Learning Event-to-Field Operators with Query-Time Aggregation

Xingyu Sha
Jinan University

Abstract

A point source in a PDE does not fill a grid. A weather station reports one value at one site, not the temperature between stations. A pickup event in a city occurs at a coordinate and a time; it is not born as an image. Yet many neural-operator pipelines first turn such sparse observations into dense surrogate fields before learning [Li et al., 2021, Lu et al., 2021, Li et al., 2020, 2023]. This step is often called preprocessing. In sparse regimes, it is already a modeling choice: it changes the object on which the operator learns.

We introduce the Dirac Neural Operator (DIRAC-NO), a measure-native neural operator for sparse event-to-field learning with learned, query-conditioned field construction rather than fixed pre-interpolation. DIRAC-NO keeps observations as atomic event measures until a query is made, then constructs field values through adaptive Dirac-kernel aggregation. Because this aggregation is a direct sum, the front end preserves source-wise additivity before nonlinear refinement and admits a learned Green-kernel view of source-to-field response.

Across controlled point-source PDE benchmarks and real sparse reconstruction tasks, DIRAC-NO characterizes when measure-native processing is beneficial relative to interpolation-first learning. On MeasureBench-1D, it achieves relative L^2 error 0.015 at $n = 128$, compared with 0.49 for FNO, and better preserves superposition structure in trained models. In 2D, it is strongest under parameter and density shifts, while ERA5 reconstruction shows that the representation advantage persists beyond analytic PDEs at high observation budgets [Hersbach et al., 2020, Rasp et al., 2024]. Boundary cases and negative-control event streams sharpen the claim rather than broaden it: the benefit appears where sparse, event-native inputs also have source-response structure. Sparse operator learning should therefore be organized not only by the neural backbone, but by the native space in which the input is represented.

1 Introduction

Sparse event-to-field problems begin with observations that do not look like fields. A point source is localized. A weather station reports one value at one site. An urban event occurs at one coordinate and one time. None of these observations supplies values between events. Yet many neural-operator pipelines first fill those gaps through rasterization, smoothing, or interpolation. Before the operator sees the data, the input space has already been changed. This mismatch appears across point-source physical systems, weather reconstruction, urban activity modeling, and event-driven data streams [Hersbach et al., 2020, Dee et al., 2011, Rasp et al., 2020, 2024, Bi et al., 2023, Lam et al., 2023, New York City Taxi and Limousine Commission, 2023, Zhang et al., 2017, Yao et al., 2018, Li et al., 2018, Liang et al., 2018, Pan et al., 2019].

Neural operators provide a powerful framework for learning maps between infinite-dimensional spaces [Kovachki et al., 2023]. Models such as Fourier Neural Operator (FNO) [Li et al., 2021],

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

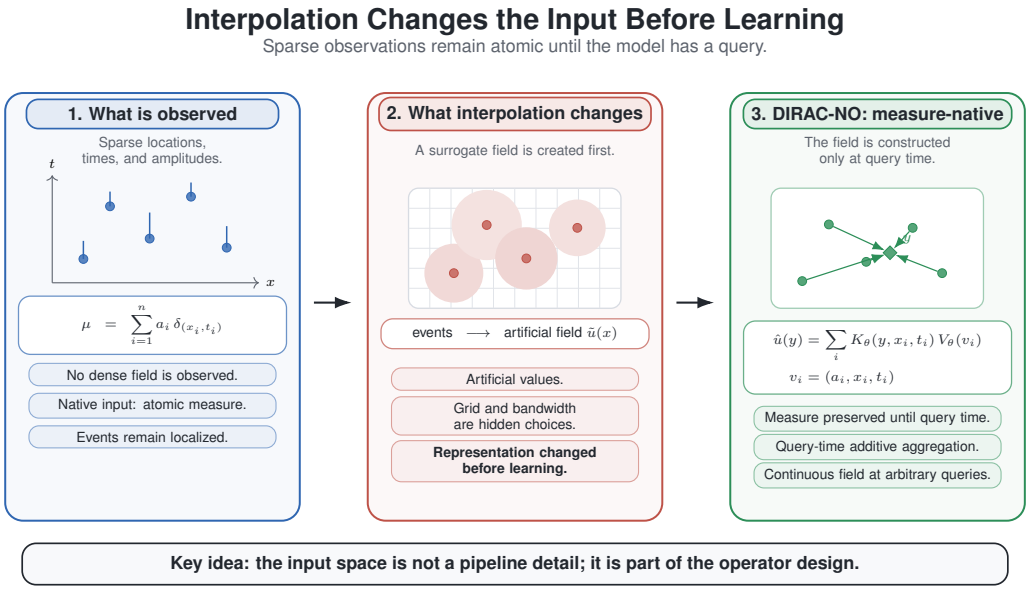


Figure 1: **Interpolation changes the input before learning.** Sparse event observations are natively atomic: they occur at locations, times, and amplitudes. Interpolation-first pipelines first rasterize or smooth these events into an artificial field, introducing values at unobserved locations and making grid resolution and bandwidth part of the modeling choice. DIRAC-NO keeps the input as an event measure until query time and constructs the field by direct measure-to-function aggregation.

DeepONet [Lu et al., 2021], Multipole Graph Neural Operator [Li et al., 2020], and Geo-FNO [Li et al., 2023] have been effective in PDE surrogate modeling and related prediction tasks. Their success, however, usually begins after an input convention has already been satisfied: the observed signal is represented as a function on a grid, a fixed sensor layout, or a geometric domain. Our question is what happens before that convention is met.

Interpolation is not preprocessing; it is the first modeling choice. Once sparse observations are rasterized, smoothed, or binned, the model no longer acts on the original event set. It acts on a constructed field whose values depend on grid resolution, interpolation bandwidth, and local observation geometry. Interpolation creates values at unobserved locations and changes how individual events interact before the learning problem is posed. In sparse event-to-field learning, the bottleneck is therefore not only the operator backbone, but the space in which the input is represented.

We introduce the Dirac Neural Operator (DIRAC-NO), a measure-native neural operator for sparse event-to-field learning that avoids fixed pre-interpolation and instead constructs fields through learned, query-conditioned aggregation. DIRAC-NO is built around one constraint: the field should be constructed only when queried. Its measure-to-function layer aggregates event contributions directly at query locations with an adaptive Dirac kernel. Because this front end is a direct sum, it preserves source-wise additivity at the aggregation stage before nonlinear refinement and admits a learned Green-kernel view of source-to-field response [Stakgold and Holst, 2011, Evans, 2010, Boullé et al., 2022, Gin et al., 2021].

Because the claim is representational, the comparison is defined not only by the backbone, but by what each model is allowed to see before learning. DIRAC-NO receives an atomic event measure and constructs fields only at query time. FNO receives a rasterized or interpolated grid field. Sequence and set baselines receive padded event sequences or context-query sets. The experiments therefore compare input spaces—measure, grid field, sequence, and context set—rather than treating preprocessing as a neutral implementation detail.

We evaluate this representation choice through a four-part evidence chain: sparsity-induced breakdown, additive structure, real sparse reconstruction, and negative controls. On MeasureBench-1D, DIRAC-NO achieves relative L^2 error 0.015 at $n = 128$, compared with 0.49 for FNO. In 2D, it is strongest under parameter and density shifts, where sparse observation geometry makes interpolation

brittle. On ERA5 weather reconstruction, the method also achieves the best accuracy at high observation budgets, showing that the representation advantage can persist beyond analytic point-source PDEs. Boundary cases and negative-control event streams sharpen the claim rather than broaden it: event-native input alone is not enough; the advantage is strongest when sparse geometry and source-response structure coincide.

Our claims. The paper makes three claims. First, sparse event-to-field learning should not automatically begin with interpolation. Second, query-time measure aggregation is a natural front end for source-driven operators, because adding a source should add its response rather than renormalize the whole input. Third, the advantage appears where the representation argument predicts: sparse, event-native, approximately additive systems. Together, these claims suggest that sparse operator learning should be organized not only by the neural backbone, but by the native space in which the input is represented.

2 Related Work

DIRAC-NO sits between neural operators, sparse-input models, and kernel views of source-to-field response. The difference is where the field is formed. Most neural operators begin after sparse observations have already been converted into a function. DIRAC-NO asks what happens before that conversion.

Function-native neural operators. Neural operators learn maps between function spaces and have become central to scientific machine learning [Kovachki et al., 2023]. FNO, DeepONet, MGNO, and Geo-FNO are effective when the input is already a field on a grid, values on a fixed sensor layout, or a signal on a geometric domain [Li et al., 2021, Lu et al., 2021, Li et al., 2020, 2023]. This convention is natural when dense fields are observed or when the mesh is part of the problem. It is less natural when the native observation is a sparse event measure. In that case, interpolation answers the representation question before operator learning begins.

Sparse inputs and source-response kernels. PointNet, Set Transformer, Neural Processes, temporal point-process models, and sequence models show that sparse or irregular observations need not be forced into fixed grids [Qi et al., 2017, Lee et al., 2019, Garnelo et al., 2018, Shchur et al., 2021, Gu and Dao, 2023]. But consuming a sparse set is not the same as constructing a continuous field from an event measure. In source-driven event-to-field problems, the aggregation rule carries structure: adding a source should add its response, not renormalize the whole input. This links DIRAC-NO to kernel and Green's function views of source-to-field response [Stakgold and Holst, 2011, Evans, 2010, Boullé et al., 2022, Gin et al., 2021], while making a different representation choice: the input remains atomic until query time.

3 Method

DIRAC-NO is built around one constraint: the input should remain atomic until the model has a query to answer. It does not first turn sparse observations into a dense surrogate field. It represents them as an event measure and constructs field values only at query locations.

3.1 Keeping the Input Atomic

Let $\Omega_x \subset \mathbb{R}^d$ be a bounded spatial domain and let

$$\Omega := \Omega_x \times [0, T]$$

be the corresponding spatiotemporal domain. Let $\mathcal{M}(\Omega)$ denote the space of finite Radon measures on Ω [Folland, 1999, Rudin, 1991]. One input sample is represented as the atomic event measure

$$\mu = \sum_{i=1}^{n} a_i \, \delta(\mathbf{x} - \mathbf{x}_i) \otimes \delta(t - t_i), \qquad a_i \in \mathbb{R},\ \mathbf{x}_i \in \Omega_x,\ t_i \in [0, T]. \tag{1}$$

Here a_i is the event amplitude, $\mathbf{x}_i$ is the spatial location, and t_i is the event time. The number of events n may vary across samples.

The learning problem is to map this atomic input to a continuous output field:

$$\mathcal{F}_\theta : \mathcal{M}(\Omega) \to L^2(\Omega_x).$$

The key point is what we do not assume. The input is not a function sampled on a grid, a mesh, or a fixed sensor layout. Field construction is postponed until a query location is specified.

3.2 Constructing Fields Only at Query Time

Given a query location $\mathbf{y} \in \Omega_x$, the measure-to-function layer constructs a response directly from the event measure:

$$A_\theta(\mu)(\mathbf{y}) = \sum_{i=1}^{n} K_\theta(\mathbf{y}, \mathbf{x}_i, t_i)\, V_\theta(a_i, \mathbf{x}_i, t_i). \tag{2}$$

Here K_θ is an adaptive query-to-event kernel and V_θ encodes event amplitude, location, and time. Eq. (2) is the central operation of the model: the field is not built before learning. It is built only when queried.

In practice, K_θ is implemented with multiple learned Gaussian-style heads with query-dependent bandwidths, while V_θ is a lightweight event-value encoder. These details control how local or broad each source response can be, but they are not the main distinction. The main distinction is that aggregation is performed over the atomic input rather than over an interpolated field. Implementation details are given in Appendix A.8.

3.3 Why Pure Summation Matters

The aggregation rule is not an implementation detail. In source-driven systems, adding a source should add its response; it should not renormalize every other source. This is why Eq. (2) uses direct summation. At the aggregation stage, source-wise contributions remain additive before any nonlinear refinement is applied.

Proposition 1 (Source-Wise Additivity of Query-Time Aggregation)**.** *For notational simplicity, suppress the time coordinate. Let*

$$E_1 = \{(a_i, x_i)\}_{i=1}^{m}, \qquad E_2 = \{(b_j, \tilde{x}_j)\}_{j=1}^{n}$$

be two disjoint event lists, and let $E_1 \sqcup E_2$ denote their concatenation as an event multiset. The aggregation layer in Eq. (2) *satisfies*

$$A_\theta(E_1 \sqcup E_2)(\mathbf{y}) = A_\theta(E_1)(\mathbf{y}) + A_\theta(E_2)(\mathbf{y}),$$

provided that aggregation is implemented by direct summation, without normalization by the event count or by the full input set.

Proof. See Appendix D for the proof, together with short arguments showing why softmax attention and mean aggregation do not preserve the same additive structure. □

Remark 1. *This is an aggregation-stage property, not a claim that the full refined model is linear end to end. In normalized attention [Vaswani et al., 2017], each contribution is rescaled by a partition function depending on the full input set. In mean aggregation, each term is divided by the number of observed events. Direct summation avoids this renormalization before later refinement layers introduce nonlinear corrections.*

3.4 A Learned Green-Kernel View

The same aggregation rule has a source-response interpretation. Consider a linear source-driven equation

$$\mathcal{L}u = \mu.$$

When a Green's function exists and the output is evaluated at a fixed target time,

$$u(\mathbf{y}) = \int_\Omega G(\mathbf{y}, \mathbf{x}, t)\, d\mu(\mathbf{x}, t) = \sum_{i=1}^{n} a_i\, G(\mathbf{y}, \mathbf{x}_i, t_i). \tag{3}$$

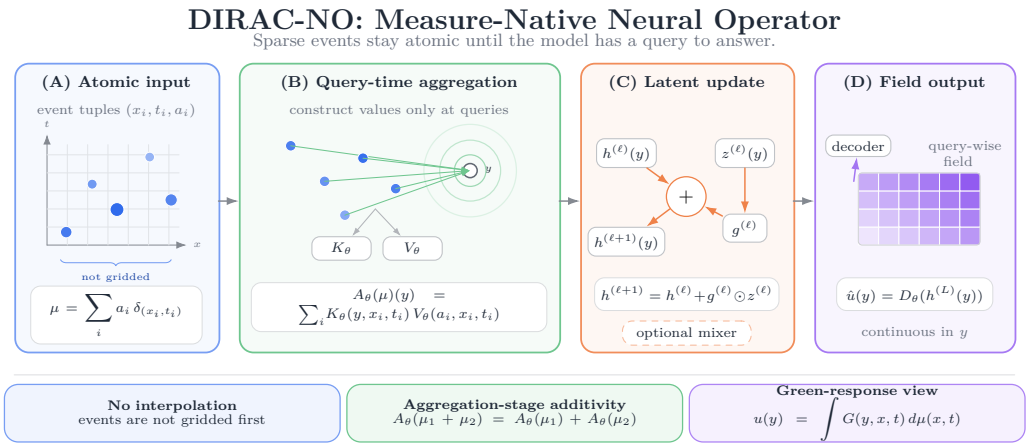


Figure 2: **DIRAC-NO architecture.** Sparse observations remain an event measure until query time. The central layer A_θ constructs field values by summing learned source-wise responses at each query location. This aggregation stage is additive before nonlinear refinement. A gated latent update, optional spectral mixing, and a continuous decoder then produce $\hat{u}(\mathbf{y})$ at arbitrary query locations.

Comparing Eq. (3) with Eq. (2), A_θ can be viewed as a learned source-response layer: K_θ models query-to-source response, while V_θ carries source amplitude, location, and time context.

This view is most natural for diffusion-dominated elliptic and parabolic systems, whose Green's functions are smooth away from the source and decay spatially [John, 1982, Evans, 2010]. We use it as an interpretive guide, not a universal claim. Benchmark-specific realization and Wasserstein stability results are given in Appendix D.

3.5 Refinement and Full Architecture

The full model $\mathcal{F}_\theta$ wraps the aggregation layer with query embeddings, gated refinement, optional spectral mixing, and a continuous decoder. Query coordinates $\mathbf{y} \in \Omega_x$ are first embedded into an initial latent state $h^{(0)}(\mathbf{y})$. At layer l, the query-time measure response is injected through

$$h^{(l+1)}(\mathbf{y}) = h^{(l)}(\mathbf{y}) + g^{(l)}(\mathbf{y}) \odot A_\theta^{(l)}(\mu)(\mathbf{y}), \qquad g^{(l)}(\mathbf{y}) = \sigma(W^{(l)} h^{(l)}(\mathbf{y})). \tag{4}$$

An optional FNO-style block [Li et al., 2021] may then capture longer-range latent-field interactions; this block is not needed for the additivity of the aggregation front end. A shallow decoder maps the final latent representation to $\hat{u}(\mathbf{y})$. Because the model is defined in query coordinates, inference can be performed at arbitrary query resolutions without retraining.

