CONENO: ENFORCING PHYSICAL INFORMATION FLOW IN NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

For finite-speed PDEs, the solution at a point depends only on data inside its backward cone. Standard neural operators need not respect this constraint: a global Fourier operator trained to low error changes its prediction when data outside the cone is edited while the reference stays fixed, and a mask applied after global mixing leaves the dependence in place. We introduce CONENO, a neural operator in which every read, including the inputs that determine its attention weights, stays inside a prescribed domain of dependence. We prove that this property is closed under composition and characterize what a Jacobian audit can and cannot certify. Under matched capacity and supervision on variable-coefficient transport and a finite-volume Burgers flow map, CONENO lowers error by 26–87% relative to convolutional and local-window Fourier operators that satisfy the same constraint. The constraint costs no in-distribution accuracy: CONENO matches an unconstrained global FNO on transport and reduces its error by 45–80% on Burgers, while having zero response to exterior interventions under which the error of global FNO grows nearly tenfold.

1 INTRODUCTION

Many-query workflows such as inverse design and control need PDE solutions for many initial data or coefficients, and neural operators answer this by learning a shared map $\mathcal{G}_\theta : \mathcal{A} \to \mathcal{U}$ across a family of inputs (Lu et al., 2019; Li et al., 2021; Kovachki et al., 2023). Spectral, convolutional, graph and attention realizations improve approximation and discretization transfer (Li et al., 2021; Raonic et al., 2023; Brandstetter et al., 2022; Hao et al., 2023; Wu et al., 2024), but the standard global realizations do not specify which input locations may influence a prediction. Physics does. For a hyperbolic equation with speed bounded by c, the solution at (x, t) depends only on the backward cone

$$\mathcal{D}_c(x,t) = \big\{(x',t') : \ 0 \le t' \le t, \ \|x - x'\| \le c(t - t')\big\}. \tag{1}$$

Two inputs that agree on $\mathcal{D}_c(x, t)$ must produce the same solution there (Courant & Hilbert, 1962; Evans, 2022; Lax, 2006):

$$a|_{\mathcal{D}_c(x,t)} = a'|_{\mathcal{D}_c(x,t)} \quad \Longrightarrow \quad \mathcal{G}(a)(x,t) = \mathcal{G}(a')(x,t). \tag{2}$$

We call this a *physical information-flow constraint*. Prediction loss does not enforce it: distant correlations can let a model fit training trajectories with exterior reads, and a cone mask applied after global mixing leaves that dependence in the hidden tokens. A surrogate that violates the constraint will respond to edits that cannot physically matter, so we measure solution error together with the response to exterior edits.

CONENO enforces the constraint by construction, throughout feature construction, mixing and decoding. Delayed compact transport builds admissible features; attention mixes them and reads raw field values with query-dependent weights. Both the values and the inputs that set those weights stay inside the prescribed cone, so the operator remains query-adaptive without becoming global.

Primary comparisons use variable-coefficient transport and a Burgers finite-volume flow map, with five seeds and matched capacity and supervision. CONENO has lower mean error than local-window Fourier and convolutional controls on all tested conditions, matches global FNO on in-distribution transport, and has the lowest error of all five operators on Burgers. Exterior interventions that leave the reference unchanged alter the unconstrained predictions and leave CONENO's fixed. Section 6.6 repeats the comparison under a disk mask on 2D advection and the wave equation.

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.

Contributions.

1. We formulate the constraint as an inductive bias and give conditions for preserving it across the computational graph, including input-dependent mixing weights.
2. We introduce CONENO: delayed compact features with attention over admissible latent tokens and raw field values.
3. Under matched supervision, CONENO is more accurate than locality-constrained alternatives on transport and Burgers, with no in-distribution penalty versus global FNO. Two-dimensional disk-mask tests repeat the comparison; Jacobian audits and finite interventions confirm the constraint.

2 RELATED WORK

Operator learning. DeepONet and the neural-operator framework give approximation foundations (Lu et al., 2019; Kovachki et al., 2023; 2021; Lanthaler et al., 2022). Practical classes are spectral (Li et al., 2021; Tran et al., 2021; Li et al., 2023), convolutional (Raonic et al., 2023; Rahman et al., 2023), graph-based (Li et al., 2020; Brandstetter et al., 2022; Pfaff et al., 2020) and attention-based (Li et al., 2022; Hao et al., 2023; Wu et al., 2024). In all of them the input support of a prediction is a by-product of the architecture: spectral and attention models mix globally, while convolutional and graph models have a receptive field set by depth rather than by the equation.

Locality and causality in learned solvers. Locality or causality has been added to learned solvers in several forms: windowed prediction, DDELD (Chen et al., 2025), propagation-aware convolution (Ma et al., 2025), causal training weights (Wang et al., 2024), causal architectures for infinite-dimensional dynamics (Galimberti et al., 2026), and learned discretizations that inherit a stencil (Kochkov et al., 2021). Dilated causal convolutions impose the temporal analogue in sequence models (Van Den Oord et al., 2016; Bai et al., 2018). These designs fix an output window or the time direction; the spatial reads that produce data-dependent weights stay unconstrained. We treat the space–time cone as a hard constraint on every read, including weight generation, and audit it with nodal Jacobians and reconstruction-compatible interventions (Appendix E).

3 PROBLEM: ADMISSIBLE INPUT DEPENDENCE

Let $\Omega \subset \mathbb{R}^d$, $I = [0, T]$, and let $\mathcal{G}^\dagger : \mathcal{A} \to \mathcal{U}$ map PDE data to the solution. We take a conservative speed bound c valid across the input family. The requirement (2) applies even to inputs unlike the training examples, provided they stay in the PDE family and the speed bound.

Definition 3.1 (Domain of dependence)**.** A set-valued map $\mathcal{D}$ is a *domain of dependence* for an operator if agreement of inputs on $\mathcal{D}(x,t)$ implies agreement of predictions at (x,t), as in (2). We call such an operator $\mathcal{D}$-preserving. The learning problem is to approximate $\mathcal{G}^\dagger$ within a $\mathcal{D}$-preserving class.

The constraint couples space and time: temporal causality still permits global spatial reads, and a fixed neighborhood does not grow with elapsed time. It is an upper bound on relevant information, not a formula for the characteristic path, so the same cone can serve advection, a bounded conservation law and the wave equation. On a periodic domain, late cones cover the initial field; early-time queries reveal forbidden spatial dependence. We study these finite-speed hyperbolic problems with a known bound. On sampled fields we audit the discrete operator on its grid, and we intervene on tasks whose reference is unchanged by an exterior edit, such as grid-aligned advection and the Burgers map (Appendix A.1).

Definition 3.2 (Nested cones)**.** A domain of dependence is *nested* if

$$(x', t') \in \mathcal{D}(x, t) \quad \Longrightarrow \quad \mathcal{D}(x', t') \subset \mathcal{D}(x, t).$$

For hyperbolic systems with speed bounded by c, the backward light cone (1) is nested (John, 2012; Evans, 2022). For the implemented discrete operator, let $I_{\theta,ij}(a) = \partial \mathcal{G}_{\theta,i}(a)/\partial a_j$, where i identifies

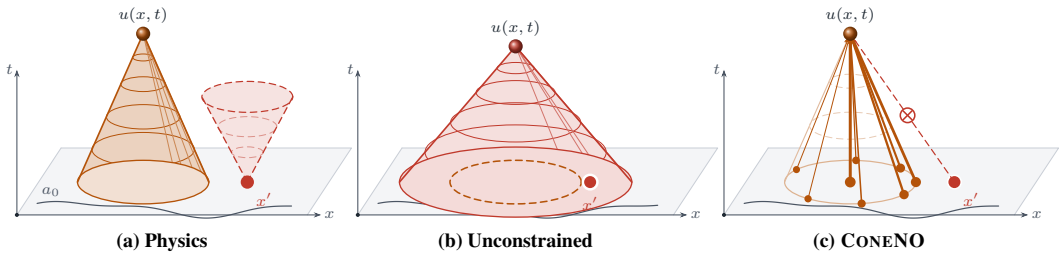


Figure 1: (a) Physics: the backward cone of $u(x,t)$ meets a_0 in a disk $\mathcal{D}_c$; the forward cone of an exterior edit x' misses the query, so x' cannot influence u. (b) An unconstrained operator opens a wider cone that contains x' (dashed: physical foot). (c) ConeNO samples only the physical foot, with learned weights on those reads, and blocks x'. Section 6.5 measures violations.

