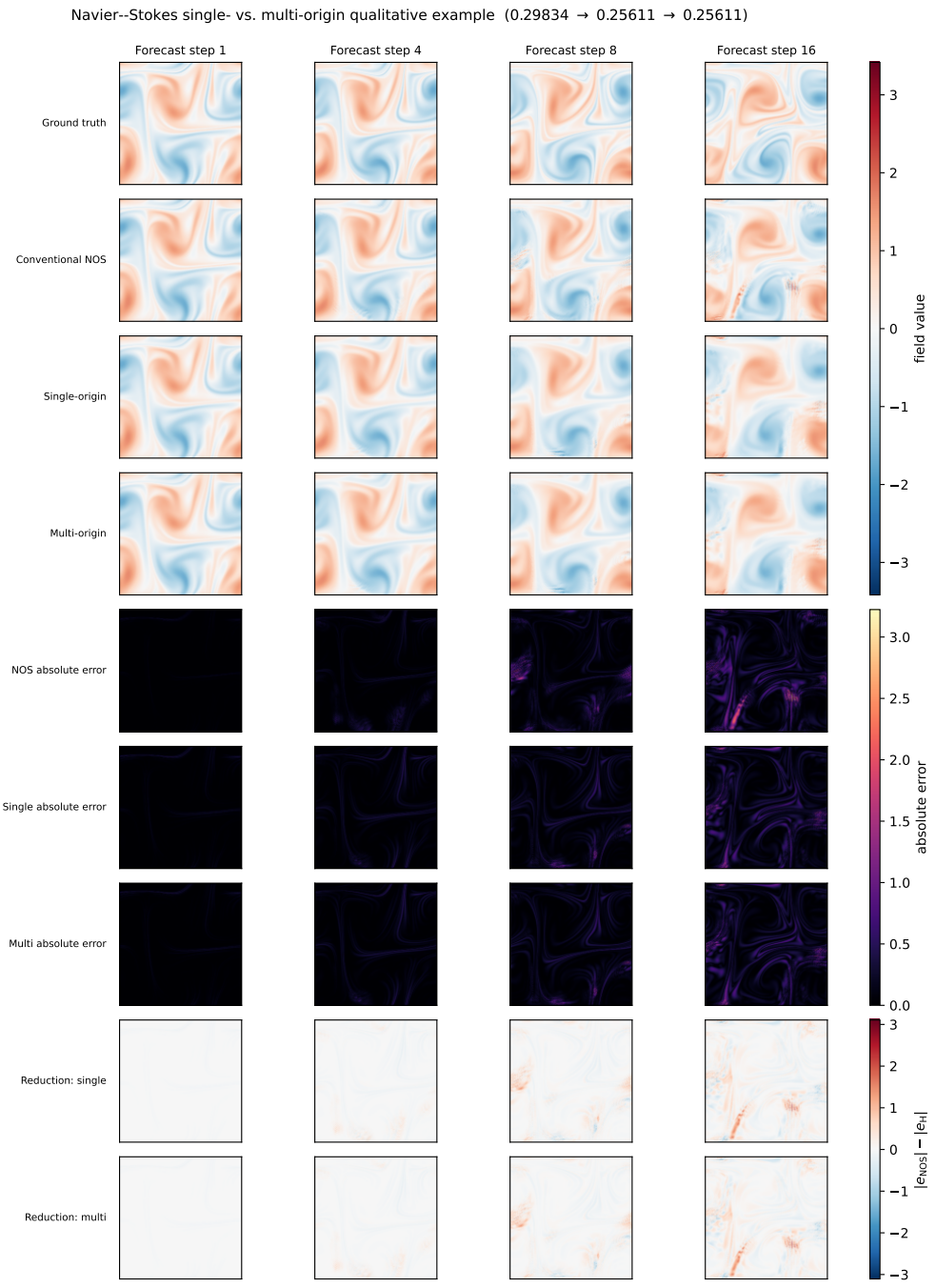


Figure 7: Maximum-improvement Navier–Stokes B4/R4 case under multi-origin $\mathcal{H} = \{5, 10, 15\}$, with single-origin shown on the same target. Columns show forecast steps 1, 4, 8, and 16. Conventional NOS uses $[12, 0]$; both selectors choose $[15, 0]$ and reduce future NRMSE from 0.29834 to 0.25611. Field, absolute-error, and signed-reduction rows use shared scales within their respective groups.

