INFINITESIMAL OPERATOR DISCOVERY FOR CONTINUOUS-DEPTH NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-depth neural networks replace a finite stack of layers with a learned vector field, creating a natural mathematical view of deep computation but obscuring which infinitesimal transformations are used after training. This work introduces Infinitesimal Operator Discovery (IOD), a diagnostic framework that analyzes the local variational dynamics of a trained continuous-depth model by projecting its Jacobian action onto constrained structural coordinates: diagonal modulation, topology-aware local mixing, low-rank global mixing, sparse nonlocal mixing, and an explicit residual that quantifies unexplained dynamics. The coordinates are deliberately structural rather than full architectural blocks, so descriptive reconstruction is separated from causal evidence. The audit combines residual fit, null baselines, repeated random controls, decision-boundary interventions, dictionary ablations, solver checks, and rank/radius sensitivity. On a controlled structured benchmark spanning 10 task families and 80 task-seed replications, IOD obtains low residual energy (mean 0.0138), operator profiles that are distinct from untrained, label-shuffled, and random-matched controls, and consistent causal support for low-rank global operators. Local operators are also task-relevant in most settings, while sparse nonlocal and residual coordinates remain negligible in the present benchmark. Boundary-focused interventions have larger effects than high-margin interventions, indicating diagnostic value for decision-sensitive examples. These results support a focused contribution: continuous-depth models can be made more transparent by analyzing their infinitesimal operator structure through a conservative, causal, and reproducible diagnostic coordinate system.

1 INTRODUCTION

Residual networks motivate a continuous view of deep computation, and Neural Ordinary Differential Equations replace a finite stack of transformations with a learned vector field (He et al., 2016; Chen et al., 2018). This formulation has influenced continuous normalizing flows, latent dynamics for irregular time series, neural controlled differential equations, graph diffusion models, implicit-depth models, and continuous-time Transformer variants (Grathwohl et al., 2019; Rubanova et al., 2019; Kidger et al., 2020; Chamberlain et al., 2021b; Bai et al., 2019; Chen et al., 2023; Tong et al., 2025). A continuous-depth predictor can be written as

$$\dot{h}(t) = f_\theta(h(t), t; x), \quad h(0) = E(x), \quad \hat{y} = D(h(T)). \tag{1}$$

The vector field is mathematically explicit, but the infinitesimal transformations learned by training are not.

This creates a structural interpretability problem. Solver steps are numerical approximations to a flow, not semantic layers. Complete neural modules are also too expressive to serve as elementary explanations for the local action of a learned vector field. The more primitive object is the Jacobian

$$J_f(h, t; x) = \frac{\partial f_\theta(h, t; x)}{\partial h}, \tag{2}$$

which governs how infinitesimal perturbations of the hidden representation are amplified, suppressed, and mixed.

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

This paper studies whether the local variational dynamics of trained continuous-depth models can be analyzed through a small set of constrained operator coordinates. IOD decomposes sampled Jacobian actions into diagonal, topology-aware local, low-rank global, sparse nonlocal, and residual terms. These coordinates are structural probes rather than identifiable modules. The explicit residual prevents every learned dynamic from being forced into a named explanation, and causal interventions distinguish descriptive fit from task relevance.

The study uses a controlled benchmark with 10 structured task families and 8 random seeds per task. The results support three claims. First, the constrained dictionary gives low residual energy, with mean residual 0.0138. Second, trained operator profiles are distinct from untrained, label-shuffled, and random-matched controls. Third, low-rank global coordinates have causal support on all 10 tasks, and local coordinates have support on 8 of 10 tasks, with larger effects near decision boundaries than on high-margin examples.

Contributions. The paper introduces a diagnostic framework for continuous-depth models based on local variational operators; defines a constrained structural dictionary with an explicit residual; separates residual reconstruction from causal task support; and reports a reproducible audit covering null baselines, random-control interventions, boundary-sensitive diagnostics, dictionary ablations, solver checks, and rank/radius robustness.

2 Related Work

Continuous-depth models. Neural ODEs parameterize the derivative of a hidden state and compute predictions by solving an initial value problem (Chen et al., 2018). Augmented Neural ODEs analyze topological limits of ordinary Neural ODE flows (Dupont et al., 2019). Continuous-time ideas also underlie continuous normalizing flows (Grathwohl et al., 2019), latent ODEs (Rubanova et al., 2019), Neural CDEs (Kidger et al., 2020), graph diffusion models (Chamberlain et al., 2021b;a), and implicit-depth equilibrium models (Bai et al., 2019). IOD does not propose a new predictor; it analyzes a trained continuous-depth predictor after training.