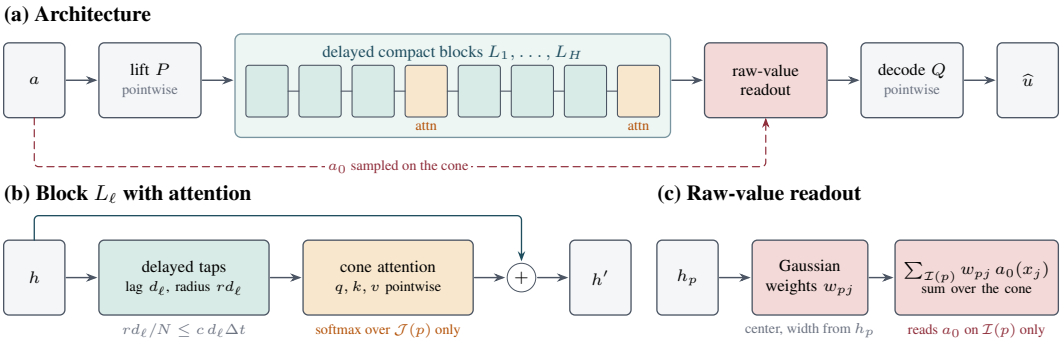


Figure 2: Architecture of ConeNO. (a) A pointwise lift, a stack of delayed compact blocks with cone attention at selected depths, a raw-value readout of the initial field, and a pointwise decoder. (b) Inside a block: delayed taps under the CFL bound, cone attention with softmax restricted to $\mathcal{J}(p)$, and a residual update. (c) The readout turns the query feature into Gaussian weights over $\mathcal{I}(p)$ and sums raw field values there.

a scalar output query and j an input coordinate and channel. Let D_i contain the admissible input indices. On the uniform grids used here, we measure the cone-violation score

$$\mathrm{CVS}(\mathcal{G}_\theta) = \mathbb{E}_{a,i}\left[\frac{\sum_{j\notin D_i} |I_{\theta,ij}(a)|^2}{\sum_j |I_{\theta,ij}(a)|^2}\right]. \tag{3}$$

The ratio is defined at probes with nonzero total influence. At such probes, a $\mathcal{D}$-preserving operator has $\mathrm{CVS} = 0$. Proposition 5.4 relates the score to finite changes. Appendix A records the influence construction.

4 ConeNO

The architecture fixes which inputs may influence a prediction; ConeNO learns how to weight them. Figure 1 illustrates the constraint. Figure 2 is the computational graph: a pointwise lift and decoder sandwich a stack of delayed compact blocks, with cone attention at selected depths and a raw-value readout of the initial field. The stacked operator is

$$\mathcal{G}_{\theta,N} = Q \circ L_H \circ \cdots \circ L_1 \circ P. \tag{4}$$

Admissible feature construction. A late attention mask cannot remove exterior information already mixed into its inputs. We therefore construct features through compact transport paths with

positive time lags. On a grid $x_i = i/N$, $t_m = m\Delta t$, a block with dilation d_ℓ reads spatial offsets up to rd_ℓ only at lag d_ℓ. Every such edge stays inside the cone when

$$\frac{rd_\ell}{N} \le c_{\mathrm{mask}} d_\ell \Delta t. \tag{5}$$

Increasing dilations expand the accessible past without same-time spatial mixing. Residual updates and channel normalization are pointwise. With a pointwise decoder, these paths are our convolutional instance of the constraint. CONENO adds the content-dependent aggregation below. Copying initial data only to its own spatial coordinate supplies early-time features without a spatial shortcut.

Attention within the admissible past. For query p, let $\mathcal{J}(p)$ be sampled past tokens inside its cone. A head computes

$$A_p^h = \sum_{q \in \mathcal{J}(p)} \alpha_{pq}^h v_q^h, \qquad \alpha_{pq}^h = \frac{\exp(e_{pq}^h)}{\sum_{q' \in \mathcal{J}(p)} \exp(e_{pq'}^h)}, \quad e_{pq}^h = \frac{(q_p^h)^\top k_q^h}{\sqrt{d_h}} + B_{pq}^h. \tag{6}$$

Queries, keys and values are pointwise projections of admissible features, so both values and weights respect the constraint. The hard mask excludes forbidden reads; an optional geometric bias B_{pq}^h changes only weights inside it. The models of Section 6 use no latent-attention bias; the supplementary linear trajectory models use the ray bias of Appendix B.3. A gated residual writes each output back to its query.

Reading raw field values. For repeated state-to-state prediction, we also attend to unprojected field values. Learned features determine where to read, while the sampled amplitudes remain available to the decoder. Let h_p be the final latent feature and $\mathcal{I}(p)$ the sampled initial positions inside its cone. Four query-conditioned heads produce

$$\bar{a}_p = \sum_{h=1}^{4} \pi_p^h \sum_{j \in \mathcal{I}(p)} w_{pj}^h a_0(x_j), \qquad \widehat{u}_p = b_p + g(\bar{a}_p - b_p). \tag{7}$$

Here $b_p = Q_0(h_p)$, $g = \tanh \gamma$ is learned, and π_p is a softmax mixture. Each w_{pj}^h is a normalized Gaussian preference with center and width predicted from h_p, restricted to $\mathcal{I}(p)$. Appendix B.2 gives the remaining equations.

Query resolution and learning. Every output location keeps its own query; only keys and values are subsampled. Residue classes avoid interpolation between hidden locations. Under integer refinement, tap and residue offsets scale together, so shared nodes are unchanged (Proposition 5.5). Training minimizes relative L^2 error. The main one-dimensional and two-dimensional models use latent and raw-value attention; the supplementary one-dimensional trajectory studies use latent attention only (Appendix D).

5 GUARANTEES

The information-flow guarantee concerns the complete computational graph: not only which values are aggregated, but also which inputs determine the aggregation weights. Proofs of the statements in this section are in Appendix C.

Assumption 5.1 (Standing)**.** *The domain of dependence $\mathcal{D}$ is prescribed independently of the input, contains its own query, and is nested as in Definition 3.2. Lifting, decoding, and channel normalization are pointwise. Every input read by the lift lies in the query's domain; same-coordinate initial-data copying satisfies this for $\mathcal{D}_c$. All layer operations are well defined.*

Proposition 5.2 (Closure of the information-flow constraint under composition)**.** *Under Assumption 5.1, suppose every update at $z = (x, t)$ reads hidden values only at coordinates in $\mathcal{D}(z)$. Any input-dependent weights, gates, and normalizers used by that update must also depend only on those values. Then the stacked operator* (4) *is $\mathcal{D}$-preserving. Compact taps satisfying* (5)*, pointwise queries/keys/values, and attention normalized only over hard-masked admissible keys meet these conditions. Raw initial values may be retained as identity channels of the lift.*

Proposition 5.2 licenses residual depth and query-adaptive mixing inside $\mathcal{D}(z)$. It excludes global normalization, unmasked spatial mixing, and any gate or kernel weight computed from the full field. Once a layer mixes space globally, a later cone mask cannot restore the guarantee.

The constraint has to be built in, because low training error does not select an admissible operator.

Proposition 5.3 (A duplicated-pattern shortcut)**.** *For linear advection with speed c on the unit torus, let the training measure be supported on half-periodic inputs, $u_0(x + 1/2) = u_0(x)$. Then $\mathcal{G}_{\mathrm{spur}}(u_0)(x,t) := u_0(x - ct + 1/2)$ is a Fourier multiplier with zero training error, and for $0 < t < 1/(4c)$ its entire influence lies outside $\mathcal{D}_c(x,t)$. A localized input that breaks the duplication produces $O(1)$ error.*

A hard-cone CONENO cannot realize $\mathcal{G}_{\mathrm{spur}}$ on the full input class; the localized-input tests of Section 6 probe this failure mode.

CVS is the Jacobian form of the constraint; finite edits test the same restriction at finite amplitude.

Proposition 5.4 (Finite out-of-cone interventions)**.** *Let $f_i : \mathbb{R}^p \to \mathbb{R}$ be a discrete scalar prediction, continuously differentiable on a neighborhood of the segment $a + s\delta$, $s \in [0,1]$. If $\delta_j = 0$ for $j \in D_i$, then*

$$|f_i(a+\delta) - f_i(a)| \le \|\delta\|_2 \int_0^1 \|\nabla_{D_i^c} f_i(a + s\delta)\|_2 \, \mathrm{d}s. \tag{8}$$

Zero exterior gradients along every admissible constant-interior segment imply invariance to finite exterior changes; a zero derivative at a single input does not.

Section 6.5 reports both: zero CVS and zero response to the exterior edits. Appendix C.4 splits paired risk into shared approximation error and an exterior-response term. The SR2 condition of Section 6 evaluates trained models on a twice-finer grid. Scaling tap, residue and readout offsets with the refinement factor makes this exact at the shared nodes.

Proposition 5.5 (Shared-node consistency under integer refinement)**.** *Let R_f restrict a field on an fN-point periodic grid to every f-th spatial node, and use the same time grid and weights at both resolutions. Scale compact tap offsets, spatial residue strides and raw-value read offsets by the integer f, with all other operations as in §4. In exact arithmetic,*

$$R_f \mathcal{G}_{\theta, fN}(a_f) = \mathcal{G}_{\theta, N}(R_f a_f) \tag{9}$$

for every fine-grid input a_f. The same statement holds for uniform refinement of both axes in the two-dimensional model.

Appendix C.5 gives the proof and float64 checks of (9).

6 EXPERIMENTS

We ask whether query-adaptive mixing helps when access, supervision and parameter count are matched, and whether the constraint costs accuracy versus unconstrained operators. One-dimensional comparisons come first, then disk-mask tests on 2D advection and the wave equation. All tasks share one protocol (Appendix D).

6.1 ONE-DIMENSIONAL SETUP

Subsections 6.2–6.5 use this setup.