4 Experiments

The experiments test the representation claim of the paper: sparse event-to-field learning should not automatically begin with interpolation. We organize the evidence around four questions. How does performance change as observation coverage decreases? Does query-time aggregation preserve the additive structure that motivates the model? Does the same representation choice remain useful beyond analytic point-source PDEs? And when event-native input is not enough, do the failures occur where the inductive bias predicts?

4.1 Experimental Protocol

Benchmarks. We evaluate DIRAC-NO on controlled source-driven benchmarks, real sparse reconstruction tasks, and negative-control event streams. MeasureBench-1D and MeasureBench-2D test point-source diffusion-reaction operators under the sparsity sweep $n \in \{128, 64, 32, 16, 8, 4\}$. ERA5 and NYC Taxi test real sparse event-to-field reconstruction [Hersbach et al., 2020, Rasp et al., 2020, 2024, New York City Taxi and Limousine Commission, 2023]. Financial streams are used only as negative controls, since they are event-native but not clearly source-driven.

Table 1: MeasureBench-1D sparsity sweep (RelL2↓, means over 3 seeds; std at $n = 128$ shown inline, full mean±std in Appendix B).

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	$\mathbf{0.015}\pm.001$	$\mathbf{0.749}$	$\mathbf{0.897}$	$\mathbf{0.964}$	$\mathbf{0.985}$	$\mathbf{0.994}$
FNO-1D	$0.490\pm.004$	1.122	1.285	1.334	1.350	1.361
Transformer	$0.040\pm.003$	0.806	1.067	1.268	1.412	1.588
DeepONet-1D	$0.051\pm.004$	0.819	1.116	1.342	1.510	1.641
Mamba-1D	$0.095\pm.008$	1.101	1.894	3.026	4.503	6.199
CNP	$0.170\pm.006$	0.959	1.426	1.910	2.482	3.199

Baselines. The baselines are grouped by input representation. FNO [Li et al., 2021] is interpolation-first: it receives a dense grid constructed before learning. Transformer [Vaswani et al., 2017] and Mamba [Gu and Dao, 2023] consume padded event sequences. CNP [Garnelo et al., 2018] uses context-query sets. DeepONet [Lu et al., 2021] receives sparse event tuples and query coordinates. DIRAC-NO is the only model that keeps the input as an atomic measure until query time. The comparison therefore asks not only which backbone is stronger, but what object each model is allowed to learn from.

Training and evaluation. All models are trained under the same optimizer, early-stopping, and sparsity protocols. The primary metric is relative L^2 error, averaged over three random seeds. Full hyperparameters, tensor shapes, and dataset splits are given in the appendix.

4.2 Performance Across Observation Budgets

The first test examines how performance changes as the number of observed events decreases. In MeasureBench, reducing the observation budget not only makes the observed event measure sparser, but also removes some causal source events that contribute to the target field. The comparison therefore evaluates how different input representations behave under progressively incomplete event coverage.

Table 1 reports the MeasureBench-1D sparsity sweep. At $n = 128$, DIRAC-NO attains RelL2 $= 0.015$, compared with 0.490 for FNO and 0.170 for CNP. As the observation budget decreases, all methods degrade. The advantage of DIRAC-NO is strongest when event coverage remains sufficiently complete, while at lower budgets the error is increasingly influenced by unobserved source contributions.

Figure 3 shows the same trend on both 1D and 2D benchmarks. The 2D setting is more diagnostic because sparse geometry is harder to hide behind a one-dimensional interpolation rule. DIRAC-NO remains strongest when event coverage is sufficiently complete, while all methods deteriorate as the contribution of unobserved sources becomes more significant. Additional sparsity curves and field-level comparisons are reported in Appendix B; structural diagnostics, boundary cases, and efficiency analyses are reported in Appendices C, E, and F.

4.3 Additive Structure Is Preserved in Trained Models

The second test targets the structural claim behind the architecture. In a source-driven event-to-field problem, adding a source should add its response; it should not renormalize the entire input. We test whether this structure remains visible after training.

For each test sample, we randomly partition the event set into two disjoint subsets μ_1 and μ_2, and compute

$$\varepsilon_{\mathrm{sup}} = \frac{\|\mathcal{F}_\theta(\mu_1 + \mu_2) - \mathcal{F}_\theta(\mu_1) - \mathcal{F}_\theta(\mu_2)\|_2}{\|\mathcal{F}_\theta(\mu_1 + \mu_2)\|_2 + \epsilon},$$

averaged over 200 test samples with 10 random splits each.

Table 2 shows that DIRAC-NO achieves the lowest violation on both benchmarks: $\varepsilon_{\mathrm{sup}} = 0.086$ in 1D and 0.121 in 2D. This connects Proposition 1 to trained behavior. The full model is not linear end

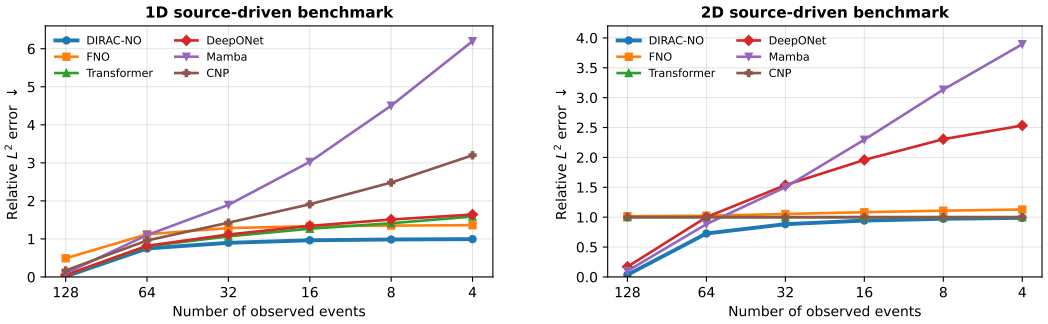


Figure 3: Performance across observation budgets. Relative L^2 error on the 1D and 2D source-driven benchmarks as the number of observed events decreases. Performance degrades for all methods as event coverage becomes incomplete. DIRAC-NO performs best when event coverage remains sufficiently complete, while lower observation budgets are increasingly affected by unobserved source contributions.

Table 2: Superposition fidelity ($\varepsilon_{\text{sup}}\downarrow$, mean$\pm$std, 200 samples $\times$ 10 splits).

Method	1D	2D
DIRAC-NO	$\mathbf{0.086}\pm.111$	$\mathbf{0.121}\pm.031$
DeepONet	$0.631\pm.089$	$0.860\pm.067$
Transformer	$0.663\pm.080$	$1.000\pm.000$
FNO	$1.021\pm.287$	$1.121\pm.255$
Mamba	$1.059\pm.075$	$1.194\pm.098$

to end, but its measure-to-function front end keeps the learned response much closer to the additive law than interpolation-first, normalized-attention, or mean-aggregation alternatives.

The ablations tell the same story. Value encoding, adaptive bandwidth, and multi-head structure all help, but replacing direct summation with mean aggregation causes the largest degradation: on the 2D benchmark at $n = 128$, RelL2 increases from 0.035 to 0.427. This is the ablation counterpart of the superposition test. Once sparse events are treated as the native input, aggregation is part of the operator design, not a secondary implementation choice. Full ablation tables and family-specific diagnostics are reported in Appendix C; harder source-driven PDE families are reported in Appendix B.

4.4 Real Data Supports the Operating Regime

The synthetic PDE benchmarks isolate the representation question, but real sparse reconstruction removes the analytic Green-function setup. We use ERA5 and NYC Taxi to test whether the same representation choice remains useful when fields and observation geometry come from data.

ERA5 is the main real-world evaluation. It remains a sparse observation-to-field reconstruction task, but the target is a geophysical field rather than an analytically generated solution. In this setting, DIRAC-NO achieves the best performance at high observation budgets, reaching RelL2 $= 0.168$ at $n = 128$. This suggests that operating directly on sparse events can remain useful beyond controlled point-source PDEs.

NYC Taxi provides a second event-to-field benchmark in a data-driven urban domain. The observations are event-native, and the target is a spatial demand field rather than a physical PDE solution. This weakens the idealized source-response structure but keeps the sparse event-to-field form. Together, ERA5 and NYC Taxi support the same reading: the benefit is strongest when the task is sparse, event-native, and at least approximately source-driven. Full real-world results are reported in Appendix B.

4.5 Negative Controls: When Event-Native Input Is Not Enough

The final experiments separate two conditions that are easy to conflate: sparse event input and source-response structure. DIRAC-NO is designed for their intersection. Event-native input alone should not be enough if the target does not behave like a field response to localized sources.

We use OOD shifts and financial event streams as diagnostic controls. Parameter shift tests transfer beyond a fixed PDE setting. Density shift changes sparse event geometry. Intensity shift moves event amplitudes outside the nominal training range. On MeasureBench-2D, DIRAC-NO performs best under parameter and density shifts, where preserving sparse geometry is central. Intensity shift remains difficult for all methods, and FNO attains lower numerical error in one setting. This pattern is informative: value-distribution shift is not the same failure mode as geometry distortion.

Financial event streams play a narrower role. They are event-native, but they do not exhibit a clear additive source-to-field mechanism. We therefore treat them as negative controls rather than primary evidence. Their role is not to claim universal superiority, but to ask whether sparsity alone is enough. The answer is no. The advantage appears where the representation argument predicts: sparse event geometry and source-response structure must coincide.

5 Scope and Limitations

The experiments define an operating regime for measure-native operator learning. DIRAC-NO is most useful when two conditions coincide: the input is naturally sparse and event-based, and the target behaves, at least approximately, like a field response to localized sources. This is where interpolation-first pipelines are most likely to change the learning problem before the operator sees it.

The boundary is equally important. Event-native input alone is not enough. Financial event streams, for example, are sparse and event-based, but they do not exhibit a clear additive source-to-field mechanism. Intensity-shift OOD stresses amplitude extrapolation rather than sparse geometry. These cases are not exceptions to the claim; they define its operating regime.

A second boundary comes from kernel geometry. The current implementation uses Gaussian-style adaptive Dirac kernels, which are well aligned with diffusion-dominated elliptic and parabolic responses but less suited to wave-dominated hyperbolic problems, where energy concentrates along characteristic fronts. Finally, the measure-to-function layer requires $\mathcal{O}(nQ)$ query-event interactions. This cost is effective in the sparse regime studied here, but may become less attractive under dense sensing or substantially finer output grids. We discuss wavefront-aware kernels and boundary cases in Appendix E, and faster query-event aggregation in Appendix F.

6 Conclusion

Sparse event-to-field learning is often treated as a data-scarcity problem: fewer observations simply make the operator harder to learn. This paper argues for a different view. In sparse source-driven systems, the first modeling decision is not only the neural backbone, but the space in which the input is represented. Interpolation answers this question before learning begins by turning atomic events into a surrogate field.

We introduced DIRAC-NO, a measure-native neural operator that keeps sparse observations as event measures until query time. Its adaptive Dirac-kernel aggregation constructs field values directly from atomic inputs, preserves source-wise additivity at the aggregation stage, and provides a learned Green-kernel view of source-to-field response. Across controlled PDEs, real sparse reconstruction tasks, structural diagnostics, and negative controls, the evidence points to the same conclusion: the advantage is strongest when sparse event geometry and source-response structure coincide.

The message is not that measure-native learning wins everywhere. It is that interpolation is not a harmless preprocessing step in sparse event-to-field problems. It is the first modeling choice. DIRAC-NO keeps that choice open until the model has a query to answer.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Nicolas Boullé, Seick Kim, Tianyi Shi, and Alex Townsend. Learning green's functions associated with time-dependent partial differential equations. *Journal of Machine Learning Research*, 23 (218):1–34, 2022.

Roger Daley. *Atmospheric Data Analysis*. Cambridge University Press, 1993.

Dick P Dee, Sakari M Uppala, Adrian J Simmons, Paul Berrisford, Paul Poli, Shinya Kobayashi, Ulf Andrae, Magdalena A Balmaseda, Gianpaolo Balsamo, Peter Bauer, et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, 2011. doi: 10.1002/qj.828.

Lawrence C Evans. *Partial Differential Equations*. American Mathematical Society, 2 edition, 2010.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2 edition, 1999.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Ali Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

Craig R Gin, Daniel E Shea, Steven L Brunton, and J Nathan Kutz. Deepgreen: Deep learning of green's functions for nonlinear boundary value problems. *Scientific Reports*, 11(1):21614, 2021.

Leslie Greengard and Vladimir Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348, 1987.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Bruce Ingleby. Global assimilation of air temperature, humidity, wind and pressure from surface stations. *Quarterly Journal of the Royal Meteorological Society*, 141(687):504–517, 2015. doi: 10.1002/qj.2372.

Fritz John. *Partial Differential Equations*. Springer, 4 edition, 1982.

Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003.

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.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753, 2019.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6755–6766, 2020.

Zongyi Li, Nikola 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.

Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3428–3434, 2018.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

New York City Taxi and Limousine Commission. Nyc taxi and limousine commission trip record data, 2023. URL `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. Accessed 2026-04-25.

Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1720–1730, 2019.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Peter A G Watson, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020MS002203.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Walter Rudin. *Functional Analysis*. McGraw-Hill, 2 edition, 1991.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4585–4593, 2021. doi: 10.24963/ijcai.2021/623.

Ivar Stakgold and Michael Holst. *Green's Functions and Boundary Value Problems*. John Wiley & Sons, 3 edition, 2011.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2588–2595, 2018. doi: 10.1609/aaai.v32i1.11836.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 1655–1661, 2017. doi: 10.1609/aaai.v31i1.10735.

The appendix makes the paper's representation-level claim checkable. The main text argues that sparse event-to-field learning should not automatically begin with interpolation: before the operator learns, the input space has already been chosen. We therefore record how sparse event measures are constructed, what each model is allowed to see before learning, and how sparsity, OOD shifts, and real-data settings are imposed. We then report the full empirical evidence behind the main claims, including complete sparsity sweeps, real reconstruction results, ablations, superposition diagnostics, and negative controls. Finally, we give the structural details behind query-time summation, additive source response, the Green-kernel view, efficiency, and boundary diagnostics. The question throughout is not whether measure-native learning wins everywhere, but when the input should remain a measure until the model has a query to answer.

A Representation Protocol: What Is Compared Before Learning?

The central comparison in this paper is not only between neural backbones. It is between input spaces. Before a model learns, a sparse observation has either remained an event measure, been converted into a grid field, been padded into a sequence, or been formatted as a context-query set. This appendix makes that choice explicit. We specify what is observed, how sparsity is imposed, and what object each model is allowed to learn from before the operator is applied. The goal is not only reproducibility, but to make the representation choice itself checkable.

A.1 What Each Benchmark Is Designed to Test

Table 3 defines the operating regime tested by each benchmark. The controlled PDE benchmarks isolate the core source-response setting where the representation argument should be strongest. ERA5 and NYC Taxi test whether the same idea remains useful once the field and observation geometry come from real data. Financial streams and hyperbolic dynamics play a different role: they are boundary cases. They keep some parts of the event-native setting while weakening the source-response structure or the kernel geometry that motivates DIRAC-NO.

Table 3: **What each benchmark is designed to test.** The expected advantage is strongest when sparse event input and source-response structure coincide. Boundary cases are included to test where the representation argument should weaken.

Benchmark	Role in the paper	Event-native input	Source-response structure	Expected advantage
MeasureBench-1D	Controlled source-response test	Present	Strong	Strong
MeasureBench-2D	Spatial stress test	Present	Strong	Strong
ERA5 Weather	Real sparse reconstruction	Present	Approximate	Moderate
NYC Taxi	Real event-to-field reconstruction	Present	Approximate	Moderate
Financial streams	Negative control	Present	Weak	Limited
Hyperbolic PDE	Kernel-geometry boundary	Present	Mismatch	Limited

A.2 What Each Model Sees Before Learning

Because the claim is representational, the input adaptation for each baseline is part of the experiment. Table 4 records the object each model receives before learning begins. An interpolation-first baseline is therefore not only a different architecture. It is a different answer to the question of how sparse events should be represented before operator learning starts.