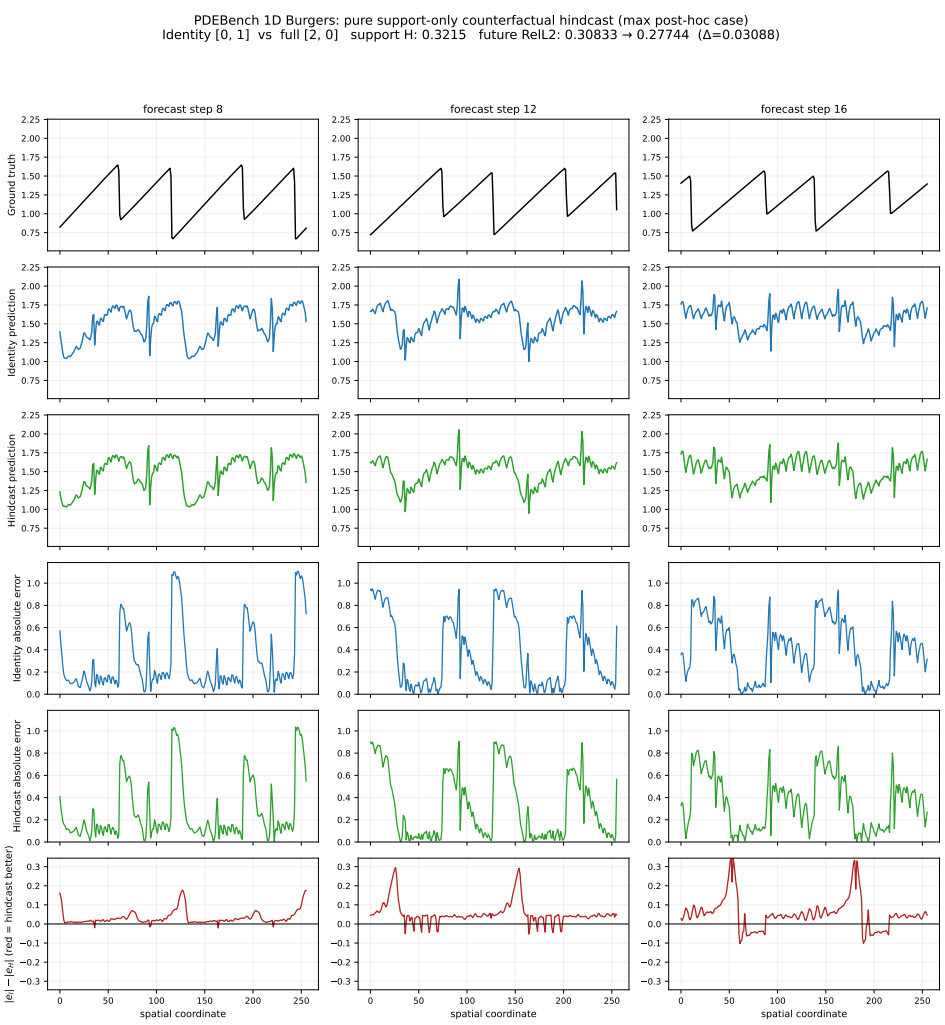


Figure 8: Maximum-improvement Burgers case under multi-origin $\mathcal{H} = \{5, 10, 15\}$. Reranking changes the composition from $[0, 1]$ to $[2, 0]$ and reduces future NRMSE from 0.30833 to 0.27744. The selected composition is the same as in the single-origin maximum case; the shared-scale error profiles localize the improvement near sharp transitions.