Tasks. Both tasks are periodic on 256 nodes, with 512 / 128 / 128 inputs. *Variable-coefficient transport* solves $q_t + c(x)q_x = 0$ with joint input (q, c) and $0.75 \le c \le 1.25$; the reference follows characteristics. *Burgers* solves $u_t + (u^2/2)_x = 0$ with $|u_0| \le 0.75$; the reference is a Godunov finite-volume map whose stencil lies inside the cone.

Models. Five operators with 259–265k parameters: global FNO; a *local-window FNO*, DDELD (Chen et al., 2025), that applies the same Fourier core to each admissible window and keeps the center; DeepONet with 256 branch sensors and a trunk on (x, τ); a *convolutional control* that keeps the compact path of CONENO and replaces each attention update by a pointwise MLP (Appendix B.4); and CONENO with latent and raw-value attention. Local-window FNO, the convolutional control and CONENO satisfy the constraint by construction.

Protocol. Each model learns the same $\tau = 1/9$ transition, 80 epochs, five seeds. Rollout has no reference resets. *ID* uses the training family; *Bump* uses localized fields; *Coeff.* (transport only) replaces speed modes 1–3 by 4–6; *SR2* doubles the grid; $T = 2$ uses eighteen self-fed steps. Errors are mean per-trajectory relative L^2 $\times 100$, excluding the initial frame, as mean $\pm$ SD over seeds. CVS follows (3). Interventions replace exterior values by a half mixture with a donor, leaving the cone and the reference unchanged.

6.2 Matched locality and mixing comparison

Table 1: Variable-coefficient transport, relative L^2 $\times\mathbf{100}$ (mean $\pm$ SD over five seeds; 128 inputs per condition).

Model	Params (k)	ID	Bump	Coeff.	SR2	$T = 2$
Global FNO	258.9	$\mathbf{1.24 \pm 0.07}$	22.65 ± 5.35	19.32 ± 0.85	$\mathbf{1.24 \pm 0.08}$	$\mathbf{2.09 \pm 0.10}$
Local-window FNO	258.9	2.17 ± 0.10	1.85 ± 0.29	12.27 ± 0.37	9.91 ± 0.58	3.61 ± 0.20
DeepONet	264.0	8.71 ± 0.13	9.49 ± 3.63	8.55 ± 0.13	8.71 ± 0.13	9.35 ± 0.15
Conv. control	264.3	2.08 ± 0.17	2.56 ± 0.61	7.98 ± 0.49	2.07 ± 0.17	3.72 ± 0.33
ConeNO	264.9	1.26 ± 0.08	$\mathbf{1.08 \pm 0.26}$	$\mathbf{5.92 \pm 0.42}$	1.25 ± 0.07	2.14 ± 0.13

Table 1 separates locality from the mixer. ConeNO has lower mean error than both restricted controls on all five conditions: 26–58% below the convolutional control and 41–87% below local-window FNO. On ID, SR2 and $T = 2$ it lies within one seed SD of global FNO; global FNO transfers poorly to bumps and coefficient shift. DeepONet is worse than the local models on every condition.

6.3 Nonlinear conservation dynamics

Table 2: Burgers finite-volume flow map, relative L^2 $\times\mathbf{100}$ (mean $\pm$ SD over five seeds; same budgets as Table 1).

Model	Params (k)	ID	Bump	SR2	$T = 2$
Global FNO	258.9	5.04 ± 0.18	5.73 ± 0.34	6.67 ± 0.12	7.10 ± 0.15
Local-window FNO	258.9	2.85 ± 0.23	1.67 ± 0.36	6.07 ± 1.46	4.02 ± 0.27
DeepONet	264.0	4.94 ± 0.08	5.27 ± 0.04	5.01 ± 0.08	5.35 ± 0.12
Conv. control	264.3	4.02 ± 0.56	4.65 ± 1.31	5.82 ± 0.50	6.09 ± 0.93
ConeNO	264.9	$\mathbf{1.86 \pm 0.18}$	$\mathbf{1.15 \pm 0.13}$	$\mathbf{3.64 \pm 0.24}$	$\mathbf{2.74 \pm 0.20}$

Table 2 uses the same five operators on the Burgers map. Here the unconstrained operators are no longer competitive: ConeNO has the lowest mean error on every condition, 37–75% below the convolutional control, 31–40% below local-window FNO, 45–80% below global FNO and 27–78% below DeepONet. All three restricted models are invariant to exterior changes (Table 3), so the ranking among them reflects the mixer, not invariance.

6.4 Accuracy throughout autoregressive rollout

The advantage over the convolutional control and local-window FNO holds at every horizon (Figure 3). At $T = 2$, error is 42% and 55% below the convolutional control on transport and Burgers; on transport ConeNO tracks global FNO, and on Burgers it stays below every baseline.

6.5 Measured information dependence

The accuracy tables compare mixers; Table 3 tests the restriction itself. Local-window FNO, the convolutional control and ConeNO have zero measured CVS and zero response, with nonzero interior sensitivity. Global FNO and DeepONet respond to exterior changes on both tasks. On transport, global FNO and ConeNO have the same clean query error (0.0067 and 0.0068); the

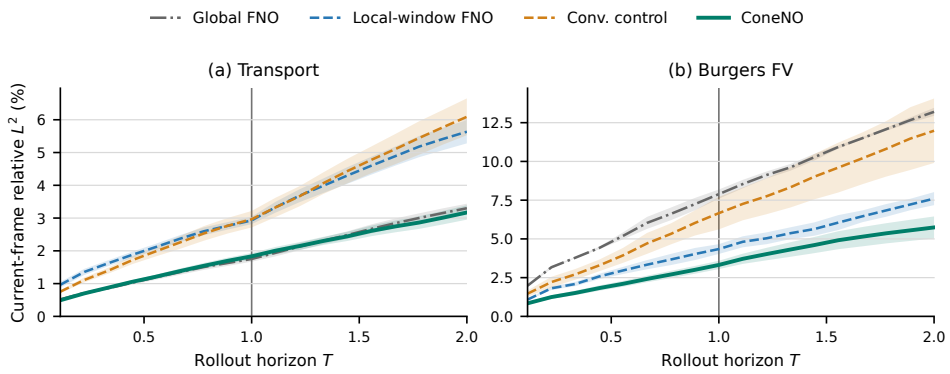


Figure 3: Current-frame relative L^2 ($\times 100$) during self-fed rollout on 128 ID inputs. Bands are SD over five seeds; the vertical mark is the training horizon.

Table 3: CVS and exterior interventions for the models of Tables 1 and 2 (mean $\pm$ SD over five seeds). Response RMS is the pooled prediction change at the query; clean and changed RMSE are absolute query errors before and after the exterior change, on the scale of the state rather than the $\times 100$ relative errors above.

Model	CVS	Response RMS	Clean RMSE	Changed RMSE
Variable-coefficient transport				
Global FNO	0.0492 ± 0.0062	0.0628 ± 0.0050	0.0067 ± 0.0004	0.0636 ± 0.0051
Local-window FNO	0	0	0.0134 ± 0.0014	0.0134 ± 0.0014
DeepONet	0.2299 ± 0.0083	0.0152 ± 0.0007	0.0727 ± 0.0010	0.0710 ± 0.0009
Conv. control	0	0	0.0104 ± 0.0006	0.0104 ± 0.0006
ConeNO	0	0	0.0068 ± 0.0005	0.0068 ± 0.0005
Burgers finite-volume flow map				
Global FNO	0.0087 ± 0.0008	0.0039 ± 0.0002	0.0067 ± 0.0005	0.0075 ± 0.0001
Local-window FNO	0	0	0.0051 ± 0.0005	0.0051 ± 0.0005
DeepONet	0.1131 ± 0.0061	0.0042 ± 0.0001	0.0130 ± 0.0001	0.0133 ± 0.0001
Conv. control	0	0	0.0053 ± 0.0007	0.0053 ± 0.0007
ConeNO	0	0	0.0033 ± 0.0002	0.0033 ± 0.0002

exterior change raises global FNO's by an order of magnitude and leaves ConeNO's unchanged. On Burgers, ConeNO has the lowest clean error of all five models. DeepONet's transport error decreases slightly after the change; Appendix C.4 accounts for that shift by the disagreement term.

6.6 Two-dimensional setup

The one-dimensional cone has an interval foot. Subsections 6.7–6.8 test the disk geometry of Figure 1 on two periodic equations, with two-dimensional models and grids.

Tasks. Both tasks use $c_{\mathrm{mask}} = 1.5$. *Advection* solves $\partial_t u + \mathbf{v} \cdot \nabla u = 0$ with $\mathbf{v} = (1,1)/\sqrt{2}$ on a $32 \times 32 \times 7$ grid; each stored step is a one-cell translation, so the nodal cone is exact (Appendix D.2.2). *Wave* solves $\partial_{tt} u = c^2 \Delta u$ with $c = 1$ on a 64×64 grid, mapping Cauchy data $(u, \partial_t u)$ through one step of length $\tau = 1/9$ (Appendix D.2.1).

Models. Four operators, matched to the width-16 backbone of Appendix B.1: global FNO-XY; a disk-window local Fourier model; the convolutional control with delayed 3×3 taps; and disk-masked ConeNO-2D with latent and raw-value attention.