Table 4: **The object each model learns from.** Only DIRAC-NO keeps the input as an atomic event measure until query time. Interpolation-first baselines answer the representation question before the operator is trained.

Model	Input representation	Preprocessing before learning	Role in comparison
DIRAC-NO	Atomic event measure	None	Measure-native operator
FNO	Dense grid field	Rasterization / interpolation	Interpolation-first operator
DeepONet	Sparse event tuples + query coordinates	Branch/trunk formatting	Operator baseline with query input
Transformer	Padded event sequence	Padding and masking	Generic sparse sequence model
Mamba	Ordered event sequence	Ordering, padding, masking	State-space sequence baseline
CNP	Context set and query set	Context-query formatting	Sparse set prediction baseline

A.3 MeasureBench-1D

MeasureBench-1D is the cleanest positive case in the paper. It is not designed for realism; it is designed for isolation. The input is natively a sparse collection of point events, the target field is generated by a known source-response law, and any interpolation-first pipeline must introduce an additional approximation before learning begins.

We simulate point-source diffusion in one spatial dimension on the domain $\Omega = [0, 1]$ over a finite time interval $[0, T]$:

$$u_t = \kappa u_{xx} - \beta u + \sum_{i=1}^{n} a_i \, \delta(x - x_i)\delta(t - t_i), \qquad x \in [0, 1], \; t \in [0, T]. \tag{5}$$

Each sample is defined by a sparse event measure

$$\mu = \sum_{i=1}^{n} a_i \, \delta(x - x_i)\delta(t - t_i),$$

where $x_i \in [0, 1]$ is the source location, $t_i \in [0, T]$ is the source time, and $a_i \in \mathbb{R}$ is the source amplitude.

For the benchmark used in the main text, we fix

$$\kappa = 0.018, \qquad \beta = 0.35.$$

Source locations, source times, and amplitudes are sampled independently from bounded distributions. The target field is evaluated on a dense set of query points at the final observation time. Since the ground-truth field is generated analytically from the corresponding Green's function, the benchmark is controlled and noise-free by construction.

This deliberately simple setting makes the representation question visible. If the input is already an atomic measure, then rasterization or smoothing is not a neutral step: it changes the object on which the operator learns. MeasureBench-1D therefore tests whether remaining in measure space changes the learned operator before any geometric or real-data complications are added.

To study robustness under missing observations, we retain only

$$n \in \{128, 64, 32, 16, 8, 4\}$$

events at test time. The same sparsity sweep is used throughout the paper.

A.4 MeasureBench-2D

MeasureBench-2D extends the same representation test to spatial geometry. The source-response law remains controlled, but the output field is larger and interactions between sparse sources are harder to hide behind a one-dimensional interpolation rule.

We consider the domain

$$\Omega = [0, 1]^2$$

and the point-source PDE

$$u_t = \kappa \Delta u - \beta u + \sum_{i=1}^{n} a_i \, \delta(x - x_i)\delta(t - t_i), \qquad x \in [0, 1]^2, \; t \in [0, T]. \tag{6}$$

Each event location $x_i \in [0, 1]^2$ is two-dimensional, while event times and amplitudes are sampled as in the 1D benchmark.

The output field is evaluated on a 32×32 query grid. As in the 1D setting, we vary the number of observed events over

$$n \in \{128, 64, 32, 16, 8, 4\}.$$

The 2D benchmark is the main synthetic stress test in the paper. It remains analytically controlled, but the sparse geometry is rich enough that the representation choice is harder to hide behind model capacity alone.

A.5 ERA5 Weather Reconstruction

ERA5 tests the same representation choice after the analytic Green-function setup is removed. The task remains sparse observation-to-field reconstruction, but the target is now a real atmospheric field rather than a synthetic source-response solution.

We use the WeatherBench2 benchmark [Rasp et al., 2024], derived from ERA5 reanalysis data [Hersbach et al., 2020, Dee et al., 2011], and focus on 2-meter air temperature. This setting is related to data-driven weather forecasting and evaluation [Rasp et al., 2020, Bi et al., 2023, Lam et al., 2023], while remaining close to the classical sparse-observation view of atmospheric reconstruction [Kalnay, 2003, Daley, 1993, Ingleby, 2015].

The original global field is downsampled to a 32×64 grid. Training covers 2015–2019, validation uses the first half of 2020, and testing uses the second half of 2020. At each time step, we sample a sparse subset of observation sites to emulate irregular weather-station measurements. Each retained station contributes one event with a spatial coordinate and an observed temperature value.

The target is the full temperature field at the same time step. As in the synthetic experiments, we evaluate performance under the sparsity sweep

$$n \in \{128, 64, 32, 16, 8, 4\},$$

where n is the number of retained station observations.

A.6 NYC Taxi Event Stream

NYC Taxi tests the event-to-field formulation in a data-driven urban domain. It is event-native from the outset: the observed input is a set of pickup events, and the target is a spatial demand field. Unlike the controlled PDE benchmarks, however, the source-response structure is weaker and less physically explicit.

We use the NYC Taxi & Limousine Commission trip-record data [New York City Taxi and Limousine Commission, 2023] and restrict attention to Manhattan. This benchmark is related to sparse spatiotemporal forecasting and urban demand prediction [Zhang et al., 2017, Yao et al., 2018, Li et al., 2018, Liang et al., 2018, Pan et al., 2019], but we cast it here as reconstruction rather than pure forecasting.

Pickup locations are mapped to continuous spatial coordinates using deterministic zone centroids. Trips are grouped into fixed temporal windows to form sparse event sets. Each sample takes the form

$$\{(\mathrm{lat}_i, \mathrm{lon}_i, t_i)\}_{i=1}^{n},$$

together with the corresponding event intensities. The target is a coarse-grained 32×32 density field representing the spatial distribution of taxi demand during the same time window.

There is no hidden analytic field before preprocessing in this benchmark. The observations are events by construction. The question is whether preserving the sparse event representation remains useful when the task is still event-to-field but less source-driven than the PDE benchmarks.

As in the other benchmarks, we vary the number of retained events over

$$n \in \{128, 64, 32, 16, 8, 4\}.$$

A.7 Financial Event Stream Setup

The financial event streams are diagnostic controls. They are event-native, but they do not provide a clean additive source-to-field mechanism. They therefore separate two conditions that are easy to conflate: being sparse and event-based, and being governed by a source-response operator.

We evaluate on four real financial event datasets: BTC spot trades, ETH spot trades, Kweichow Moutai OHLC sequences, and Kweichow Moutai tick-event streams. Each sample is represented as a sparse event sequence rather than as a gridded time series, and we use the same observation-budget sweep

$$n \in \{128, 64, 32, 16, 8, 4\}.$$

These benchmarks should not be read as finance-specific endorsements of DIRAC-NO. Their role is diagnostic. In contrast to source-driven PDEs and sparse reconstruction tasks, the mapping from

sparse trades or ticks to the prediction target does not naturally decompose into a superposition of localized event responses. The financial setting therefore asks whether event-native input alone is enough.

The answer we expect is no. If the representation argument is correct, DIRAC-NO should be most useful when sparse event geometry and source-response structure coincide. Financial streams test the boundary of that claim rather than serving as primary evidence. In particular, Mamba-1D is omitted when the number of available training samples falls below 2000; these cases are marked explicitly in the corresponding result tables.

A.8 Shared Implementation Protocol

This subsection fixes the implementation protocol used across experiments. The purpose is not to list engineering choices in isolation. It is to make the comparison auditable: all models are trained from the same underlying sparse observations, but they are not allowed to see those observations in the same form. The protocol therefore separates three ingredients: the native event set, the representation adapter given to each baseline, and the shared optimization and evaluation rules.

DIRAC-NO architecture. Unless otherwise stated, DIRAC-NO uses $H = 8$ adaptive kernel heads with learnable base bandwidths and latent width $d_{\mathrm{model}} = 128$. The event-value encoder is a two-layer MLP with LayerNorm and GELU activations. The decoder is a shallow MLP that maps query-wise latent features to scalar field values. In 2D settings, an optional spectral mixing block is applied after the measure-to-function layer to capture longer-range spatial correlations. This block acts after query-time aggregation and is not part of the additive measure-to-function front end.

Where the representation choice enters. All models are trained on the same underlying sparse observations. What differs is the object each model receives before learning begins. FNO [Li et al., 2021] receives a rasterized or binned grid field; the sparse event set has already been converted into a function. Transformer [Vaswani et al., 2017] and Mamba [Gu and Dao, 2023] receive padded event sequences with binary masks. CNP [Garnelo et al., 2018] receives a context-query set. DeepONet [Lu et al., 2021] receives sparse event tuples in the branch network and query coordinates in the trunk network. Only DIRAC-NO keeps the input as an atomic measure until query time. Thus, the comparison is not only between backbones; it is between the objects from which those backbones are asked to learn.

Optimization. All models are trained under a shared optimization protocol. We use AdamW [Loshchilov and Hutter, 2019] with learning rate 10^{-3} and weight decay 10^{-4}. Training uses cosine annealing with warmup followed by warm restarts [Loshchilov and Hutter, 2017], batch size 32, and gradient clipping with maximum norm 1.0 [Pascanu et al., 2013]. Runs continue for up to 500 epochs with early stopping based on validation performance.

Evaluation. The primary metric is relative ℓ_2 error,

$$\varepsilon = \frac{\|\hat{u} - u\|_2}{\|u\|_2},$$

computed on the full target field. Headline numbers are averaged over three random seeds. We report in-distribution sparsity sweeps and OOD diagnostics using the shift types defined in Appendix B. The same observation budgets,

$$n \in \{128, 64, 32, 16, 8, 4\},$$

are used throughout, so sparsity remains a common axis of comparison.

Reproducibility units. All figures are generated from stored result summaries and saved model outputs. PDE benchmarks, ERA5 experiments, NYC Taxi experiments, financial experiments, OOD studies, and ablation runs are organized in separate result directories. This organization mirrors the evidence trail of the paper: controlled source-response tests, real sparse reconstruction, structural diagnostics, negative controls, and efficiency analyses can be reproduced independently.

Table 5: **Shared training protocol.** These settings standardize optimization across architectures; the main experimental difference is the representation each model receives before learning.

Component / setting	Value
Number of kernel heads H	8
Latent width d_{model}	128
Optimizer	AdamW
Learning rate	10^{-3}
Weight decay	10^{-4}
Batch size	32
Warmup	30 epochs
Maximum training length	500 epochs
Early stopping patience	80 epochs
Gradient clipping	max norm 1.0
Random seeds	3
Sparsity levels	$128, 64, 32, 16, 8, 4$

A.9 Dataset Splits and Observation Protocols

Table 6 records how each benchmark is split and how sparse observations are imposed. The synthetic PDE benchmarks are analytically generated and noise-free by construction. ERA5, NYC Taxi, and the financial streams are derived from real data and require explicit subsampling. Across all settings, the observation budget is controlled directly, so changes in performance can be read against the same sparsity axis rather than against different dataset protocols.

Table 6: **Dataset splits and observation protocols.** Each benchmark is evaluated under the same observation-budget sweep, while the observation unit changes with the native data source.

Benchmark	Train / Val / Test	Observation unit	Sparse observation rule
MeasureBench-1D	synthetic split	point source event	retain $n \in \{128, 64, 32, 16, 8, 4\}$ events
MeasureBench-2D	synthetic split	point source event	retain $n \in \{128, 64, 32, 16, 8, 4\}$ events
ERA5	2015–2019 / 2020-H1 / 2020-H2	weather station	subsample stations to target budget n
NYC Taxi	temporal split	pickup event	subsample observed pickups to target budget n
Financial streams	temporal split	trade / tick / OHLC event	subsample observations to target budget n

A.10 Input and Output Tensor Shapes

Although the conceptual input is always a sparse event measure, the tensors passed to different models are intentionally different. This is not an implementation nuisance. It records the representation decision made before learning: the same sparse observation may remain atomic, become a sequence, become a context set, or be converted into an interpolated grid. Table 7 makes this distinction explicit.

Table 7: **Tensor forms used by each representation.** The same sparse observations are exposed to different models as measures, sequences, context-query sets, or interpolated grid fields.

Component	Symbolic shape	Meaning	Used by
Sparse event tensor	$[B, n, d_{\text{feat}}]$	event tuples (x, t, a) or (x, y, t, a)	DIRAC-NO / Transformer / Mamba / DeepONet
Query tensor	$[B, Q, d_{\text{coord}}]$	query coordinates	DIRAC-NO / DeepONet
Rasterized grid input	$[B, C, H, W]$ or $[B, C, N]$	interpolated or binned field	FNO
Output field	$[B, Q, 1]$ or $[B, H, W, 1]$	target scalar field	all methods

B Full Evidence Trail

This section expands the empirical evidence behind the main text. The purpose is not to add disconnected results. It is to make the representation claim traceable across the regimes in which it should hold, weaken, or fail. The guiding question is the same throughout: how does preserving sparse observations in their native event form affect performance across different levels of event coverage?

We organize the evidence in three layers. First, the full sparsity sweeps test whether interpolation-first learning collapses earlier as the observation budget shrinks. Second, OOD shifts separate geometry changes from value-distribution changes. Third, real-data and negative-control results test whether the operating regime predicted by the representation argument is visible beyond analytic PDEs.

B.1 Full Sparsity Sweeps: Where Breakdown Begins

Tables 8 and 9 report the complete sparsity sweeps for the two source-driven PDE benchmarks. The 1D benchmark is the cleanest case: the source-response law is controlled, and the degradation pattern is easy to inspect. The 2D benchmark is more diagnostic because sparse geometry is harder to hide behind a one-dimensional interpolation rule.

The point of these tables is not that DIRAC-NO remains accurate under arbitrarily severe sparsity. No method does. The point is sharper: preserving the native event representation postpones collapse relative to pipelines that first turn sparse events into dense surrogate fields. The learned operator still fails when too little information remains, but it fails later and in the regime predicted by the representation argument.

Table 8: **MeasureBench-1D complete sparsity sweep.** Relative L^2 error (↓), mean ± std over three seeds. The clean 1D setting isolates the cost of changing an atomic source measure into a preconstructed field before learning.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	$\mathbf{0.015 \pm 0.001}$	$\mathbf{0.749 \pm 0.001}$	$\mathbf{0.897 \pm 0.001}$	$\mathbf{0.964 \pm 0.000}$	$\mathbf{0.985 \pm 0.000}$	$\mathbf{0.994 \pm 0.000}$
FNO-1D	0.490 ± 0.004	1.122 ± 0.132	1.285 ± 0.117	1.334 ± 0.116	1.350 ± 0.106	1.361 ± 0.102
Transformer	0.040 ± 0.003	0.806 ± 0.004	1.067 ± 0.020	1.268 ± 0.082	1.412 ± 0.158	1.588 ± 0.285
DeepONet-1D	0.051 ± 0.004	0.819 ± 0.027	1.116 ± 0.057	1.342 ± 0.084	1.510 ± 0.108	1.641 ± 0.115
Mamba-1D	0.095 ± 0.008	1.101 ± 0.025	1.894 ± 0.051	3.026 ± 0.088	4.503 ± 0.131	6.199 ± 0.159
CNP	0.170 ± 0.006	0.959 ± 0.016	1.426 ± 0.039	1.910 ± 0.080	2.482 ± 0.147	3.199 ± 0.229

Table 9: MeasureBench-2D submitted-baseline sparsity sweep. Relative L^2 error (↓). The Transformer and CNP values shown here correspond to the original submitted checkpoints, which were subsequently diagnosed as near-degenerate predictors; strengthened-baseline results are reported separately.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	**0.035**	**0.727**	**0.883**	**0.943**	**0.973**	**0.989**
FNO-2D	1.017	1.024	1.053	1.083	1.109	1.128
Transformer	1.000	1.000	1.000	1.000	1.000	1.000
DeepONet-2D	0.170	1.000	1.536	1.958	2.305	2.534
Mamba-2D	0.100	0.883	1.501	2.296	3.135	3.894
CNP	1.000	1.000	1.000	1.000	1.000	1.000

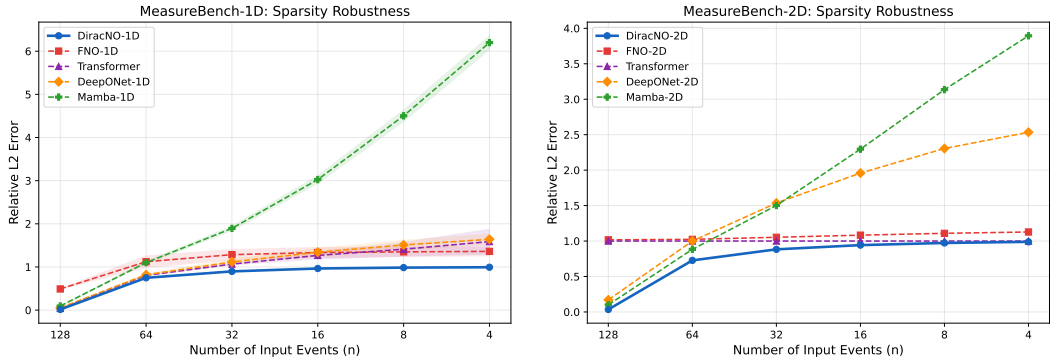


Figure 4: **Full sparsity sweeps.** In both the 1D and 2D source-driven benchmarks, DIRAC-NO delays breakdown as the number of observed events decreases. The figure should be read as a breakdown diagnostic rather than as a claim that any method remains accurate under extreme sparsity.