Protocol. Conditions are ID, Bump, SR2, $T = 2$ and CVS, as in Section 6.1. Advection is trained on the stored trajectory and evaluated on 64 test fields; $T = 2$ doubles the stored length. Wave learns

the $\tau = 1/9$ step with 512 / 128 / 128 inputs, relative L^2 on $(u, \partial_t u)$, and eighteen-step $T = 2$. Both report mean $\pm$ SD over five seeds.

6.7 GRID-ALIGNED 2D ADVECTION

Table 4 and Figure 4 apply the comparison to the translation map. CONENO-2D is lowest on ID, Bump, SR2 and $T = 2$: 28–89% below FNO-XY, 29–41% below the convolutional control and 17–59% below local-window FNO. The restricted models have CVS 0; FNO-XY has CVS 0.411.

Table 4: Grid-aligned 2D advection, relative L^2 $\times\mathbf{100}$ on 64 test trajectories (mean $\pm$ SD over five seeds; matched width-16 backbone).

Model	Params (k)	ID	Bump	SR2	$T = 2$	CVS
FNO-XY	139.4	2.54±0.21	26.41±2.83	68.2±9.14	24.7±1.86	0.411
Local-window FNO	138.6	3.41±0.24	7.92±0.91	17.6±2.08	11.3±1.17	0
Conv. control	139.5	3.06±0.27	9.85±1.21	11.4±1.46	7.62±0.89	0
CONENO-2D	140.1	**1.82±0.13**	**6.57±0.58**	**7.18±0.69**	**5.44±0.47**	0

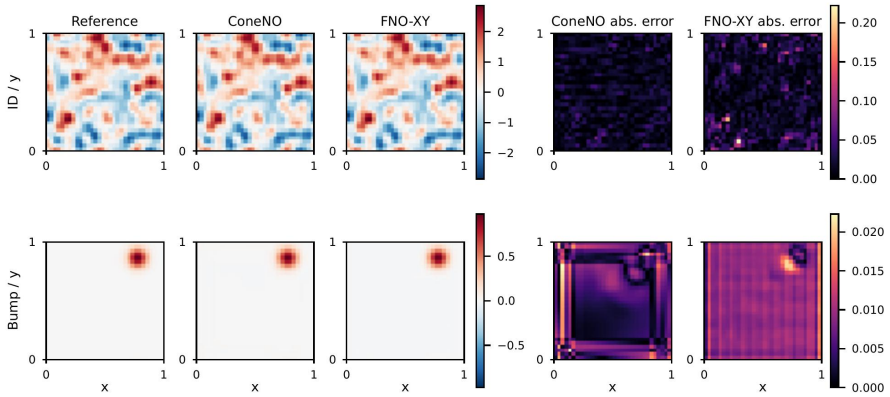


Figure 4: First ID and bump samples of grid-aligned 2D advection at the final stored time.

6.8 2D WAVE

Table 5 applies the same comparison to the wave map. Among the restricted models CONENO-2D is lowest on every condition, and under refinement and rollout it is lowest of all four operators (58% and 40% below FNO-XY on SR2 and $T = 2$). FNO-XY is lowest on ID and bump, with CVS 0.318.

Table 5: 2D wave, relative L^2 $\times\mathbf{100}$ on $(u, \partial_t u)$ (mean $\pm$ SD over five seeds; 128 inputs per condition; matched width-16 backbone).

Model	Params (k)	ID	Bump	SR2	$T = 2$	CVS
FNO-XY	140.3	**0.42±0.03**	**3.87±0.41**	12.64±1.87	6.95±0.52	0.318
Local-window FNO	139.8	1.34±0.09	4.62±0.55	9.83±1.24	7.41±0.83	0
Conv. control	139.5	2.91±0.24	8.15±1.02	16.72±2.13	18.34±2.76	0
CONENO-2D	140.1	0.96±0.07	4.05±0.38	**5.27±0.61**	**4.18±0.44**	0

7 CONCLUSION

A prescribed domain of dependence can be enforced throughout a neural operator, including in the inputs that set its mixing weights, without giving up query-adaptive aggregation. Within that constraint, how admissible information is combined still matters: at matched capacity and supervision, CONENO is more accurate than convolutional and local-window Fourier mixing on both one-dimensional tasks, and among the restricted operators on both two-dimensional tasks. Against unconstrained operators the constraint costs no in-distribution accuracy on transport, improves substantially on Burgers and grid-aligned 2D advection, and removes every measured exterior response. On the 2D wave map CONENO-2D is lowest under refinement and rollout.

Limitations. The constraint needs a known conservative speed bound on a periodic uniform grid; nonperiodic domains and unknown speeds are left open.

AI USE STATEMENT

Generative AI tools were used only to aid and polish the writing. They were not used to develop the conceptual framework, derive mathematical claims, design experiments, implement models, generate data, or interpret results. The authors remain responsible for the text and all scientific claims.

REPRODUCIBILITY STATEMENT

Section 5 states the assumptions and claims behind the guarantees. Appendix C proves Propositions 5.2–5.5 and records the remaining derivations. Appendices B.1 and D specify the architectures, data construction, training settings, evaluation protocols, and numerical checks. Appendix A.1 explains how input reconstruction affects the interpretation of nodal influence. Code, configuration files and the generated result tables will be released upon publication.

REFERENCES

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural PDE solvers. *arXiv preprint arXiv:2202.03376*, 2022.

Jiangce Chen, Wenzhuo Xu, Zeda Xu, Noelia Grande Gutierrez, Sneha Prabha Narra, and Christopher McComb. Enforcing the principle of locality for physical simulations with neural operators. *Journal of Computational Physics*, 538:114131, 2025.

Richard Courant and David Hilbert. *Methods of mathematical physics*, volume 2. Interscience, 1962.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *Constructive Approximation*, pp. 1–67, 2026.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *International Conference on Machine Learning*, pp. 12556–12569. PMLR, 2023.

Fritz John. *Partial differential equations*. Springer Science & Business Media, 2012.

Dmitrii Kochkov, Jamie A Smith, Ayya Alieva, Qing Wang, Michael P Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

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.

Samuel Lanthaler, Siddhartha Mishra, and George E Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *Transactions of Mathematics and its Applications*, 6(1):tnac001, 2022.

Peter D Lax. *Hyperbolic partial differential equations*, volume 14. American Mathematical Soc., 2006.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *arXiv preprint arXiv:2205.13671*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for PDEs on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

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.

Pingchuan Ma, Haoyu Yang, Zhengqi Gao, Duane S Boning, and Jiaqi Gu. PIC2O-Sim: A physics-inspired causality-aware dynamic convolutional neural operator for ultra-fast photonic device time-domain simulation. *APL Photonics*, 10(3), 2025.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023.

Bogdan Raonic, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel De Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. *Advances in Neural Information Processing Systems*, 36:77187–77200, 2023.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized Fourier neural operators. *arXiv preprint arXiv:2111.13802*, 2021.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421: 116813, 2024.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. *arXiv preprint arXiv:2402.02366*, 2024.

A Additional Preliminaries

An evolutionary PDE on $\Omega \times (0, T]$ may be written as

$$\partial_t u = \mathcal{F}(u, \nabla u, \nabla^2 u, a), \qquad u(\cdot, 0) = u_0, \tag{10}$$

where a may contain u_0, coefficient fields, or forcing. Pointwise statements use continuous fields, or discrete samples with specified reconstruction, so that the query values are well defined. Error norms are evaluated in L^2.

For a differentiable scalar query $f_z(a) = \mathcal{G}(a)(z)$, Definition 3.1 implies

$$Df_z(a)[h] = 0 \qquad \text{if } h|_{\mathcal{D}(z)} = 0. \tag{11}$$

The reverse implication requires this condition throughout each admissible constant-interior fiber, with paths along which the fundamental theorem of calculus applies. A derivative that vanishes at only a finite set of inputs is insufficient. Proposition 5.4 gives the discrete version used to interpret the audits. Diffusive equations have no finite cone and lie outside this setting.

For a trained discrete operator and a scalar query i, the empirical influence in (3) is

$$I_{\theta,ij}(a) = \lim_{\varepsilon \to 0} \frac{\mathcal{G}_{\theta,i}(a + \varepsilon e_j) - \mathcal{G}_{\theta,i}(a)}{\varepsilon}, \tag{12}$$

where e_j changes one nodal input component. Reverse-mode differentiation computes these derivatives when its backward rules faithfully differentiate every input-dependent forward operation. Stop-gradient or surrogate backward rules require separate checks (Appendix E). Channelwise squared gradients are summed when computing CVS. Finite nodal perturbations provide an independent check.

A.1 Reference reconstruction and nodal support

Let R_N reconstruct a continuous input from N nodal values and S_N sample the solution. A spectral reference implements $S_N \mathcal{G}^\dagger R_N$, not $\mathcal{G}^\dagger$ on independently localized physical data. Trigonometric reconstruction is global: changing a grid value outside a query cone can change $R_N a$ inside it. Thus a nonzero nodal CVS need not imply a violation by the continuous PDE.

For periodic unit-speed advection on an even grid, the derivative of the spectral reference at (x_q, t) is the cardinal interpolation row

$$J_{qj}(t) = \frac{1}{N}\left[1 + 2\sum_{k=1}^{N/2-1} \cos(2\pi k z_{qj}) + \cos(\pi N z_{qj})\right], \quad z_{qj} = x_q - t - x_j. \tag{13}$$

When Nt is an integer, this reduces to a cyclic shift and its support is the exact characteristic footpoint. At fractional shifts the row generally has nonzero entries outside the nodal cone. Nodal CVS therefore measures a discrete information-flow constraint. Physical intervention claims require a reconstruction-compatible test, such as grid-aligned advection or a locally reconstructed reference.

B Design Details and Controls

B.1 Processor implementation

The implementation uses real-valued compact taps and no spatial interpolation of hidden states. Let $z_\ell(i, m) \in \mathbb{R}^w$ be a hidden vector. A dilation-d_ℓ block first computes

$$\begin{aligned} b_\ell(i,m) &= W_\ell z_\ell(i,m) + b_\ell^{\mathrm{pt}} \\ &\quad + \mathbf{1}_{m \ge d_\ell}\left[\sum_{j=-r}^{r} K_{\ell,j} z_\ell(i - j d_\ell, m - d_\ell) + b_\ell^{\mathrm{sp}}\right], \end{aligned} \tag{14}$$

with periodic spatial indices and zero padding before the initial time, followed by

$$\widetilde{z}_\ell = z_\ell + \mathrm{GELU}\big(\mathrm{LN}_{\mathrm{channel}}(b_\ell)\big). \tag{15}$$

Equation (5) bounds dependency speed. All 1D models use eleven delayed blocks with dilations $1, 1, 2, 2, 4, 4, 8, 8, 16, 16, 32$. The one-dimensional models of Section 6 use width 40 and radius 6 (Appendix D.1.2); the supplementary trajectory models use radius 4 with width 60 for linear advection and 40 for Hopf–Lax Burgers (Table 8). Both two-dimensional tasks use the width-16 backbone below.

Attention follows blocks 6 and 11. All spatial coordinates and output times act as queries. Keys and values are sampled every $s_t = 4$ time steps within each spatial residue modulo $s_x = 4$:

$$\mathcal{J}(i,m) = \left\{(j,n): \begin{array}{l} j \equiv i \pmod{s_x}, \quad n \in s_t\mathbb{N}_0, \quad n \le m, \\ d_{\mathbb{T}}(x_i, x_j) \le c_{\mathrm{mask}}(t_m - t_n) \end{array}\right\}. \tag{16}$$

At each token, query/key/value projections act on the channel-normalized vector. Four attention heads form the residual

$$z_{\ell+1}(p) = \widetilde{z}_\ell(p) + \gamma_\ell O_\ell \,\mathrm{concat}_h\, A_p^h. \tag{17}$$

The learned scalar gate γ_ℓ starts at zero. Other blocks set $z_{\ell+1} = \widetilde{z}_\ell$. Only admissible keys enter the softmax denominator.

The pointwise lift uses $(a(x_i), x_i, t_m)$. A pointwise two-layer head predicts the solution. On a grid with fN points, integer f, spatial tap offsets become fjd_ℓ and residue strides become fs_x, preserving physical tap and key positions at shared nodes.

In 2D, taps use a delayed 3×3 stencil and attention uses a toroidal disk mask. Both two-dimensional tasks use width 16, dilations $(1, 1, 2, 2)$, attention after blocks 2 and 4, and spatial/temporal strides 2.

B.2 Raw-value cone attention

For a scalar state a_0 and query $p = (x_i, t)$, the readout in (7) uses the final admissible feature $h_p \in \mathbb{R}^{40}$. A channel-only layer normalization and a linear map produce twelve scalars, grouped as $(\eta_p^h, \xi_p^h, \rho_p^h)$ for four heads. Define

$$\mu_p^h = c_{\mathrm{mask}} t \tanh \eta_p^h, \qquad s_p^h = \mathrm{clip}(e^{\xi_p^h}, 1/(2N_0), 0.1), \qquad \pi_p^h = \frac{e^{\rho_p^h}}{\sum_{k=1}^4 e^{\rho_p^k}}, \tag{18}$$

where $N_0 = 256$. On the unit torus, the raw-value attention weights are

$$w_{pj}^h = \frac{\exp\left[-\frac{1}{2}\left(d_{\mathbb{T}}^{\pm}(x_j - x_i - \mu_p^h)/s_p^h\right)^2\right]}{\sum_{k \in \mathcal{I}(p)} \exp\left[-\frac{1}{2}\left(d_{\mathbb{T}}^{\pm}(x_k - x_i - \mu_p^h)/s_p^h\right)^2\right]}, \qquad j \in \mathcal{I}(p). \tag{19}$$

The center, width and mixture weight depend on the query feature. Coefficient information enters through the admissible latent stream.

For a grid with fN_0 nodes, $\mathcal{I}(p)$ contains indices $(i + fk) \bmod fN_0$ with integer offsets $-R \le k \le R$, where $R = \min(\lfloor c_{\mathrm{mask}} t N_0 \rfloor, N_0/2)$. When the window covers the full circle, one duplicate endpoint is omitted. Thus refinement preserves physical read locations and every raw input read lies inside the cone. Values are the unnormalized scalar state samples, without a learned value projection. All quantities determining their weights are functions of h_p, whose input dependence is already admissible. This verifies the hypotheses of Proposition 5.2, including the weight-generation paths.

The pointwise decoder fuses the latent prediction b_p and raw attention output as in (7). Its scalar γ starts at zero. Readout query weights start at zero; biases initialize head centers evenly from $-0.8 c_{\mathrm{mask}} t$ to $0.8 c_{\mathrm{mask}} t$, widths at 0.03 and head mixtures uniformly. The readout adds 573 parameters and is trained jointly with the rest of the network.

B.3 Ray bias

The linear advection model uses, for admissible key q and query p,

$$B_{pq}^h = -\frac{\left[d_{\mathbb{T}}^{\pm}(x_q - x_p - v_h(t_p - t_q))\right]^2}{2s_h^2}, \qquad v_h = c_{\mathrm{mask}} \tanh \alpha_h, \tag{20}$$

with $s_h = \mathrm{clip}(\exp \beta_h, 10^{-3}, 0.5)$. Sixteen input-independent scalars set the head velocities and widths; they cannot enlarge the hard mask. The matched transport and Burgers models, and both two-dimensional models, set $B_{pq}^h = 0$.

B.4 Convolutional control

At each attention position, the convolutional control replaces the attention update by a pointwise $w \to 2w \to w$ MLP with GELU, matching the parameter count of the query/key/value and output projections. Channel normalization, residual gates and the compact path are retained. This is the convolutional control of Section 6 and the pointwise capacity control of Table 9; the compact path alone, with attention positions removed, is the convolutional instance of the constraint described in Section 4.

C PROOFS

C.1 PROOF OF PROPOSITION 5.2

Fix a query z and two inputs agreeing on $\mathcal{D}(z)$. For each $z' \in \mathcal{D}(z)$, the lift reads only inputs in $\mathcal{D}(z') \subseteq \mathcal{D}(z)$, so its values agree. Suppose this holds for a hidden layer. The update at z' reads coordinates in $\mathcal{D}(z')$; nestedness and the induction hypothesis make all these hidden values equal. The weights, gates and normalizers computed from them are equal as well. Each update and its pointwise residual therefore agree. Induction through the layers and the pointwise decoder proves invariance at z.

For a compact tap, (5) bounds spatial displacement by c times elapsed time. Along a path, distances obey the triangle inequality and time lags add, keeping its initial input within the query cone. Pointwise queries, keys and values with softmax over admissible keys satisfy the same read restriction. In (7), the retained field values lie inside the initial cone and the weights depend on an admissible query feature, so this readout obeys the restriction as well.

The path argument also holds for periodic geodesic distance. Channel-only normalization and input-independent ray parameters add no spatial reads.

C.2 PROOF OF PROPOSITION 5.3

The solution is $u_0(x - ct)$, which equals $u_0(x - ct + 1/2)$ on half-periodic inputs, so the training error is zero. Translation by $1/2$ multiplies Fourier mode k by $(-1)^k$, hence $\mathcal{G}_{\mathrm{spur}}$ is a multiplier. Its derivative at (x, t) is point evaluation at $x - ct + 1/2$, at toroidal distance $1/2 - ct$ from x; for $t < 1/(4c)$ this exceeds the cone radius ct, so the read is exterior and the discrete CVS equals one. A unit-amplitude narrow bump near $x - ct$ with no copy near $x - ct + 1/2$ has true value one and predicted value zero.

C.3 PROOF OF PROPOSITION 5.4

Set $g(s) = f_i(a + s\delta)$. Since $\delta|_{D_i} = 0$, the fundamental theorem of calculus gives

$$f_i(a + \delta) - f_i(a) = \int_0^1 \langle \nabla_{D_i^c} f_i(a + s\delta), \delta_{D_i^c} \rangle \, \mathrm{d}s.$$

Cauchy–Schwarz yields (8). If the exterior gradient vanishes along the segment, the integral is zero. Vanishing at one endpoint is insufficient: $f(b) = b^2$ has $f'(0) = 0$ but $f(1) \neq f(0)$.

Semigroup composition. For an autonomous PDE, nested cone support is closed under state-to-state rollout $S_{n\tau} = S_\tau^n$, so the constraint holds at every rollout step.