B.2 OOD Diagnostics: Geometry Shift Is Not Value Shift

The OOD experiments are diagnostic controls, not a single undifferentiated robustness score. The sparsity sweeps ask what happens when fewer events are observed. The OOD shifts ask a different question: which part of the event-to-field problem has changed, and which inductive bias should matter under that change?

We use three shifts.

OOD-Param. The underlying PDE parameters are changed at test time while the source distribution is otherwise kept similar. This probes whether the learned operator transfers beyond a single fixed parameter setting.

OOD-Density. The spatial density pattern of source events is changed. Test-time events become more clustered or more unevenly distributed than in training. This shift probes robustness to sparse event geometry, the regime where preserving the input as an event measure should matter most.

OOD-Intensity. The event-amplitude distribution is changed. This includes heavier-tailed intensity patterns than those seen during training. Unlike density shift, this does not primarily distort sparse geometry; it stresses the value pathway and amplitude extrapolation.

These shifts separate failure modes that are easy to conflate. Parameter shift tests operator transfer, density shift tests sparse geometry, and intensity shift tests value extrapolation. The representation argument predicts that DIRAC-NO should be strongest when sparse geometry matters most, and less decisive when the main failure mode is amplitude extrapolation.

Tables 10 and 11 report the complete OOD results at $n = 128$. The 1D and 2D cases should be read differently. In 1D, FNO attains lower absolute OOD error, but it also begins from much weaker in-distribution fit. This makes the comparison less diagnostic: a model that underfits the ID operator has less learned structure to lose under shift. In 2D, where sparse geometry is harder to hide behind a one-dimensional interpolation rule, DIRAC-NO is best under parameter and density shift. Intensity shift remains difficult for all methods and is one setting where FNO is lower. This is the intended reading: geometry shift and value-distribution shift are different failure modes.

Table 10: **OOD generalization on MeasureBench-1D.** Relative L^2 error ($\downarrow$), mean $\pm$ std over three seeds at $n = 128$. FNO has lower absolute OOD error in 1D, but this should be read together with its much weaker in-distribution fit.

Method	OOD-Param	OOD-Density	OOD-Intensity
DIRAC-NO	1.035 ± 0.004	1.193 ± 0.003	1.080 ± 0.003
FNO-1D	0.793 ± 0.010	0.600 ± 0.003	0.512 ± 0.003
Transformer	3.115 ± 0.036	2.690 ± 0.028	2.948 ± 0.032
DeepONet-1D	1.183 ± 0.013	1.069 ± 0.003	1.109 ± 0.007
Mamba-1D	52.692 ± 0.489	24.617 ± 0.277	59.241 ± 0.804

Table 11: **OOD generalization on MeasureBench-2D.** Relative L^2 error ($\downarrow$), mean $\pm$ std over three seeds at $n = 128$. DIRAC-NO is strongest under parameter and density shift, where preserving sparse event geometry is central. Intensity shift stresses event amplitudes and remains a boundary case.

Method	OOD-Param	OOD-Density	OOD-Intensity
DIRAC-NO	$\mathbf{0.932 \pm 0.008}$	$\mathbf{0.932 \pm 0.009}$	1.201 ± 0.011
FNO-2D	1.097 ± 0.002	1.170 ± 0.007	$\mathbf{1.035 \pm 0.005}$
Transformer	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
DeepONet-2D	6.024 ± 0.052	6.774 ± 0.037	7.821 ± 0.029
Mamba-2D	26.738 ± 0.403	16.302 ± 0.682	10.501 ± 0.334

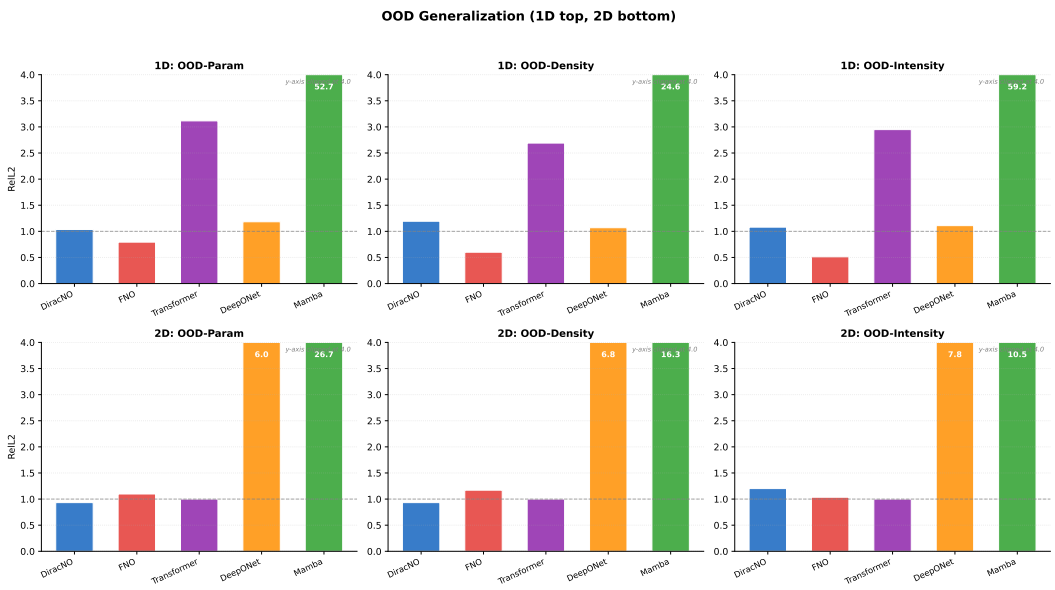


Figure 5: **OOD diagnostics.** Parameter and density shifts test operator transfer and sparse-geometry robustness; intensity shift stresses event values. The y-axis is clipped for readability, with very large Mamba and DeepONet values annotated directly.

B.3 Why Relative OOD Degradation Must Be Read with ID Fit

Absolute OOD error is the primary quantity, but relative degradation explains a common failure mode in interpreting OOD tables. When two models begin from very different in-distribution errors, the same OOD number can mean different things. We therefore also report

$$\Delta_{\mathrm{OOD}} = \frac{\text{OOD error} - \text{ID error}}{\text{ID error}}.$$

This ratio is diagnostic, not primary. It distinguishes models that are genuinely stable from models that appear stable because they already underfit the in-distribution operator.

This distinction is most visible on MeasureBench-1D. FNO begins with much higher ID error than DIRAC-NO. Its relative degradation can therefore appear small even when its absolute predictions remain less accurate in-distribution. A model with low ID error has more learned structure to lose under distribution shift; a model that already fits poorly can appear insensitive for the wrong reason. OOD behavior should therefore be read together with ID fit.

Table 12: **Relative OOD degradation on MeasureBench-1D.** Relative degradation is a diagnostic ratio, not the primary metric. A weak ID fit can make a model appear less sensitive under shift because there is less learned structure to lose.

Method	ID error	Δ-Param	Δ-Density	Δ-Intensity
DIRAC-NO	0.015	68.0×	78.5×	71.0×
FNO	0.490	0.62×	0.22×	0.04×
Transformer	0.040	76.9×	66.3×	72.7×
DeepONet	0.051	22.2×	20.0×	20.8×
Mamba	0.095	554×	258×	623×

B.4 Real Sparse Reconstruction: When the Analytic PDE Is Removed

The controlled PDE benchmarks isolate the representation question, but they do so under an analytic source-response law. ERA5 and NYC Taxi remove that assumption. They test whether the measure-native formulation remains useful when the field, observation geometry, and noise structure come from data. These results should be read through the operating regime of the paper: the advantage should be strongest when the task remains sparse, event-native, and at least approximately source-driven.

B.4.1 ERA5: Real Geophysical Field Reconstruction

ERA5 is the main real-world reconstruction test. The observations emulate sparse weather-station measurements, while the target is a real geophysical field rather than a synthetic diffusion solution. The task is therefore no longer generated by the analytic Green-function process used in MeasureBench, but it remains a sparse observation-to-field problem.

Table 13 reports the full sparsity sweep. DIRAC-NO is strongest at the highest observation budget, reaching RelL2 $= 0.168$ at $n = 128$. As the observation budget decreases, the ranking becomes more benchmark-dependent. This is expected: once too little information remains, missing observations dominate the problem for all methods. The conclusion is therefore precise rather than universal: measure-native learning is most useful before the task becomes information-starved.

Table 13: **ERA5 full sparsity results.** Relative L^2 error ($\downarrow$), mean $\pm$ std over three seeds. DIRAC-NO is strongest at the highest observation budget, while lower-budget regimes become increasingly information-limited.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	$\mathbf{0.168 \pm 0.002}$	0.253 ± 0.007	0.420 ± 0.015	0.536 ± 0.010	0.597 ± 0.006	0.625 ± 0.006
FNO-2D	0.189 ± 0.003	0.448 ± 0.008	0.725 ± 0.011	0.880 ± 0.010	0.957 ± 0.011	0.982 ± 0.012
Transformer	0.175 ± 0.002	$\mathbf{0.194 \pm 0.002}$	$\mathbf{0.214 \pm 0.001}$	$\mathbf{0.253 \pm 0.005}$	0.364 ± 0.026	0.637 ± 0.089
DeepONet	0.236 ± 0.003	0.251 ± 0.003	0.266 ± 0.002	0.288 ± 0.003	$\mathbf{0.323 \pm 0.009}$	$\mathbf{0.376 \pm 0.027}$
Mamba	0.187 ± 0.002	0.267 ± 0.012	0.615 ± 0.110	0.973 ± 0.094	1.039 ± 0.101	0.964 ± 0.061

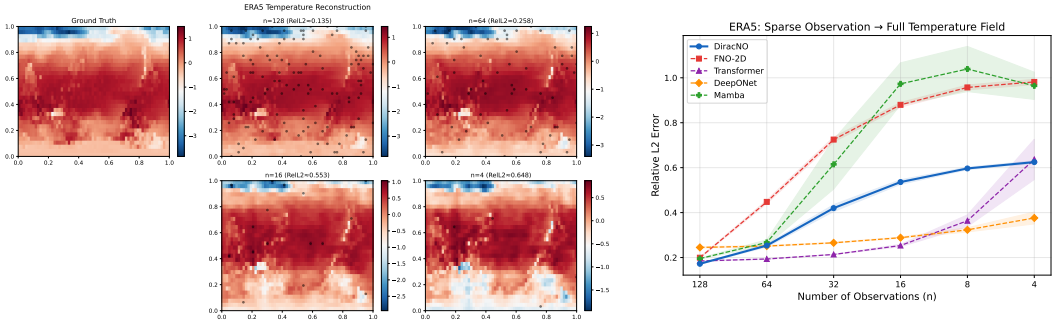


Figure 6: **ERA5 reconstruction.** The left panel shows representative reconstructions from sparse station observations. The right panel shows error as the number of observed stations decreases.

B.4.2 NYC Taxi: Event-to-Density Reconstruction

NYC Taxi provides a second real-data test with weaker physical structure. The observations are discrete pickup events and the target is a spatial demand-density field. Unlike ERA5, the task is not naturally tied to a classical field equation. It remains event-to-field, but its source-response structure is less explicit.

Table 14 reports the full sparsity sweep. DIRAC-NO performs best or near-best at higher observation budgets and remains competitive as sparsity increases. As in ERA5, the claim is not that any method avoids degradation under extreme sparsity. The claim is that preserving the event-native representation is useful before missing information dominates the task.

Table 14: **NYC Taxi full sparsity results.** Relative L^2 error ($\downarrow$), mean $\pm$ std over three seeds. The task is event-native but less physically explicit than the PDE benchmarks, so it tests the moderate operating regime of the method.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	$\mathbf{0.160 \pm 0.000}$	$\mathbf{0.252 \pm 0.013}$	$\mathbf{0.638 \pm 0.011}$	$\mathbf{0.873 \pm 0.030}$	0.948 ± 0.033	0.978 ± 0.033
FNO-2D	0.162 ± 0.001	0.468 ± 0.021	0.771 ± 0.030	0.897 ± 0.028	0.954 ± 0.022	0.982 ± 0.017
Transformer	0.901 ± 0.170	0.901 ± 0.170	0.903 ± 0.166	0.918 ± 0.144	0.992 ± 0.035	1.118 ± 0.152
DeepONet	0.934 ± 0.008	0.934 ± 0.008	0.934 ± 0.008	0.934 ± 0.008	$\mathbf{0.935 \pm 0.008}$	$\mathbf{0.935 \pm 0.008}$
Mamba	0.575 ± 0.066	0.687 ± 0.052	1.036 ± 0.081	1.659 ± 0.137	2.254 ± 0.230	2.342 ± 0.453

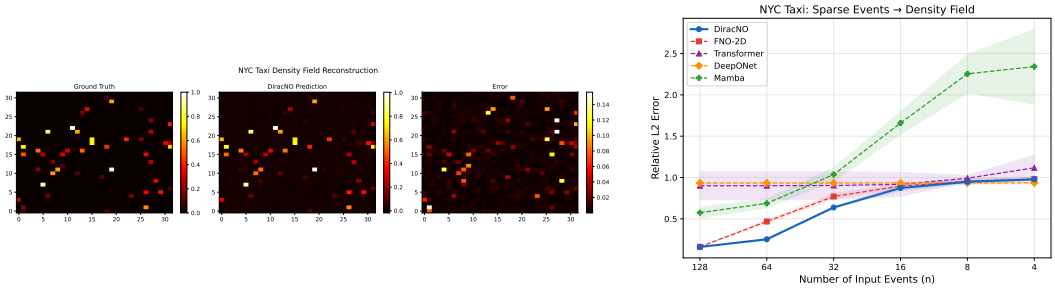


Figure 7: **NYC Taxi reconstruction.** The left panel shows a representative density-field prediction and error. The right panel reports error as the number of observed pickup events decreases.

B.5 Harder Source-Driven PDE Families

The default synthetic benchmark is intentionally clean: it isolates the representation question by combining sparse point sources with a diffusion-dominated response. A natural stress test is to keep the same event-native input but make the source-to-field map less symmetric. If the advantage of DIRAC-NO comes from the measure-native formulation rather than from an easy isotropic benchmark, it should persist when source responses are shaped by transport or direction-dependent diffusion.

We therefore evaluate two harder 2D PDE families. The advective family adds stronger transport, so localized sources generate fields shaped by directional drift. The anisotropic family changes the diffusion geometry itself, so responses spread at different rates across directions. These experiments are not separate application benchmarks. They are controlled stress tests of the same representation claim: sparse source-driven fields should be learned from the event measure rather than from an interpolated surrogate grid.

B.5.1 Advective 2D Family

The advective family strengthens transport relative to the default 2D benchmark. The response is no longer shaped only by local diffusion around each source; it also depends on directional drift. This makes the target field more sensitive to transport geometry and less aligned with isotropic smoothing.

Using the stored family-specific checkpoints, DIRAC-NO attains RelL2 $= 0.097$ at $n = 128$, compared with 1.027 for FNO-2D, 1.000 for Transformer, 0.177 for DeepONet-2D, and 0.109 for Mamba-2D. More importantly, the gain is not confined to the densest regime. Across the full sparsity sweep, DIRAC-NO remains the strongest overall curve and degrades later as the number of observed events decreases. This indicates that the measure-native advantage is not limited to isotropic diffusion; it persists when sparse event geometry must be combined with directional transport.

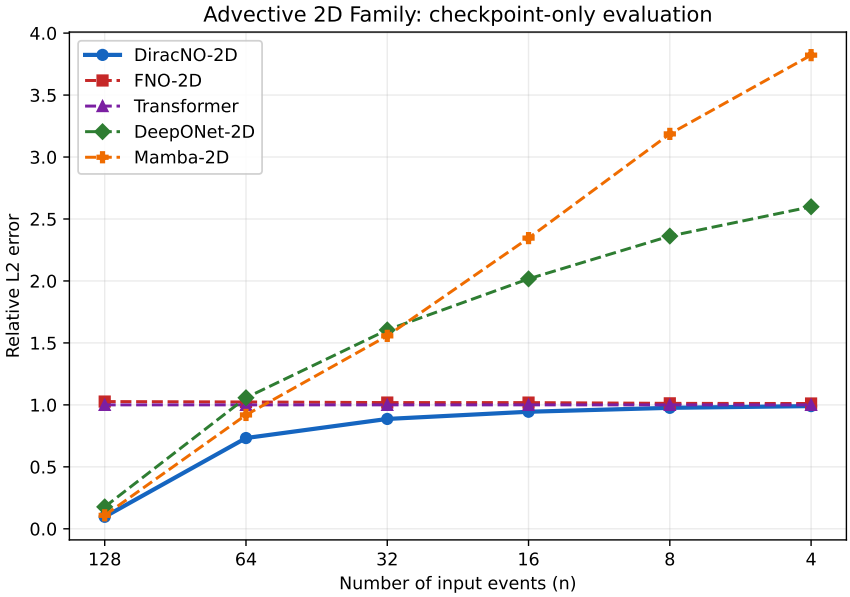


Figure 8: **Advective 2D source-response family.** Stronger transport makes the response more direction-dependent. DIRAC-NO remains the strongest model across the sparsity sweep, suggesting that the measure-native advantage is not limited to the default isotropic benchmark.

B.5.2 Anisotropic Diffusion 2D Family

The anisotropic family stresses the source-response map in a different way. Instead of adding drift, it breaks isotropy in the diffusion operator itself. Fields spread at different rates across spatial directions, producing elongated and directionally biased responses even when the source configuration remains sparse.

This family also probes the kernel geometry of DIRAC-NO. The current adaptive Dirac kernel is more flexible than a fixed isotropic smoother, but it is still a relatively simple local response family. Anisotropic diffusion therefore tests whether keeping the input as a measure continues to help before a richer directional kernel becomes necessary.

Using the stored family-specific checkpoints, DIRAC-NO reaches RelL2 $= 0.060$ at $n = 128$, compared with 1.023 for FNO-2D, 1.000 for Transformer, 0.194 for DeepONet-2D, and 0.137 for Mamba-2D. As in the advective family, the full sparsity curve remains consistently strongest for

DIRAC-NO. The result supports a precise conclusion: the benefit of operating in measure space extends beyond the default isotropic benchmark when the problem remains sparse and source-driven.

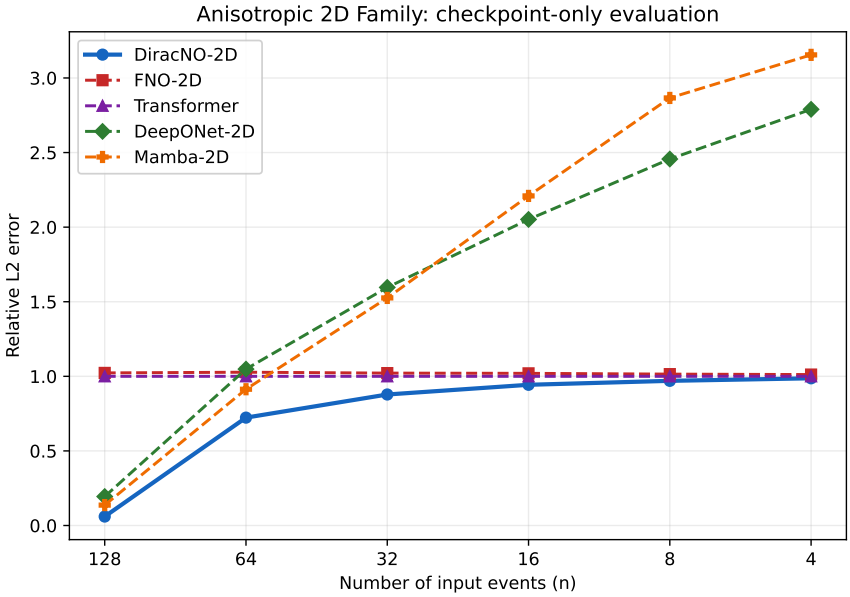


Figure 9: **Anisotropic 2D source-response family.** Direction-dependent diffusion produces elongated responses that are less aligned with simple isotropic smoothing. DIRAC-NO remains strongest across the sparsity sweep, indicating that the advantage persists under harder source-response geometry.

B.6 Field-Level and Scaling Diagnostics

Scalar errors summarize performance, but they do not show how reconstructions fail. We therefore include field-level comparisons and scaling summaries to visualize the same representation effect from different angles.

Field-level comparisons. Figure 10 compares full field predictions on MeasureBench-2D. The visual pattern matches the numerical results: DIRAC-NO tends to preserve coarse global structure even when fine detail is lost, whereas interpolation-first baselines more often show oversmoothing, distortion, or artifacts.

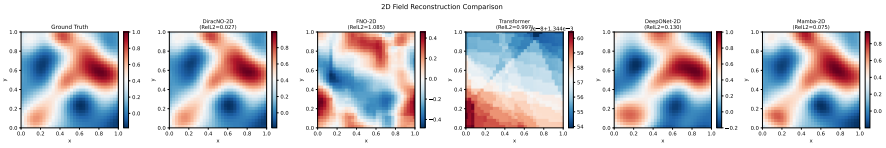


Figure 10: **Field-level comparison on MeasureBench-2D.** DIRAC-NO better preserves large-scale target structure, while baseline errors are more often dominated by oversmoothing or interpolation artifacts.

B.7 Additional Sparsity Diagnostics

The tables above report the numerical sparsity sweeps. The following figures show the same phenomenon from complementary angles. They are not separate claims. They are diagnostics for where the approximation enters the pipeline: before learning, through sparse-to-grid conversion, or at query time, through measure-to-function aggregation.

DIRAC-NO versus FNO. Figure 11 isolates the strongest interpolation-first comparison. FNO is competitive when the effective sampling density is high enough that the surrogate grid still

contains useful information. The gap becomes clearer as observations become sparse. In that regime, interpolation error is introduced upstream of operator learning, before the model has a chance to decide how each event should contribute to a queried field value.

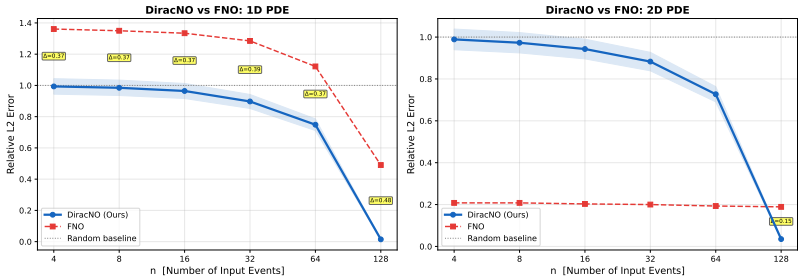


Figure 11: **DIRAC-NO versus FNO.** FNO first converts sparse events into a surrogate grid field. The comparison becomes most diagnostic when observations are sparse: interpolation error enters before operator learning, whereas DIRAC-NO keeps events atomic until query time.

Scaling with observation budget. Figure 12 summarizes how quickly different models deteriorate as observed events are removed. We do not interpret the trend as a universal scaling law. Its role is narrower: it visualizes the breakdown regime. Interpolation-first methods must construct a field before learning begins, so their error reflects both the surrogate-field approximation and the learned operator. Measure-native aggregation postpones field construction until a query is made, which shifts where the approximation enters.

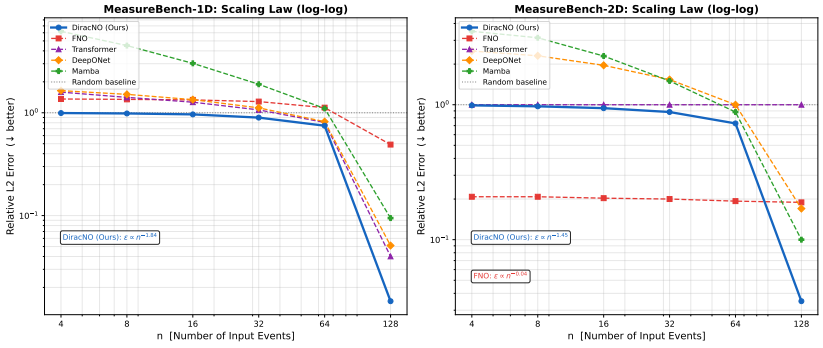


Figure 12: **Scaling with observation budget.** Error grows as observed events are removed. The diagnostic question is where the approximation enters: before learning through sparse-to-grid conversion, or at query time through measure aggregation.

Empirical transition under extreme sparsity. Figure 13 visualizes the transition from a representation-limited regime to an information-limited regime. This is not a mathematically sharp phase transition. It is an empirical summary of the breakdown pattern. When enough events remain, the choice of input representation matters: keeping events atomic delays collapse. Once too few events remain, the missing information dominates the task and all methods degrade.

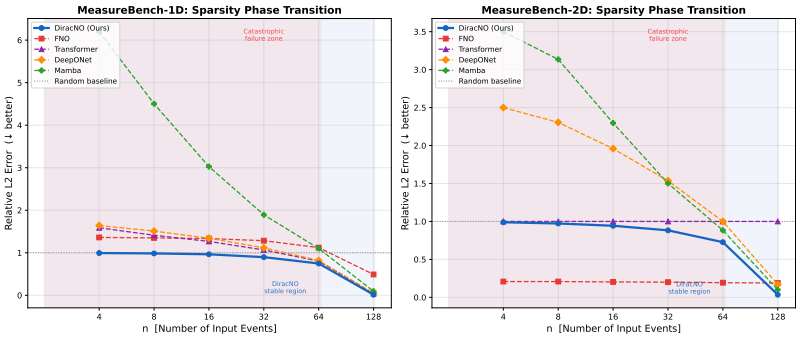


Figure 13: **Empirical transition under extreme sparsity.** The performance gap widens when sparse-to-grid construction becomes unreliable. Under more extreme sparsity, all methods enter an information-limited regime.

Additional ERA5 visualizations. Figure 14 complements the ERA5 sparsity curves with qualitative reconstructions. The purpose is not to claim that measure-native learning removes uncertainty under sparse sensing. It does not. Rather, the figure shows how the representation choice changes the failure mode: some baselines become unstable or overly smooth as station observations are removed, while DIRAC-NO more often preserves coarse temperature structure before the task becomes information-starved.

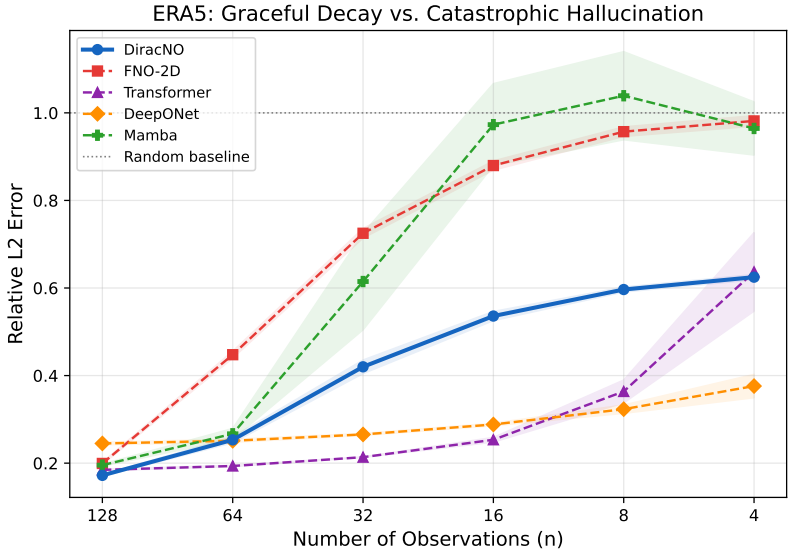


Figure 14: **ERA5 qualitative results under increasing sparsity.** As station observations are removed, baselines tend to become unstable or overly smooth. DIRAC-NO more reliably preserves coarse spatial structure before the reconstruction problem becomes information-limited.

C Mechanism Diagnostics and Alternative Explanations

This section asks whether the empirical gains can be explained away by simpler factors. The goal is not component bookkeeping. It is to test the structural claims behind the method: sparse events should remain atomic, aggregation should not renormalize source contributions, event geometry and event values should be separated, and fields should be constructed only at query time.

We organize the diagnostics around the most direct alternative explanations. Is the gain only model size? Is the FNO gap caused by a weak interpolation front end? Which parts of the measure-to-function layer are responsible for the effect? Does the trained model behave more additively, as the

theory predicts? And does the learned front end recover a source-response structure that resembles the Green-kernel interpretation?

C.1 Is the Gain Only Model Size?

A first alternative explanation is that DIRAC-NO wins simply because it has more parameters. We therefore compare against a substantially smaller version of the model. A Small DIRAC-NO with only 6,929 parameters ($d_{\text{model}} = 32$, $H = 4$) achieves RelL2 $= 0.199$ on MeasureBench-1D at $n = 128$, comparable to CNP with 91K parameters and RelL2 $= 0.170$. On MeasureBench-2D, the separation is larger: Small DIRAC-NO reaches RelL2 $= 0.458$, while CNP collapses to approximately 1.0.

This comparison does not make the models identical in every architectural respect. It rules out the simplest explanation. The observed advantage is not only a parameter-count effect. A small measure-native model still benefits from the same representation choice: the sparse input remains an event measure until the field is queried.

C.2 Is the FNO Gap Caused by a Weak Interpolation Front End?

A second alternative explanation is that FNO underperforms because we used a weak sparse-to-grid front end. If this were the main issue, stronger interpolation should close the gap. We therefore retrain FNO on MeasureBench-1D with three front ends: nearest binning, Gaussian smoothing, and Epanechnikov-kernel density estimation.

Table 15 shows that changing the interpolation rule does not close the gap with DIRAC-NO. The conclusion is not that one interpolation rule is uniquely bad. The issue is upstream: once sparse events are converted into a surrogate grid field, the object on which the operator learns has already been changed.

Table 15: **Effect of interpolation front end on FNO.** Relative L^2 error ($\downarrow$) on MeasureBench-1D, mean over three seeds. Stronger interpolation front ends do not close the gap with DIRAC-NO, supporting the claim that the issue is the interpolation-first representation rather than a single poor interpolation rule.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	**0.015**	**0.748**	**0.897**	**0.964**	**0.985**	**0.994**
FNO-Nearest	0.582	0.913	1.046	1.112	1.144	1.150
FNO-Gaussian	0.762	0.970	1.019	1.033	1.075	1.125
FNO-KDE	0.867	0.975	1.054	1.080	1.120	1.091

C.3 Which Design Choices Matter?

The ablations in Table 16 and Figure 15 test which parts of the measure-to-function layer are necessary. Value encoding, adaptive bandwidth, and multi-head structure all contribute. The largest failure mode is mean aggregation, which replaces source-wise summation with event-count normalization. This is the structural ablation that matters most for the paper: once sparse events are treated as the native input, aggregation is part of the operator design rather than a secondary implementation detail.

Table 16: **Extended ablation on MeasureBench-2D.** Relative L^2 error ($\downarrow$). Mean aggregation is the most damaging change because it renormalizes event contributions instead of summing source responses.

Variant	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
Full (Ours)	**0.035**	**0.727**	**0.883**	**0.943**	**0.973**	**0.989**
Fixed Bandwidth	0.072	0.788	0.938	1.007	1.040	1.044
Single Head	0.095	0.736	0.891	0.949	0.978	0.994
No Value Encoding	0.385	0.785	0.915	0.968	0.999	1.016
Mean Aggregation	0.427	1.061	1.837	3.164	5.076	6.979

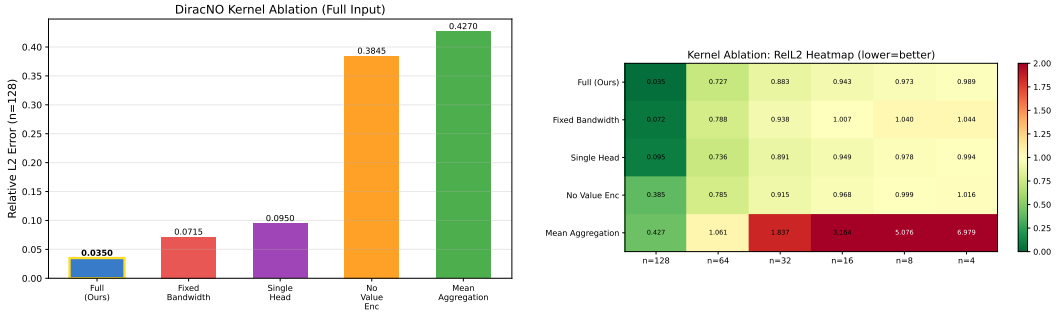


Figure 15: **Ablation diagnostics on MeasureBench-2D.** Left: full-input ablation at $n = 128$. Right: Relative L^2 error across sparsity levels. The mean-aggregation variant fails most sharply, consistent with the source-additivity argument.