C.4 EXTERIOR RESPONSE AND PAIRED PREDICTION ERROR

For an input pair with common reference y, write the predictions as f_0, f_1, their midpoint as $\bar{f} = (f_0 + f_1)/2$, and the response as $r = f_1 - f_0$. Then

$$\mathcal{R}_{\mathrm{pair}} := \tfrac{1}{2}\mathbb{E}\big[(f_0 - y)^2 + (f_1 - y)^2\big] = \mathbb{E}\big[(\bar{f} - y)^2\big] + \tfrac{1}{4}\mathbb{E}[r^2]. \tag{21}$$

The identity splits paired MSE into shared approximation error and a disagreement term $\frac{1}{4}\mathbb{E}[r^2]$. On the Hopf–Lax interventions of Table 11, the disagreement term accounts for 10.6–42.5% of paired MSE for the Fourier models. On the matched interventions of Table 3, it accounts for 48% for global FNO on transport and for 1–8% in the three other responding cases, whose exterior response is small relative to their approximation error. Models with zero response have paired error equal to their shared approximation error.

C.5 Proof of Proposition 5.5

Identify coarse node i with fine node fi. Their physical coordinates and lifted data agree after applying R_f. A fine compact edge from fi reads $fi - fjd_\ell = f(i - jd_\ell)$, which is the corresponding coarse edge, with the same temporal lag and weight. Pointwise residuals, activations, channel normalization and biases therefore commute with restriction.

The attention residue class of fi modulo fs_x contains exactly the nodes corresponding to the coarse residue class of i modulo s_x. The physical key coordinates, sampled times and hard-mask decisions match. By induction the hidden values at those nodes agree, so their query/key/value projections, row-normalized attention weights, output projections and gates agree. Raw-value read positions also coincide under restriction: their samples, query features and physical offsets are equal, hence so are (18)–(19) and the pointwise fusion. Induction through all blocks and the decoder proves (9). In two dimensions the latent-attention argument applies to both coordinates and the disk distances.

If two fine inputs agree at the old nodes, their restrictions are equal, so (9) forces the shared-node predictions to agree. Numerical tests vary the inserted values independently. Six float64 tests cover scalar and two-channel 1D models and the two-dimensional backbone at factors 2 and 3, with nonzero attention gates. Shared-node discrepancies are below 2×10^{-16}; exterior-sublattice gradients at old queries are zero, while predictions at added nodes change.

D Experimental Protocol and Reproducibility

All experiments share one protocol. Each operator maps the initial data to the solution over its training horizon, and longer horizons are reached by feeding predictions back without reference resets. On the one-dimensional tasks and 2D wave the training horizon is one step of length $\tau = 1/9$; on grid-aligned 2D advection it is the stored seven-frame trajectory. Models within a comparison have matched parameter counts. Training uses AdamW with weight decay 10^{-4}, batch size eight, cosine decay and gradient clipping at norm one, with 80 epochs at initial learning rate 0.001 (40 epochs at 0.003 for 2D advection). The loss is relative L^2 at the target. Five independent seeds control initialization and minibatch order; one-dimensional tasks and 2D wave use 512 / 128 / 128 training / validation / test inputs, and 2D advection is evaluated on 64 test fields. Errors exclude the initial frame, are first averaged within a seed, and are reported as mean and sample SD across seeds. Every table reports mean per-trajectory relative L^2

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\|\widehat{u}_j - u_j\|_2}{\|u_j\|_2 + 10^{-6}}$$

over n test trajectories, scaled by 100. Conditions are ID, Bump, Coeff. (transport), SR2 and $T = 2$; SR2 doubles the spatial grid without further training; $T = 2$ rolls out to time 2, eighteen self-fed steps on the one-step tasks and twice the stored trajectory on 2D advection. Jacobian audits use probes with nonzero total influence and report seed-level aggregates. Finite interventions replace exterior values toward a donor field while keeping every value inside the query cone unchanged. Rollout uses the prefix execution of Appendix D.1.5.

D.1 One-dimensional tasks

D.1.1 Variable-coefficient reference

The transport task is the periodic nonconservative equation $q_t + c(x)q_x = 0$, with joint input (q, c) and piecewise-linear reconstruction of both channels, $0.75 \le c \le 1.25$. The travel coordinate $F(x) = \int_0^x c(s)^{-1}\,\mathrm{d}s$ extends by $F(x+1) = F(x) + F(1)$, and the reference is

$$q(x,t) = q_0\big(F^{-1}(F(x) - t)\big)\,.$$

Segment integrals and inverses use analytic logarithm and exponential formulas. Coefficient-shift tests replace the three training speed-field modes by modes 4–6 at the same amplitude bounds. The conservative mask speed is $c_{\mathrm{mask}} = 1.5$.

D.1.2 OPERATORS ON TRANSPORT

Table 1 compares five operators on the transport reference of Appendix D.1.1, with state pairs separated by $\tau = 1/9$. Rollout feeds back predicted q with c held fixed.

Architectures. CONENO has width 40, compact radius six, eleven delayed blocks, latent attention after blocks 6 and 11, and the raw-value readout of Appendix B.2. The convolutional control keeps the compact path and replaces each latent-attention update by a pointwise MLP. The two models have 264904 and 264331 parameters. The global and local-window FNOs share a width-44 Fourier core (258893 parameters). The local model applies that core to a radius-$\lfloor 1.5\tau N \rfloor$ window and retains the center prediction. DeepONet has 264037 parameters, with a 256-sensor branch and a trunk on (x, τ).

Evaluation. Each condition uses 128 inputs. SR2 refines the ID fields to 512 nodes; $T = 2$ uses eighteen self-fed calls.

Information dependence. Jacobians use both input channels at six space–time queries. Finite interventions replace exterior values of both fields toward a cyclic donor, with fraction 0.5. The reference at each query is unchanged. Results are in Table 3.

D.1.3 OPERATORS ON THE BURGERS FLOW MAP

Table 2 uses the five architectures of Appendix D.1.2 on periodic $u_t + (u^2/2)_x = 0$. Inputs are cell averages. Each supervised transition spans $\tau = 1/9$.

The Godunov update is

$$q_j^{m+1} = q_j^m - \frac{\Delta t}{\Delta x}\big[F(q_j^m, q_{j+1}^m) - F(q_{j-1}^m, q_j^m)\big], \qquad F(a,b) = \begin{cases} \min_{v\in[a,b]} v^2/2, & a \le b, \\ \max_{v\in[b,a]} v^2/2, & a > b. \end{cases} \tag{22}$$

Each macro step uses a fixed number of equal substeps so that the cone of speed 1.5 contains the numerical stencil. Training pairs belong to this discrete map. Initial fields follow Appendix D.3.2. Evaluation uses 128 inputs per condition, nine calls to $T = 1$ and eighteen to $T = 2$. Exterior interventions replace the state outside a radius-$1/6$ cone by an equal mixture with a donor field; the numerical reference is unchanged (Table 3).

Table 6: Burgers flow-map relative L^1 ($\times 100$), same checkpoints as Table 2.

Model	Params (k)	ID	Bump	SR2	$T = 2$
Global FNO	258.9	2.14 ± 0.03	5.52 ± 1.43	2.67 ± 0.04	2.60 ± 0.03
Local-window FNO	258.9	1.03 ± 0.03	1.41 ± 0.24	2.31 ± 0.56	1.26 ± 0.06
DeepONet	264.0	4.43 ± 0.09	6.79 ± 0.15	4.50 ± 0.09	4.88 ± 0.14
Conv. control	264.3	1.46 ± 0.22	3.24 ± 1.50	2.06 ± 0.18	1.89 ± 0.33
CONENO	264.9	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.98 \pm 0.21}$	$\mathbf{1.12 \pm 0.09}$	$\mathbf{0.84 \pm 0.04}$

Table 7: Burgers flow-map final-frame relative L^2 ($\times 100$), same checkpoints as Table 2.

Model	Params (k)	ID	Bump	SR2	$T = 2$
Global FNO	258.9	7.89 ± 0.30	10.07 ± 1.11	9.47 ± 0.28	13.21 ± 0.26
Local-window FNO	258.9	4.35 ± 0.32	2.53 ± 1.00	9.62 ± 2.37	7.60 ± 0.42
DeepONet	264.0	5.69 ± 0.14	6.12 ± 0.14	5.74 ± 0.14	6.96 ± 0.24
Conv. control	264.3	6.66 ± 1.05	7.83 ± 2.25	8.35 ± 0.99	11.99 ± 2.08
CONENO	264.9	$\mathbf{3.32 \pm 0.27}$	$\mathbf{2.08 \pm 0.11}$	$\mathbf{4.46 \pm 0.31}$	$\mathbf{5.74 \pm 0.71}$

D.1.4 ROLLOUT ERROR OVER THE FULL HORIZON

Figure 5 uses the same 128 ID inputs and checkpoints as Tables 1 and 2. Each model advances its own state for eighteen steps of length $\tau = 1/9$.