C.4 Family-Specific Ablations

The harder PDE families provide a second structural diagnostic. If the gains came from a single accidental implementation choice, the ablation ranking should be unstable once transport or anisotropy is introduced. Instead, the same pattern persists.

On the advective family, the full model attains RelL2 $= 0.091$ at $n = 128$, compared with 1.015 without adaptive bandwidth, 1.309 without the spectral mixing block, 1.044 without gating, 1.213 without iterative refinement, 1.080 without value encoding, and 1.118 under mean aggregation. On the anisotropic family, the full model reaches 0.056, compared with 1.177 without adaptivity, 1.003 without the spectral block, 1.384 without gating, 1.420 without iterative refinement, 1.074 without value encoding, and 1.110 under mean aggregation.

These ablations reinforce the same conclusion as the default benchmark. The performance does not come from one isolated module. It comes from the coordinated measure-to-function design: adaptive local response, event-value encoding, direct aggregation, and nonlinear refinement only after query-time field construction.

C.5 Superposition Fidelity Beyond the Main Table

The superposition-fidelity experiment connects the structural argument to the behavior of trained models. It does not claim that the full network is exactly linear end to end. Instead, it asks whether the learned response remains closer to the additive law when events are combined through a direct measure-native summation rule.

For each test sample, we partition the event set into two disjoint subsets μ_1 and μ_2 and compute

$$\varepsilon_{\mathrm{sup}} = \frac{\|\mathcal{F}(\mu_1 + \mu_2) - \mathcal{F}(\mu_1) - \mathcal{F}(\mu_2)\|_2}{\|\mathcal{F}(\mu_1 + \mu_2)\|_2 + \epsilon}.$$

A small value means that the model responds to the union of sparse events much like the sum of its responses to the two parts. In this sense, $\varepsilon_{\mathrm{sup}}$ is the empirical counterpart of Proposition 1.

The remaining nonzero error is expected. Proposition 1 applies to the aggregation rule itself. Once gated residual updates, spectral mixing, and the decoder are added, the full network is no longer exactly linear in the input measure. The experiment shows instead that departures from additivity remain much smaller when sparse events enter through direct measure-native summation.

C.6 Green-Response Recovery

The Green-kernel view in the main text is structural: it explains why query-time summation is aligned with source-driven operators. A natural diagnostic is whether the learned front end recovers a response resembling the underlying single-source Green curve.

We isolate single-event test cases from the 1D benchmark and compare the model's recovered single-event response curve with the analytic Green response under the corresponding sample-specific PDE parameters. This is a diagnostic rather than a theorem about exact kernel identification.

Table 17: **Superposition fidelity across sparsity levels on MeasureBench-1D.** Lower ε_{sup} indicates closer agreement with source-wise additive response.

Method	$n = 128$	$n = 64$	$n = 32$	$n = 16$	$n = 8$	$n = 4$
DIRAC-NO	**0.083**	**0.088**	**0.068**	**0.068**	**0.083**	**0.127**
FNO	1.017	0.682	0.477	0.407	0.422	0.484
Transformer	0.656	0.723	0.817	1.050	1.446	1.747
DeepONet	0.627	0.672	0.693	0.717	0.745	0.754
Mamba	1.061	1.117	1.202	1.249	1.287	1.198

Figure 16 shows representative examples. Across six single-event probes, the recovered response achieves mean RelL2 $= 0.109$ and mean correlation $= 0.992$ relative to the analytic Green curve. This makes the Green-kernel view visible in the trained model rather than only in the architecture diagram.

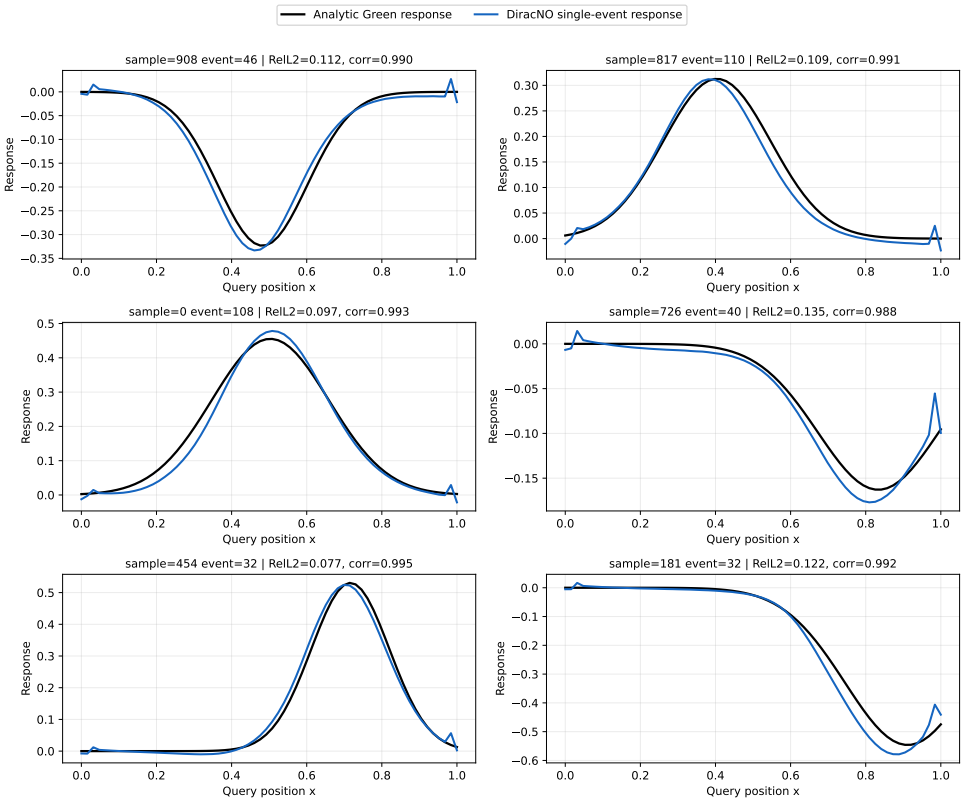


Figure 16: **Single-event Green-response recovery.** Each panel compares the analytic Green response for a held-out single event with the response recovered by the trained DIRAC-NO model.

C.7 White-Box Kernel Diagnostics

The main text argues that DIRAC-NO postpones breakdown under decreasing observation budgets. Figure 17 provides a white-box view of the learned front end by visualizing kernel scales and representative predictions under varying sparsity.

The left panel plots learned bandwidths for several kernel heads. The heads do not collapse to one common scale. They separate into different effective spatial ranges: some remain local and respond strongly to nearby events, while others operate at broader scales and stabilize reconstruction when

local evidence is weak. The right panel shows representative predictions under different sparsity levels. The model does not fail abruptly as observations are removed; errors mainly appear as gradual loss of fine structure rather than unstable oscillation or complete collapse.

This figure is not a proof. Its role is diagnostic: it makes the architectural story visible by showing that measure-native aggregation and multi-scale kernels leave a coherent footprint in the trained model.

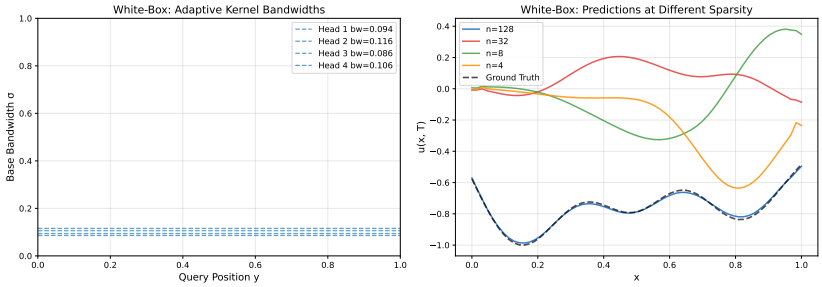


Figure 17: **White-box kernel diagnostics.** Left: learned kernel bandwidths across query locations, showing head specialization to different effective spatial scales. Right: representative predictions under varying sparsity, illustrating gradual loss of fine structure as observations become scarce.

D Structural Claims: What Is Guaranteed and What Is Not

This section formalizes the structural claims used in the main text. The purpose is not to prove that DIRAC-NO is universally optimal. The purpose is narrower and tied to the representation argument of the paper. We show that query-time direct summation preserves source-wise additivity at the aggregation stage, that normalized aggregation generally does not, and that the measure-to-function front end contains the classical Green source-to-field formula as a special case. We then record benchmark-specific representability and stability statements for the smooth diffusion setting used in MeasureBench.

The distinction between the aggregation front end and the full model matters. The aggregation layer is written as A_θ. The full predictor, after gated updates, optional spectral mixing, and decoding, is written as F_θ. The additivity result below applies to A_θ, not to the nonlinear full model.

D.1 Additivity of Query-Time Summation

We restate the measure-to-function layer from Eq. (2). Given an event list

$$E = \{(a_i, x_i, t_i)\}_{i=1}^{n},$$

the aggregation front end computes

$$A_\theta(E)(y) = \sum_{i=1}^{n} K_\theta(y, x_i, t_i)\, V_\theta(a_i, x_i, t_i).$$

Define the per-event contribution

$$\psi_\theta(y; a, x, t) := K_\theta(y, x, t)\, V_\theta(a, x, t).$$

Then

$$A_\theta(E)(y) = \sum_{(a,x,t)\in E} \psi_\theta(y; a, x, t).$$

Let

$$E_1 = \{(a_i, x_i, t_i)\}_{i=1}^{m}, \qquad E_2 = \{(b_j, \tilde{x}_j, \tilde{t}_j)\}_{j=1}^{n}$$

be two event lists, and let $E_1 \sqcup E_2$ denote their concatenation as an event multiset. Then

$$\begin{aligned} A_\theta(E_1 \sqcup E_2)(y) &= \sum_{i=1}^{m} \psi_\theta(y; a_i, x_i, t_i) + \sum_{j=1}^{n} \psi_\theta(y; b_j, \tilde{x}_j, \tilde{t}_j) \\ &= A_\theta(E_1)(y) + A_\theta(E_2)(y). \end{aligned}$$

Thus direct summation is exactly additive over concatenated event inputs at the aggregation stage.

What this does and does not claim. This is the structural property used by the architecture: adding events adds their encoded contributions without renormalizing the events already present. It is not a claim that the full model F_θ is linear end to end. After the aggregation stage, gated residual updates, spectral mixing, and the decoder may introduce nonlinear corrections.

If one requires strict linearity with respect to signed Radon-measure addition, then the value pathway must also be linear in the source amplitude, for example $V_\theta(a, x, t) = a\,\phi_\theta(x, t)$. The implemented model uses a learned value encoder. The exact claim we rely on is therefore source-wise additivity of the direct aggregation rule, not global linearity of the full network.

D.2 Why Normalized Aggregation Breaks Source Additivity

The previous property depends on direct summation. Normalized aggregation rules generally change the contribution of existing events when new events are inserted. This is why replacing direct summation with mean aggregation is a structural ablation rather than a cosmetic implementation change.

Softmax attention. For a query q and key-value pairs $\{(k_i, v_i)\}_{i=1}^n$, standard attention computes

$$\mathrm{Attn}(q, \{(k_i, v_i)\}_{i=1}^n) = \sum_{i=1}^n \frac{\exp(q^\top k_i/\sqrt{d})}{\sum_{j=1}^n \exp(q^\top k_j/\sqrt{d})}\, v_i.$$

The denominator depends on the full input set. Inserting, deleting, or duplicating one event changes the weights assigned to all other events.

A one-event counterexample is enough. For a single event with score s and value v,

$$\mathrm{Attn}(q, \{(k, v)\}) = v.$$

If the same event is duplicated, then

$$\mathrm{Attn}(q, \{(k, v), (k, v)\}) = \frac{e^s}{e^s + e^s} v + \frac{e^s}{e^s + e^s} v = v.$$

A source-additive response would assign the duplicated input twice the contribution, $2v$. Softmax attention therefore implements a different response law from source summation.

Mean aggregation. For mean aggregation,

$$\overline{A}_\theta(E)(y) = \frac{1}{|E|} \sum_{(a,x,t)\in E} \psi_\theta(y; a, x, t).$$

Let E_1 contain m events and E_2 contain n events. Then

$$\overline{A}_\theta(E_1 \sqcup E_2)(y) = \frac{1}{m+n} \left[\sum_{(a,x,t)\in E_1} \psi_\theta(y; a, x, t) + \sum_{(b,\tilde{x},\tilde{t})\in E_2} \psi_\theta(y; b, \tilde{x}, \tilde{t}) \right],$$

whereas

$$\overline{A}_\theta(E_1)(y) + \overline{A}_\theta(E_2)(y) = \frac{1}{m} \sum_{(a,x,t)\in E_1} \psi_\theta(y; a, x, t) + \frac{1}{n} \sum_{(b,\tilde{x},\tilde{t})\in E_2} \psi_\theta(y; b, \tilde{x}, \tilde{t}).$$

These expressions coincide only in special degenerate cases. Mean aggregation therefore does not preserve source-additive response. This explains the ablation pattern in Appendix C.3: when direct summation is replaced by mean aggregation, the model no longer implements the source-response rule that motivates the measure-native front end.

D.3 Green-Kernel Realization as a Special Case

The Green-kernel view in the main text can be stated as a compatibility result. It explains why query-time summation is a natural front end for source-driven systems.

Proposition 2 (Green source-to-field realization)**.** *Suppose a linear source-driven PDE admits a Green's function G such that, for an atomic source measure*

$$\mu = \sum_{i=1}^{n} a_i \, \delta_{(x_i,t_i)},$$

the solution at query location y satisfies

$$u(y) = \int_{\Omega} G(y,x,t) \, d\mu(x,t) = \sum_{i=1}^{n} a_i \, G(y,x_i,t_i).$$

If the measure-to-function layer is chosen so that

$$K_\theta(y,x_i,t_i) = G(y,x_i,t_i), \qquad V_\theta(a_i,x_i,t_i) = a_i,$$

then A_θ exactly recovers the discrete Green representation:

$$A_\theta(\mu)(y) = u(y).$$

Proof. Substituting the stated choices into the aggregation rule gives

$$A_\theta(\mu)(y) = \sum_{i=1}^{n} K_\theta(y,x_i,t_i) \, V_\theta(a_i,x_i,t_i) = \sum_{i=1}^{n} G(y,x_i,t_i) \, a_i = u(y).$$

□

This proposition is not a universal approximation theorem. It shows that the architecture is compatible with the classical source-to-field formula: a Green response is a special case of query-time measure aggregation.

D.4 Benchmark-Specific Exact Representability

The next statement is deliberately benchmark-specific. It applies to the analytic diffusion benchmarks used to isolate the representation question. It should not be read as a claim that the finite implementation exactly represents arbitrary PDE solution operators.

Let $\mathcal{F}_{\mathrm{true}}$ denote the data-generating operator induced by the analytic diffusion Green kernel. Let $\mathcal{H}_{\mathrm{DIRAC}}$ denote the idealized measure-to-function class underlying DIRAC-NO, with downstream nonlinear blocks omitted or set to the identity and with sufficient capacity to represent the benchmark Green kernel.

Theorem 1 (Benchmark-specific exact representability)**.** *For the synthetic point-source diffusion benchmarks considered in this paper, the ground-truth operator satisfies*

$$\mathcal{F}_{\mathrm{true}} \in \mathcal{H}_{\mathrm{DIRAC}}.$$

Equivalently, there exists a measure-to-function parameter setting $\theta^\star$ such that the resulting predictor exactly reproduces the benchmark operator on every admissible sparse event input in the benchmark class.

Proof. For the benchmark class considered here, the target field is generated by a diffusion Green kernel. For an atomic source input

$$E = \{(a_i,x_i,t_i)\}_{i=1}^{n},$$

the benchmark operator takes the form

$$\mathcal{F}_{\mathrm{true}}(E)(y) = \sum_{i=1}^{n} a_i \, G(y,x_i,t_i),$$

where G is the corresponding analytic Green kernel.

Choose

$$K_{\theta^\star}(y,x_i,t_i) = G(y,x_i,t_i), \qquad V_{\theta^\star}(a_i,x_i,t_i) = a_i.$$

Then

$$A_{\theta^\star}(E)(y) = \sum_{i=1}^{n} K_{\theta^\star}(y, x_i, t_i)\, V_{\theta^\star}(a_i, x_i, t_i) = \sum_{i=1}^{n} G(y, x_i, t_i)\, a_i = \mathcal{F}_{\mathrm{true}}(E)(y).$$

Thus the synthetic diffusion operator lies inside the idealized measure-to-function hypothesis class. □

The theorem is benchmark-specific by design. Its role is not to claim universal expressivity. It shows that, in the controlled diffusion benchmark used to test the representation question, the proposed front end is not mismatched to the target operator at the hypothesis-class level.

D.5 Wasserstein Stability for Marked Event Measures

We finally record a stability statement for the smooth diffusion benchmark. The claim is again local to the benchmark regime. It formalizes the intuition that, when the single-event response is smooth in location, time, and amplitude, small changes in the marked event measure produce controlled changes in the output field.

Because event amplitudes are part of the observed event tuple, we work with marked empirical event distributions. For a fixed event budget n, define

$$\nu_E = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}, \qquad z_i = (x_i, t_i, a_i).$$

In multiple spatial dimensions, $z_i = (\mathbf{x}_i, t_i, a_i)$. Let the ground metric be

$$d(z, z') = \|x - x'\|_1 + |t - t'| + |a - a'|$$

in one dimension, with the replacement $\|\mathbf{x} - \mathbf{x}'\|_1$ in multiple spatial dimensions. Let W_1^d denote the associated 1-Wasserstein distance between marked empirical distributions.

Theorem 2 (Benchmark-specific Wasserstein stability)**.** *For the smooth diffusion benchmark on the compact admissible parameter range, there exists a constant $C_{\mathrm{stab}} > 0$ such that, for any two fixed-budget event lists E and E',*

$$\|\mathcal{F}_{\mathrm{true}}(E) - \mathcal{F}_{\mathrm{true}}(E')\|_{L^\infty(\Omega_x)} \le n\, C_{\mathrm{stab}}\, W_1^d(\nu_E, \nu_{E'}).$$

Proof. For each query point y, define the single-event response function

$$\Phi_y(z) = a\, G(y, x, t), \qquad z = (x, t, a).$$

On the compact benchmark domain and admissible parameter range, the regularized diffusion response used in data generation is uniformly Lipschitz in the event variables. Hence there exists $C_{\mathrm{stab}} > 0$ such that

$$|\Phi_y(z) - \Phi_y(z')| \le C_{\mathrm{stab}}\, d(z, z')$$

for all admissible z, z' and all query points y.

The benchmark operator can be written as

$$\mathcal{F}_{\mathrm{true}}(E)(y) = \sum_{i=1}^{n} \Phi_y(z_i) = n \int \Phi_y(z)\, d\nu_E(z).$$

Let π be any coupling between ν_E and $\nu_{E'}$. Then

$$\begin{aligned}
|\mathcal{F}_{\mathrm{true}}(E)(y) - \mathcal{F}_{\mathrm{true}}(E')(y)| &= n \left| \int \Phi_y(z)\, d\nu_E(z) - \int \Phi_y(z')\, d\nu_{E'}(z') \right| \\
&= n \left| \int \big(\Phi_y(z) - \Phi_y(z')\big)\, d\pi(z, z') \right| \\
&\le n \int |\Phi_y(z) - \Phi_y(z')|\, d\pi(z, z') \\
&\le n\, C_{\mathrm{stab}} \int d(z, z')\, d\pi(z, z').
\end{aligned}$$

Taking the infimum over all couplings π gives

$$|\mathcal{F}_{\mathrm{true}}(E)(y) - \mathcal{F}_{\mathrm{true}}(E')(y)| \le n\, C_{\mathrm{stab}}\, W_1^d(\nu_E, \nu_{E'}).$$

Taking the supremum over $y \in \Omega_x$ proves the result. □

This stability statement is intentionally local: fixed event budget, bounded amplitudes, compact spatiotemporal domain, and the regularized smooth diffusion response used in data generation. It should not be read as a universal stability theorem for arbitrary signed measures, discontinuous responses, or singular hyperbolic kernels.

D.6 Why Interpolation Can Distort Additivity

The paper's central claim is not that every interpolation rule is mathematically nonlinear. Some rasterization or kernel-smoothing operators can be linear. The claim is more basic: interpolation changes the object on which the operator is trained before learning begins.

Let $\mathcal{I}$ denote an interpolation, rasterization, or sparse-to-grid operator that maps an event input to a dense surrogate field. An interpolation-first pipeline has the form

$$\hat{u} = \mathcal{N}\big(\mathcal{I}(E)\big),$$

where $\mathcal{N}$ is a neural operator acting on functions.

If $\mathcal{I}$ is nonlinear, normalized, or dependent on local sample density, then in general

$$\mathcal{I}(E_1 \sqcup E_2) \neq \mathcal{I}(E_1) + \mathcal{I}(E_2).$$

In this case, source-wise additivity has already been altered before the neural operator is trained. The effect is especially direct for normalized or density-adaptive front ends: adding an event can change the weights, bandwidths, or normalization factors assigned to events that were already present.

Even when $\mathcal{I}$ is linear, the representation has still changed. The operator no longer sees the atomic event measure. It sees a constructed field whose values depend on the chosen grid, kernel, bandwidth, and discretization rule. A linear interpolation may preserve algebraic additivity of the surrogate field, but it does not preserve the native support of the observations or the fact that no values were observed between events.

This is the formal version of the paper's representation claim. The comparison is not only between neural architectures. It is between two ways of posing the input space: one in which the sparse observation is first converted into a function, and one in which the input remains an event measure until a query is made.

D.7 What the Theory Establishes

The results above establish a specific alignment between DIRAC-NO and the source-driven settings studied in the paper. Query-time summation preserves source-wise additivity at the aggregation stage. Normalized attention and mean aggregation generally do not. Classical Green source-to-field maps are contained as special cases of measure-to-function aggregation. For the smooth diffusion benchmarks, the true operator is exactly representable by the idealized front end and stable to perturbations of marked event inputs.

These statements do not claim universal superiority. They identify the structural regime in which the representation argument applies: sparse event-to-field problems where localized inputs act, at least approximately, as sources of a continuous response field.

E Boundary Diagnostics: Where the Claim Should Weaken

The main text identifies an operating regime for measure-native operator learning: sparse event-native input together with an approximately source-driven response law. This section tests the boundary of that regime. The goal is not to weaken the argument. It is to make the inductive bias falsifiable.

The prediction is simple. If sparse observations behave like localized sources of a field response, keeping them as measures should help. If event-native input remains but the source-response structure

weakens, the advantage should weaken. If the response geometry changes from smooth diffusion to wavefront propagation, then the current Gaussian-style kernel family should no longer be the right front end. Boundary cases are therefore part of the evidence trail: they show where the representation argument stops.

E.1 When Should Measure-Native Learning Help?

Table 18 summarizes the operating regime suggested by the experiments. MeasureBench-1D and MeasureBench-2D are the cleanest positive cases: the inputs are sparse events, the source-response structure is strong, and the advantage is large. ERA5 and NYC Taxi weaken the idealized PDE structure while retaining sparse event-to-field form, so the expected advantage is more moderate. Financial streams and wave-dominated settings mark diagnostic boundaries: the input may remain event-native, but either the source-response structure is weak or the kernel geometry is mismatched.

This table is not a defensive scope disclaimer. It is a prediction map. Measure-native learning should help when sparse observations behave like localized sources of a field response; it should be limited when sparsity appears without that structure.

Table 18: **When should measure-native learning help?** The advantage is predicted to be strongest when event-native input and source-response structure coincide. Boundary cases test where this prediction should weaken.

Regime	Event-native input	Source-response structure	Predicted advantage	Observed pattern
MeasureBench-1D	Present	Strong	Strong	$33\times$ lower RelL2 vs. FNO
MeasureBench-2D	Present	Strong	Strong	$29\times$ lower RelL2 vs. FNO
ERA5 Weather	Present	Approximate	Moderate	Lowest error at $n = 128$
NYC Taxi	Present	Approximate	Moderate	Best or near-best at high/moderate budgets
Financial streams	Present	Weak	Limited	No uniform advantage
OOD-Param (2D)	Present	Strong	Moderate	Best among all methods
OOD-Density (2D)	Present	Strong	Moderate	Best among all methods
OOD-Intensity (2D)	Present	Value-shifted	Limited	FNO lower numerical error
Hyperbolic PDEs	Present	Kernel mismatch	Limited	Wavefront mismatch
Seismic waves	Present	Kernel mismatch	Limited	Competitive but not dominant

E.2 Financial Streams: Event-Native Input Is Not Enough

The financial benchmarks are negative controls. They retain sparse event-native input, but they do not exhibit a clear additive source-to-field mechanism. They therefore test whether the representation advantage of DIRAC-NO is triggered by sparsity alone. The answer is no, and this is informative.

The empirical pattern is consistent with this diagnostic role. No single method dominates uniformly, and in several cases all methods cluster near RelL2 ≈ 1.0. This does not contradict the main result. It separates two conditions that coincide in the controlled PDE benchmarks but need not coincide in general: sparse event input and additive source-response structure.

Table 19 reports a compact summary at $n = 128$ together with the area under the sparsity-error curve. Full sparsity curves are shown in Figure 18.

Table 19: **Financial event-stream benchmarks.** RelL2 at $n = 128$ and AUC ($\downarrow$). These datasets are event-native but not clearly source-driven. Entries marked "—" indicate that Mamba-1D was not trained because fewer than 2000 samples were available.

Method	BTC n=128	BTC AUC	ETH n=128	ETH AUC	OHLC n=128	OHLC AUC	Tick n=128	Tick AUC
DIRAC-NO	1.0266	129.6	1.0000	124.0	**0.9978**	**124.0**	**1.0000**	**124.0**
FNO-1D	**0.9418**	**122.5**	1.0094	125.5	1.0007	125.0	1.0094	125.5
Transformer	1.0127	136.1	1.0032	124.4	1.0001	**124.0**	1.0032	124.4
DeepONet-1D	0.9949	146.3	1.0031	124.4	1.0006	124.4	1.0031	124.4
Mamba-1D	—	—	**0.9149**	**117.8**	—	—	1.0012	124.3

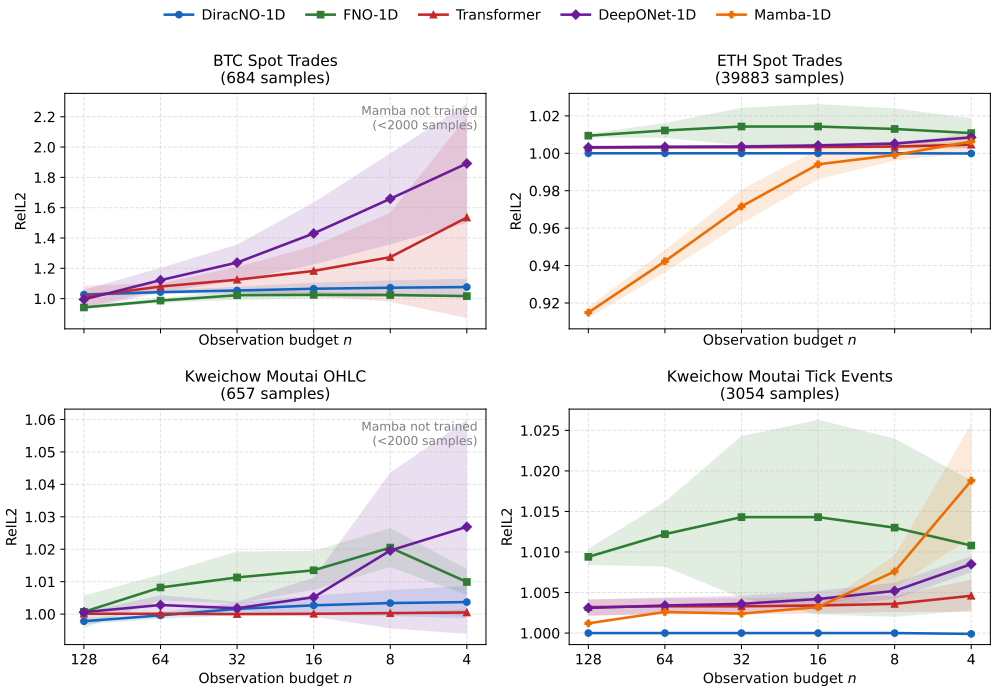


Figure 18: **Financial streams as negative controls.** These datasets retain sparse event-native input but lack a clear additive source-to-field mechanism. The weak and non-uniform advantage shows that event-native input alone is not sufficient to trigger the measure-native benefit.

E.3 Cross-Regime Summary

Figure 19 places the benchmark families side by side. Its purpose is not to claim a universal ranking across all domains. It is to show where the measure-native advantage is strongest and where it weakens.

The comparison reinforces the operating-regime story. In the synthetic PDE benchmarks, sparse localized events and additive source structure are both explicit, so DIRAC-NO has a clear advantage. In ERA5 and NYC Taxi, observations remain sparse and irregular, but the target structure is less cleanly source-driven, so the advantage becomes more moderate. In the financial benchmarks, event-native input remains but the additive source-to-field interpretation largely disappears; the advantage weakens accordingly.

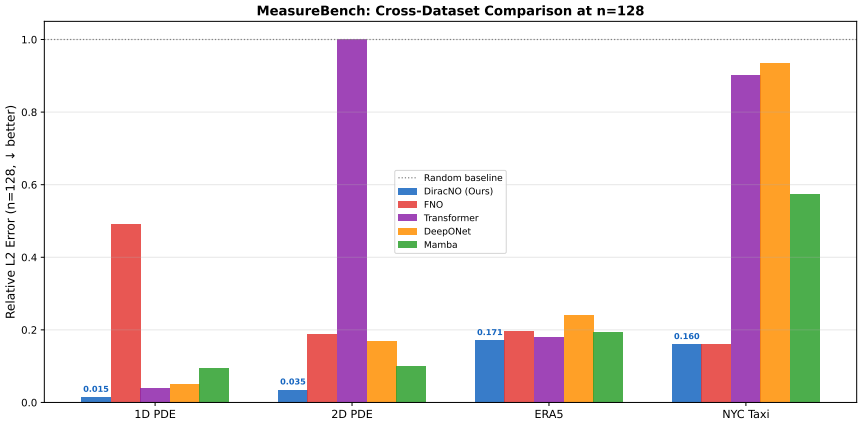


Figure 19: **Cross-regime comparison.** The measure-native advantage is clearest in sparse event-to-field settings with stronger source-response structure, and weaker when event-native input appears without that structure.

E.4 Hyperbolic PDEs: Gaussian Kernels Are Not Wavefront Kernels

The main text emphasizes that DIRAC-NO is most naturally aligned with additive, diffusion-dominated operators. Hyperbolic systems make the boundary of that alignment clear. Consider a linear wave equation of the form

$$u_{tt} - c^2 \Delta u = \mu.$$

Its Green's function differs qualitatively from the Green's functions associated with elliptic or parabolic problems. For diffusion-type equations, the response spreads smoothly and decays away from the source. For hyperbolic equations, the response propagates along characteristic cones or wavefronts [John, 1982, Evans, 2010, Stakgold and Holst, 2011]. The singular support of the solution is therefore concentrated on a lower-dimensional geometric structure rather than on a smooth spatial envelope.

This difference matters directly for the current kernel family. In DIRAC-NO, the measure-to-function layer is built from adaptive Gaussian-style kernels,

$$K_\theta(y, x, t) = \sum_{h=1}^{H} \alpha_h \exp\left(-\frac{\|y - x\|^2}{2\sigma_h(y)^2} \right),$$

with positive bandwidths $\sigma_h(y) > 0$. Such kernels are well suited to smooth, spatially decaying propagators, but they are less natural for sharply localized fronts, cone-supported singularities, or strongly directional transport. In practice, the mismatch can appear as oversmoothing of wavefronts, spatial smearing outside the physically relevant propagation region, or a need for many narrow heads to approximate a structure that is fundamentally directional rather than isotropic.