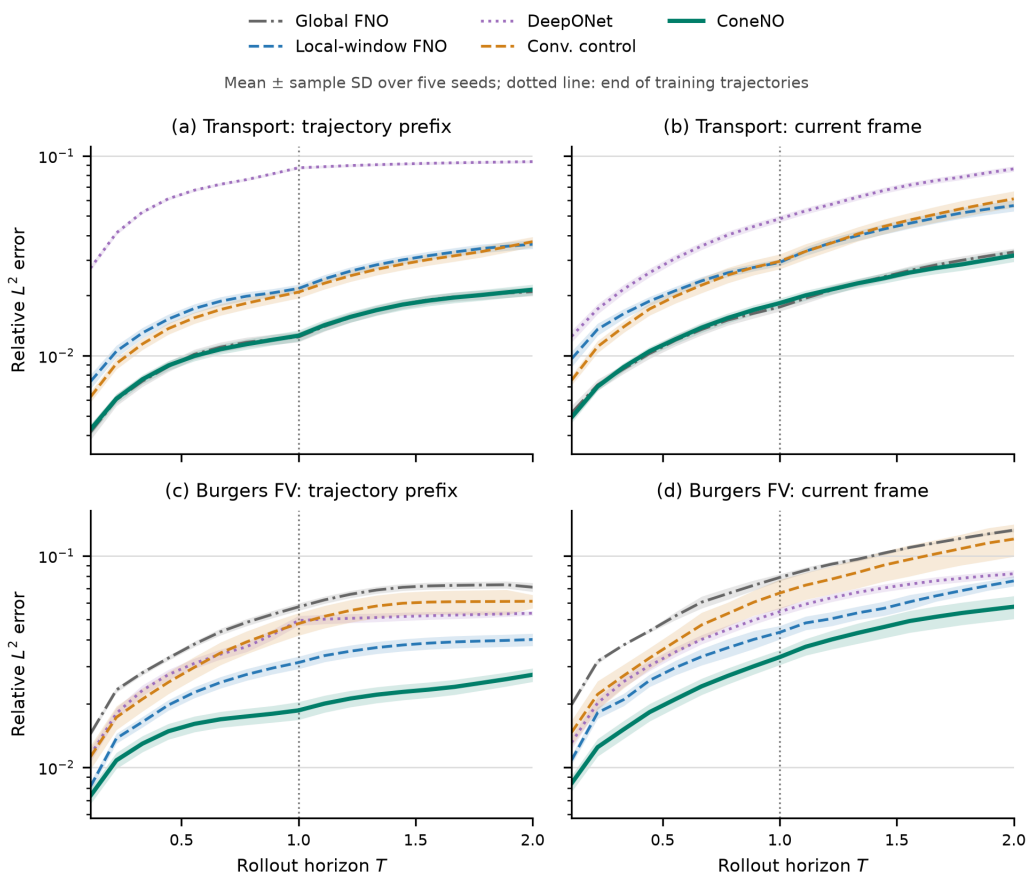


Figure 5: Self-fed rollout error at $T = k/9$ on the two matched tasks. Bands are SD over five seeds; the dotted vertical line marks the training horizon.

For horizon $T_k = k\tau$, the left panels aggregate the positive-time trajectory, while the right panels evaluate only its current frame:

$$E_{\text{prefix}}(k) = \frac{1}{128}\sum_{i=1}^{128}\frac{\|\widehat{u}_i^{1:k} - u_i^{1:k}\|_2}{\rho(u_i^{1:k})}, \tag{23}$$

$$E_{\text{frame}}(k) = \frac{1}{128}\sum_{i=1}^{128}\frac{\|\widehat{u}_i^{k} - u_i^{k}\|_2}{\rho(u_i^{k})}. \tag{24}$$

The denominators follow the task metrics: $\rho(v) = \|v\|_2 + 10^{-6}$ for transport and $\rho(v) = \max\{\|v\|_2, 10^{-12}\}$ for Burgers. Each curve first averages input-wise ratios within a seed, then reports the mean and sample SD across the five seeds. The initial frame is excluded. At $T = 2$, the prefix errors reproduce the corresponding main-table entries.

ConeNO has lower prefix and current-frame errors than the convolutional and local-window Fourier controls at every evaluated horizon. On transport it lies within seed variation of global FNO; on Burgers it is below all four baselines throughout the rollout.

D.1.5 Prefix execution

A $\tau = 1/9$ call only needs frames $0, \ldots, 7$. On this prefix, delayed taps and cone attention use the same keys, values and softmax sets as the corresponding nodes of a 64-frame graph, so the retained outputs agree. Independent spatial residue groups may be packed along the batch axis. The compact-path models in Tables 1 and 2 train and roll out with this prefix.

D.2 TWO-DIMENSIONAL TASKS

D.2.1 2D WAVE

Table 5 uses the wave equation on the periodic unit square,

$$\partial_{tt} u = c^2 \Delta u, \tag{25}$$

with $c = 1$. The operator maps Cauchy data $(u, v) = (u, \partial_t u)$ through one step of length $\tau = 1/9$. On the unit torus the Fourier modes $\mathbf{n} \in \mathbb{Z}^2$ rotate at $\omega_{\mathbf{n}} = 2\pi c|\mathbf{n}|$. For $\mathbf{n} \neq \mathbf{0}$,

$$\widehat{u}_{\mathbf{n}}(\tau) = \widehat{u}_{\mathbf{n}}(0)\cos(\omega_{\mathbf{n}}\tau) + \omega_{\mathbf{n}}^{-1}\widehat{v}_{\mathbf{n}}(0)\sin(\omega_{\mathbf{n}}\tau), \tag{26}$$

$$\widehat{v}_{\mathbf{n}}(\tau) = -\omega_{\mathbf{n}}\widehat{u}_{\mathbf{n}}(0)\sin(\omega_{\mathbf{n}}\tau) + \widehat{v}_{\mathbf{n}}(0)\cos(\omega_{\mathbf{n}}\tau), \tag{27}$$

and $\widehat{u}_{\mathbf{0}}(\tau) = \widehat{u}_{\mathbf{0}}(0) + \tau\widehat{v}_{\mathbf{0}}(0)$, $\widehat{v}_{\mathbf{0}}(\tau) = \widehat{v}_{\mathbf{0}}(0)$. The exact domain of dependence is the disk of radius ct. The conservative mask speed is $c_{\mathrm{mask}} = 1.5$.

Training, validation and test use 512 / 128 / 128 fields on a 64×64 grid. ID draws bandlimited (u_0, v_0) with the same eight-mode family as one-dimensional linear advection, independently in each coordinate, then L^2-normalizes the pair. Bump uses a localized isotropic Gaussian $u_0(\mathbf{x}) = \exp[-d_{\mathbb{T}^2}(\mathbf{x}, \boldsymbol{\mu})^2/(2\sigma^2)]$ with $\boldsymbol{\mu} \sim U[0,1)^2$, $\sigma = 0.05$ and $v_0 = 0$. SR2 evaluates the same checkpoints on 128×128 without retraining. $T = 2$ uses eighteen self-fed steps. Errors are mean per-trajectory relative L^2 $\times 100$ on (u, v), reported as mean and sample SD across five seeds. CVS and exterior interventions use the disk cone of radius $c_{\mathrm{mask}}\tau$ at a query; the Fourier reference is unchanged under an exterior edit.

Architectures. All four models use the two-dimensional backbone of Appendix B.1: delayed 3×3 taps, dilations $(1, 1, 2, 2)$, and width 16. CONENO-2D places disk-masked attention, with raw-value readout, after blocks 2 and 4, spatial/temporal strides 2, and $B^h_{pq} = 0$. The convolutional control keeps that compact path and replaces each attention update by the pointwise MLP of Appendix B.4. Global FNO-XY uses a two-dimensional Fourier core whose width is set so that the parameter count matches CONENO-2D. Local-window FNO applies the same core to the disk of radius $\lfloor c_{\mathrm{mask}}\tau N \rfloor$ and retains the center prediction. The matched counts are those of Table 5.

D.2.2 GRID-ALIGNED 2D ADVECTION

Table 4 uses

$$\partial_t u + \mathbf{v} \cdot \nabla u = 0, \qquad \mathbf{v} = (1, 1)/\sqrt{2},$$

on the periodic unit square with a $32 \times 32 \times 7$ grid. Each stored step is a one-cell translation, so the nodal cone of Appendix A.1 is exact. The conservative mask speed is $c_{\mathrm{mask}} = 1.5$.

The training horizon is the stored seven-frame trajectory, which the model predicts from the initial data. Table 8 lists the optimizer and the CONENO-2D backbone. Evaluation uses 64 test fields. ID draws the bandlimited family of the one-dimensional linear trajectories, independently in each coordinate. Bump uses a localized isotropic Gaussian with $\sigma = 0.05$. SR2 evaluates the same checkpoints on 64×64 without retraining. $T = 2$ evaluates a stored trajectory of twice the training horizon. Errors are mean per-trajectory relative L^2 $\times 100$. CVS uses the disk cone at each query.

Architectures. The four operators are those of Section 6.6. CONENO-2D has 140.1k parameters and uses disk-masked latent and raw-value attention on the width-16 backbone of Appendix B.1. The convolutional control (139.5k) keeps that compact path and replaces each attention update by the pointwise MLP of Appendix B.4. Local-window FNO (138.6k) applies a two-dimensional Fourier core to the query disk and retains the center prediction. Global FNO-XY uses a matched two-dimensional Fourier core (139.4k). Errors in Table 4 are mean and sample SD.

D.3 SUPPLEMENTARY TRAJECTORY STUDIES

The nodal influence map, the attention controls and the Hopf–Lax Burgers audit train one-dimensional models on whole space–time trajectories. CONENO uses latent attention only, with the configurations in Table 8.