This limitation is constructive. It does not suggest that measure-native learning is misguided; it shows that measure-native aggregation still needs the right response geometry. Hyperbolic PDEs therefore identify a natural next step: wavefront-aware kernels, anisotropic local metrics, or hybrid designs that combine event-measure aggregation with physics-aware propagators.

E.5 Seismic Boundary Case

We also evaluate the method on a seismic benchmark, which should be read as a boundary case rather than as a new domain of dominance. Unlike the diffusion-dominated source-driven settings emphasized in the main paper, seismic wave propagation is hyperbolic and shaped by oscillatory fronts and finite-speed transport. In this regime, the current Gaussian-like measure-native kernel is not especially well matched to the underlying propagator.

The empirical pattern is consistent with this interpretation. On Task A, DIRAC-NO remains competitive but does not dominate the strongest baselines; DeepONet attains the lowest error, while

DIRAC-NO, Transformer, and Mamba form a closely clustered second tier. Under the harder Task C OOD shifts, all methods degrade severely, and the advantage of the measure-native front end does not persist. These results sharpen the claim of the paper: measure-native aggregation alone is not sufficient for wave-dominated operators, and richer wavefront-aware kernels are likely needed in this regime.

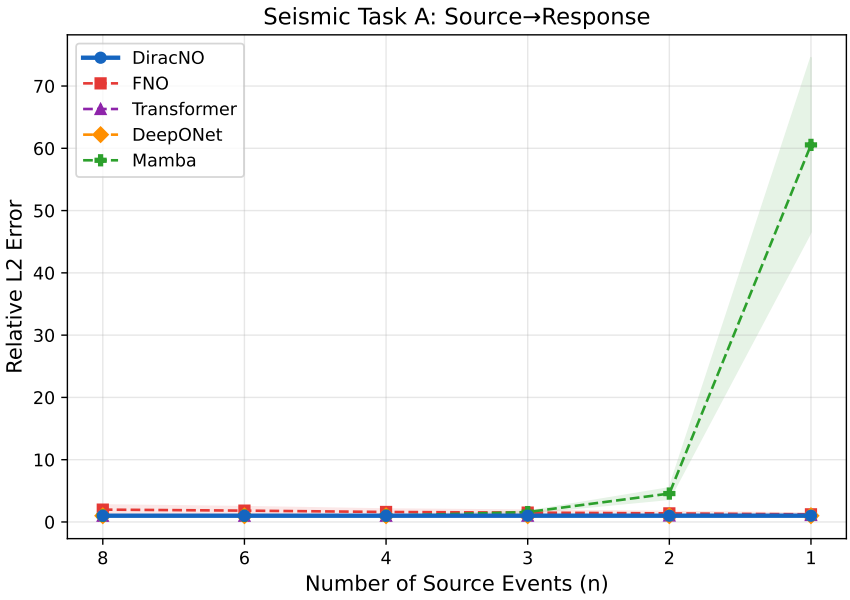


Figure 20: **Seismic Task A as a boundary case.** DIRAC-NO remains competitive but does not dominate the strongest baselines, consistent with the mismatch between Gaussian-like measure-native kernels and wave-dominated propagators.

E.6 What the Boundary Cases Establish

The boundary cases separate three conditions that are easy to conflate. Sparse event input alone is not enough. Source-response structure matters. Kernel geometry also matters. Financial streams weaken the source-response assumption; intensity-shift OOD stresses the value pathway; hyperbolic and seismic settings change the response geometry from smooth spatial decay to wavefront propagation.

The conclusion is therefore not that DIRAC-NO is a universal sparse-data method. It is that the representation advantage appears where the paper predicts it should appear: sparse event-to-field problems in which localized observations act approximately as sources of a continuous response field.

F Efficiency and Practical Frontier

The measure-native formulation changes where computation enters the pipeline. Interpolation-first methods spend effort constructing a dense surrogate field before learning begins. DIRAC-NO instead spends effort at query time, when each output location aggregates responses from the observed events. This section makes that tradeoff explicit. The goal is not to claim a new asymptotic result, but to identify when the direct query-event interaction is practical, where it becomes the limiting factor, and what kinds of acceleration would matter outside the sparse regime studied in this paper.

F.1 Where the Query-Event Cost Enters

Suppose one sample contains n observed events and the operator is queried at Q output locations. For each query-event pair, the measure-to-function layer evaluates a kernel weight and combines it with a latent event representation. Treating the number of heads H and the latent width d_{model} as

architectural constants, the leading-order cost is

$$\mathrm{Cost}_{\mathrm{M2F}} = \mathcal{O}(nQ).$$

More explicitly, the kernel evaluation contributes

$$\mathcal{O}(nQH),$$

and latent accumulation contributes

$$\mathcal{O}(nQd_{\mathrm{model}}).$$

The cost is therefore linear in both the number of observed events and the number of query points.

In the 2D settings, we optionally apply an FNO-style spectral block after the measure-to-function layer. If the latent field is defined on a regular grid with $Q = S_x S_y$ locations and width W, the FFT-based mixing cost is approximately

$$\mathrm{Cost}_{\mathrm{spec}} = \mathcal{O}(WQ \log Q),$$

up to constants determined by the number of retained Fourier modes [Li et al., 2021]. Once the latent field has been formed, this term is independent of the input event count.

The benchmark regime studied in this paper is intentionally sparse:

$$n \leq 128, \qquad Q \leq 2048.$$

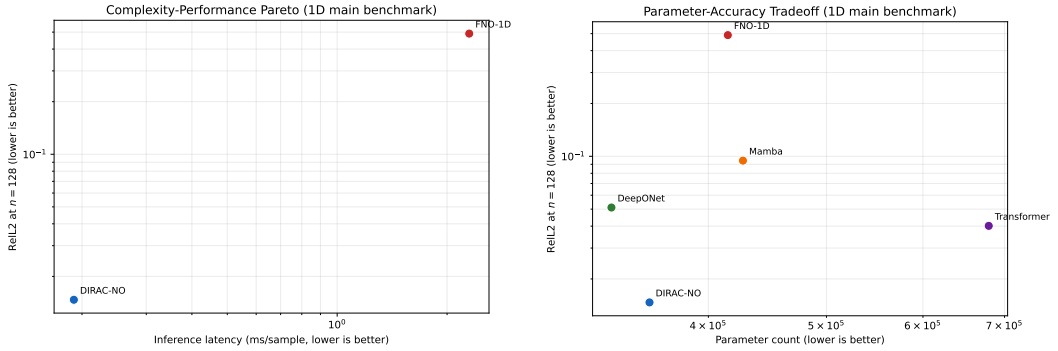
In this regime, direct $\mathcal{O}(nQ)$ aggregation is computationally manageable. The relevant tradeoff is not feasibility versus infeasibility. It is where approximation enters the pipeline. Interpolation-first methods approximate the input before learning by constructing a dense surrogate field. DIRAC-NO keeps the input atomic and pays for query-event interactions only when field values are requested.

F.2 Complexity-Performance Pareto

The practical question is not whether DIRAC-NO contains an explicit query-event interaction. It does. The question is whether, in the sparse regime studied here, that interaction gives a favorable tradeoff between computational cost and achieved accuracy.

Figure 21 makes this comparison explicit on the main 1D benchmark at $n = 128$. The key point is not only that DIRAC-NO is more accurate. It remains on a competitive practical frontier while delivering that accuracy. In the latency comparison, DIRAC-NO achieves RelL2 $= 0.0147$ with inference latency 0.19 ms/sample, whereas FNO-1D attains RelL2 $= 0.4904$ with latency 2.30 ms/sample. In the parameter-accuracy comparison, DIRAC-NO achieves the best dense-regime accuracy with $357{,}897$ parameters, remaining within the same scale as FNO-1D $(414{,}977)$, below Transformer $(679{,}425)$ and Mamba $(427{,}009)$, and only modestly above DeepONet $(333{,}057)$.

These plots are not a universal efficiency theorem. They support a narrower claim: in sparse event-to-field regimes, measure-native aggregation is not an accuracy gain purchased at impractical cost. It is a different placement of computation, and in the sparse regimes evaluated here, that placement is practically competitive.



(a) Latency versus accuracy.

(b) Parameter count versus accuracy.

Figure 21: **Complexity-performance Pareto plots.** On the 1D main benchmark at $n = 128$, DIRAC-NO combines the strongest accuracy with competitive latency and model size. The measure-native gain is not simply an artifact of excessive computation.

F.3 Model Complexity and Training Cost

The computational tradeoff is not whether DIRAC-NO has extra cost. It does. The question is whether the extra cost is visible, controlled, and aligned with the representation claim. Table 20 reports parameter count, training cost, and inference latency under standardized hardware and training settings.

Table 20: **Model complexity and runtime.** Training cost is reported on ERA5 using a single A100 GPU. Inference latency is measured on the 1D benchmark at $n = 128$, batch size 32, on an RTX 4070.

Method	Parameters	Early stop	Total time	Per epoch	Inference
DIRAC-NO	28.6 M	ep 150	∼104 min	∼41 s	0.19 ms
FNO-2D	9.5 M	ep 150	∼119 min	∼48 s	2.30 ms
Transformer	0.68 M	ep 375	∼75 min	∼12 s	—
DeepONet	0.33 M	ep 500	∼66 min	∼8 s	—
Mamba	0.34 M	ep 150	∼380 min	∼152 s	—
CNP	0.09 M	ep 500	∼8 min	∼1 s	0.03 ms

Reading the table. DIRAC-NO's higher per-epoch cost relative to Transformer and DeepONet reflects the $\mathcal{O}(nQ)$ query-event interaction in the measure-to-function layer. In the sparse regime considered here ($n \leq 128,\ Q \leq 2048$), this overhead remains modest relative to FNO. Mamba's substantially higher cost stems from its sequential state-space scan, which parallelizes less efficiently than attention- or convolution-based operations on GPU. FNO stops early because its validation error plateaus quickly under the selected training protocol.

F.4 Interpolation Cost versus Query-Time Aggregation

A standard way to apply grid-based neural operators to sparse observations is to first convert the event measure into a dense field through interpolation, binning, or kernel smoothing, and only then apply an operator learner such as FNO. Abstractly, this pipeline has the form

$$\hat{u} = \mathcal{N}\big(\mathcal{I}(\mu)\big),$$

where μ is the sparse input measure, $\mathcal{I}$ is an interpolation operator, and $\mathcal{N}$ is a neural operator acting on functions.

This strategy is useful because it allows architectures designed for function-valued inputs to be reused. At the same time, it changes the representation before learning begins. The original input is atomic:

$$\mu = \sum_{i=1}^{n} a_i\, \delta(x - x_i)\delta(t - t_i).$$

After interpolation, the model no longer acts on this event measure. It acts on a dense surrogate field defined on a chosen grid. The behavior of that surrogate field depends on interpolation bandwidth, grid resolution, and local event geometry.

In high-observation regimes, the difference may be small. In sparse regimes, it can be decisive. The event configuration contains information not only in the observed values, but also in where events are present and where they are absent. Once events are rasterized to a grid, part of this structure is smoothed or renormalized before the operator is trained. For source-driven systems, the issue is sharper: if sparse events are first mapped through a nonlinear or normalized interpolation rule, source-wise additivity may already be distorted before the neural operator is applied.

DIRAC-NO avoids this bottleneck by acting directly on the atomic input and constructing output values only at query time. The experiments should therefore be read not only as model comparisons, but as evidence that input representation can become a decisive part of sparse operator learning.

F.5 Acceleration Paths

The current implementation uses direct query-event aggregation. This is adequate for the sparse regime studied here. If the same formulation is applied to dense sensing scenarios with much larger

event counts, the $\mathcal{O}(nQ)$ term will eventually dominate. Two classical acceleration strategies are especially relevant.

Random feature approximations. For Gaussian-like kernels, random Fourier features [Rahimi and Recht, 2007] can replace pairwise kernel evaluation with a low-dimensional inner-product approximation. This can reduce the cost from $\mathcal{O}(nQ)$ to approximately

$$\mathcal{O}((n+Q)D)$$

for fixed feature dimension D.

Hierarchical kernel summation. Fast multipole methods [Greengard and Rokhlin, 1987] and related hierarchical schemes exploit smoothness and decay in kernel interactions to reduce the effective cost of all-pairs summation. These methods are especially attractive in the low-dimensional spatial settings considered here.

These accelerations lie beyond the scope of the current paper because the benchmark regime is intentionally sparse. Their role here is to show that the computational cost of measure-native aggregation is not conceptually tied to the naive implementation.

F.6 Constructive Extensions

The boundary cases in Appendix E suggest several natural extensions. Each extension changes the response geometry rather than the core representation principle: the input remains an event measure, but the kernel family becomes better matched to the target operator.

Wavefront-aware kernels. Instead of isotropic Gaussian envelopes, one could use kernels parameterized by propagation direction and wave speed, so that responses concentrate near characteristic fronts.

Oscillatory basis functions. Traveling-wave solutions are often better represented by oscillatory or wavelet-like bases than by purely Gaussian kernels. Incorporating such bases could make the aggregation layer more compatible with hyperbolic response structure.

Finite-speed causality constraints. Another possibility is to encode causality directly in the kernel by enforcing zero response outside the physically admissible propagation region. This would move part of the hyperbolic inductive bias from training into the architecture.

These directions lie beyond the current paper. Their importance is interpretive: they show that the limitations of DIRAC-NO are informative rather than arbitrary, and they indicate what a next-generation measure-native operator would need to change.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction accurately describe the paper's main contribution as measure-native operator learning for sparse event-to-field reconstruction, and they explicitly state the scope conditions under which the method performs best, including the boundary cases for intensity-shift OOD and financial event streams.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The paper includes a dedicated Limitations section discussing the mismatch of the current Gaussian-kernel formulation to wave-dominated hyperbolic PDEs, as well as the $\mathcal{O}(nQ)$ query-event interaction cost at high event counts or fine output grids.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: The assumptions for the formal results are stated in the main text, and the full proofs of the superposition, exact-representability, and Wasserstein-stability results are provided in the appendix.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The paper specifies the benchmark families, sparsity levels, baselines, training setup, evaluation metrics, and the main experimental protocols in Section 4 and Appendix A, including dataset splits, observation protocols, and tensor formats.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: An anonymized supplementary code package is provided with the submission. It includes the main training and evaluation scripts, data-generation scripts for the synthetic benchmarks, and data-preparation scripts for ERA5 and NYC Taxi, together with a README describing the minimal reproduction path and required dependencies.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: Section 4.1 and Appendix A specify the training setup, optimizer, learning rate, weight decay, batch size, early stopping, data splits, observation budgets, and the main benchmark-specific details needed to interpret the results.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: The main quantitative results are reported as averages over three random seeds, with standard deviations reported inline or in the appendix. The paper states that the primary metric is averaged over three random seeds, and the appendix provides the full mean $\pm$ std tables.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: Appendix A.2 reports model complexity, training cost, and inference latency, including hardware information such as a single A100 GPU for training-cost measurements and an RTX 4070 for inference timing.

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?
   Answer: [Yes]
   Justification: This work is a scientific machine learning study on operator learning for sparse event-to-field reconstruction. It does not involve deception, manipulation of human subjects, unsafe data practices, or intentionally harmful applications, and it adheres to the NeurIPS Code of Ethics.

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative

societal impacts of the work performed?
Answer: [No]
Justification: The paper does not include a dedicated broader-impact discussion. The work is primarily methodological and scientific, and the manuscript focuses on technical scope, limitations, and applicability rather than a standalone societal-impact section.

11. **Safeguards**
Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [N/A]
Justification: The paper does not introduce a high-risk generative model, scraped internet dataset, or deployment setting requiring special release safeguards.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [No]
Justification: Existing datasets and baseline methods are properly credited in the paper, but the manuscript does not comprehensively enumerate the licenses and detailed terms of use for every external asset.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [Yes]
Justification: The paper introduces new synthetic benchmark families and an anonymized supplementary code package. The benchmark definitions, sparsity protocols, data-generation procedures, tensor formats, and reproduction steps are documented in Appendix A and in the supplementary README.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [N/A]
Justification: The paper does not involve crowdsourcing or human subjects. All experiments use synthetic or publicly available scientific/event datasets.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [N/A]
Justification: No human subjects are involved in this research. IRB approval is not required.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
Answer: [N/A]
Justification: LLMs are not used as part of the core methodology. They were used only for writing assistance and do not impact the scientific contributions, methodology, or experimental results.