Linear initial data. On $x_j = j/N$, define

$$g_K(x) = \sum_{k=1}^{K} \big[A_k \cos(2\pi kx) + B_k \sin(2\pi kx)\big], \qquad A_k, B_k \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Advection uses $u_0 = g_8/(\mathrm{std}_{\mathrm{grid}}(g_8) + 10^{-8})$. Linear trajectories use exact Fourier evolution of the sampled inputs, with the reconstruction qualification in Appendix A.1. The linear bump field is $u_0(x) = \exp[-d_{\mathbb{T}}(x,\mu)^2/(2\sigma^2)]$, with $\mu \sim U[0,1)$ and $\sigma = 0.05$.

Fourier baselines and optimization. The 1D Fourier baselines have four blocks, each applying GELU to the sum of a spectral map and a pointwise linear branch. FNO-X retains 16 nonnegative spatial FFT bins at width 48 or 114. FNO-ST uses width 30, two 16-bin spatial bands and eight nonnegative temporal bins. All 1D models use the optimizer settings above.

Table 8: CONENO configurations for the one-dimensional trajectory studies and grid-aligned 2D advection (Table 4); 2D wave is specified in Appendix D.2.1. Attention positions use one-based block indices.

Setting	1D	2D adv.
Width / compact radius	60 / 4 (linear); 40 / 4 (Hopf–Lax Burgers)	16 / 1
Dilations	$(1,1,2,2,4,4,8,8,16,16,32)$	$(1,1,2,2)$
Attention positions	6, 11	2, 4
Heads / spatial stride / time stride	4 / 4 / 4	4 / 2 / 2
Ray bias	On (linear); off otherwise	Off
Epochs / batch size / initial LR	80 / 8 / 0.001	40 / 8 / 0.003
Optimization	AdamW, cosine decay, gradient clip 1	Same

Nodal influence. Figure 6 shows the audited quantity of Appendix A for one advection query. Both Fourier baselines have exterior sensitivity, whereas CONENO is sensitive only inside the prescribed interval.

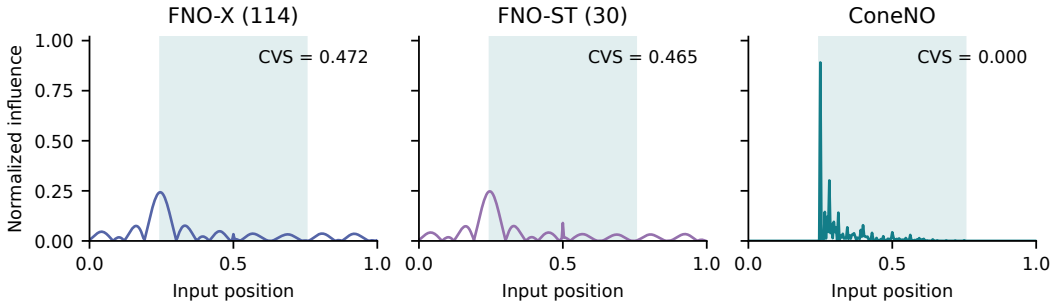


Figure 6: Normalized input influence at $(x,t) = (1/2, 16/63)$ for one advection query. FNO-X and FNO-ST are the spatial and space–time Fourier trajectory baselines. Shading marks the admissible initial interval.

D.3.1 ATTENTION CONTROLS ON LINEAR ADVECTION

Table 9 varies the mixer on linear advection trajectories. The compact path alone has no attention positions; the pointwise capacity control adds the MLP of Appendix B.4 at those positions. At matched parameter count, attention without ray bias lowers bump error relative to the pointwise control, and the 16 ray scalars lower it further, to 33% below the pointwise control. On ID and SR2 the pointwise control and CONENO differ by at most 0.03 points.

Table 9: Linear advection attention controls, relative L^2 $\times 100$ on the same 128 test inputs and training budget. Parameters in thousands.

Model	Params (k)	ID	Bump	SR2	CVS
Compact path only	402.8	0.80	29.08	0.74	0
Pointwise capacity control	432.4	0.68	25.86	0.65	0
Attention, no ray bias	432.4	0.76	22.69	0.73	0
ConeNO	432.4	0.66	17.40	0.64	0

D.3.2 Hopf–Lax Burgers trajectories

The inviscid equation $u_t + (u^2/2)_x = 0$ is posed on the unit torus. Nodal values define centered piecewise-constant cells. A float64 Hopf–Lax minimizer supplies point samples of the entropy solution, with speed bounded by 0.75. The speed-one nodal cone covers those cells at every positive stored time.

All three models use the 1D split and optimizer of Appendix D, with multiple training seeds. ConeNO uses latent attention at width 40. Table 10 reports 128 inputs per condition.

Table 10: Hopf–Lax Burgers trajectories, relative L^2 $\times 100$ on 128 inputs per test (mean over seeds). Parameters in thousands.

Model	Params (k)	ID	Bump	SR2	CVS
FNO-X (48)	307.7	9.90	9.50	9.86	0.00653
FNO-ST (30)	349.1	11.64	10.70	11.56	0.01030
ConeNO (40)	193.9	**7.10**	**8.05**	**7.32**	0

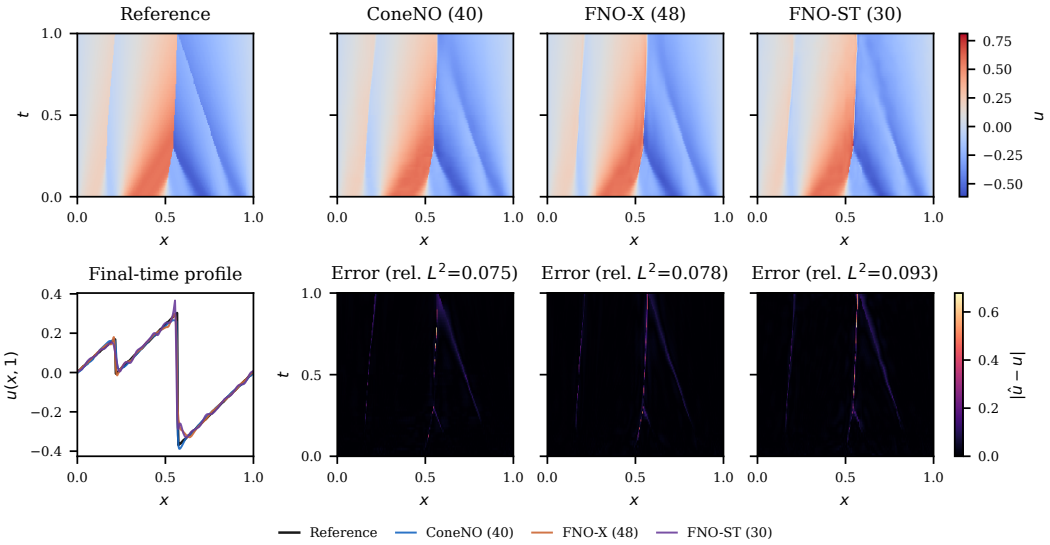


Figure 7: First Hopf–Lax Burgers ID input and final-time errors.

Exterior interventions. Exterior cells are replaced by a convex combination with a donor field, keeping $|u_0| \le 0.75$ and the interior unchanged, at nine queries for each of 128 inputs. The entropy reference is unchanged. At replacement fraction 0.5, ConeNO has zero response and pooled query RMSE 0.01100, whereas FNO-X rises from 0.01410 to 0.02342 and FNO-ST from 0.01270 to 0.02172 (Table 11).

Table 11: Exterior interventions on Hopf–Lax Burgers: 128 fields, nine queries each, aggregated over seeds.

Model	Clean RMSE	α	Response RMS	Perturbed RMSE
FNO-X (48)	0.01410	0.25	0.01002	0.01652
	0.01410	0.5	0.01954	0.02342
	0.01410	1	0.03903	0.04122
FNO-ST (30)	0.01270	0.25	0.01104	0.01504
	0.01270	0.5	0.02044	0.02172
	0.01270	1	0.03321	0.03371
CONENO (40)	0.01100	0.25	0	0.01100
	0.01100	0.5	0	0.01100
	0.01100	1	0	0.01100

E RELATION TO NEARBY ARCHITECTURES

Windowed and propagation-aware designs (Chen et al., 2025; Ma et al., 2025) restrict the region that feeds a prediction or the kernel that propagates it. The local-window FNO of Section 6.1 is DDELD (Chen et al., 2025) in the matched setting: the global Fourier core is applied to each admissible window and the center prediction is kept. Convolutional and graph operators (Raonic et al., 2023; Brandstetter et al., 2022; Pfaff et al., 2020) bound the receptive field by depth; the convolutional control of Appendix B.4 is that class with the cone bound (5) imposed layer by layer. CONENO differs from both in constraining the inputs of the data-dependent weights as well as the aggregated values.

CNO (Raonic et al., 2023) establishes continuous-discrete equivalence on bandlimited spaces. Proposition 5.5 is the corresponding identity for nested integer refinements with fixed physical read locations.

Whole-map auditing. If $f_i(a) = s(a)\phi(a_{D_i}/s(a))$ with a global scale $s > 0$, exterior derivatives can appear through s even when ϕ is local. Stop-gradients can hide forward dependence. Finite interventions check the forward response directly.