Dynamical views of architectures and discovery. Neural architectures have long been connected to discretizations of ordinary and partial differential equations (E, 2017; Haber & Ruthotto, 2017; Lu et al., 2018; Ruthotto & Haber, 2020). SINDy and PDE-FIND identify governing equations by sparse regression over candidate libraries (Brunton et al., 2016; Rudy et al., 2017), and PDE-Net learns differential operators from data (Long et al., 2018). IOD borrows the discipline of restricted libraries, residual accounting, and misspecification tests, but the target is different: it builds a diagnostic coordinate system for a learned neural vector field rather than inferring a physical law.

3 Method

3.1 Operator Coordinates

For a sampled trajectory point (x, t), IOD approximates the local Jacobian by

$$J_f \approx J_{\mathrm{diag}} + J_{\mathrm{loc}} + J_{\mathrm{lr}} + J_{\mathrm{sp}} + J_{\mathrm{res}}. \tag{3}$$

The diagonal family contains matrices with zero off-diagonal entries and captures coordinate-wise modulation. The local family

$$\mathcal{C}_{\mathrm{loc}}(\rho) = \{A : A_{ij} = 0 \text{ if } \mathrm{dist}(i, j) > \rho\} \tag{4}$$

uses the fixed order or neighborhood structure of the representation. The low-rank global family

$$\mathcal{C}_{\mathrm{lr}}(r) = \{A : \mathrm{rank}(A) \le r\} \tag{5}$$

captures dense interactions constrained to a low-dimensional subspace. The sparse nonlocal family contains at most s off-band long-range entries. The residual is the remaining operator after fitting the constrained coordinates.

The implementation uses sequential constrained projections of the current residual. Because projection order and dictionary size can affect the fit, the audit includes order-sensitivity checks and

rank/radius sweeps. The residual energy is

$$E_{\mathrm{res}} = \frac{\|J_{\mathrm{res}}\|_F^2}{\|J_f\|_F^2 + \epsilon}. \tag{6}$$

Component energy shares are defined analogously. A low residual is descriptive evidence that the chosen coordinates compress the local operator; it is not by itself evidence of a task mechanism.

This distinction is important mathematically. If two vector fields are uniformly close on a trajectory tube and one is Lipschitz in the hidden state with constant L, then Gronwall's inequality bounds the final-state deviation by $(\epsilon/L)(e^{LT} - 1)$ for $L > 0$, and by ϵT for $L = 0$. Thus local approximation errors can control trajectory errors under standard conditions. The bound does not identify which component matters for the classifier, so IOD adds causal interventions.

3.2 Causal Support

For a component m, the intervention modifies the local Jacobian action along sampled perturbation directions by

$$J_f v \mapsto J_f v - \lambda J_m v, \tag{7}$$

where λ is selected from a fixed intervention grid. Each named intervention is compared with repeated random controls matched to the perturbation scale of $J_m v$. The reported effect is

$$\Delta_m = \Delta L_m - \mathbb{E}[\Delta L_{\mathrm{rand}}], \tag{8}$$

where ΔL_m is the loss increase under the named component intervention. A task supports a component when the seed-level confidence interval for Δ_m exceeds the fixed practical margin $\tau = 0.015$.

Boundary-sensitive diagnostics use correctly classified examples with small positive logit margin,

$$\mathrm{margin}(x) = z_y(x) - \max_{c \neq y} z_c(x), \tag{9}$$

and compare them with high-margin correctly classified examples. Larger intervention effects near the boundary indicate that the coordinate is informative for fragile decisions, where interpretability is most operationally useful.

3.3 Audit Design

The audit is designed to avoid interpreting a good reconstruction as a mechanism. Synthetic recovery checks whether the estimator behaves correctly when the true operator coordinates are known. Misspecification checks remove a true coordinate from the dictionary and test whether the residual exposes the omission. Null baselines compare trained profiles with untrained, label-shuffled, random-matched, coordinate-permuted, and orthogonal-basis controls. Causal interventions test whether removing a named coordinate changes the model more than a repeated matched random perturbation. Boundary-focused interventions test whether the coordinate is informative where the classifier is fragile. Dictionary ablations, solver variants, and rank/radius sweeps test whether the conclusion depends on one projection setting.

These checks deliberately target different failure modes. Low residual without null separation could reflect architecture or initialization. Null separation without intervention effects could reflect a descriptive but behaviorally irrelevant profile. Intervention effects without random matching could reflect perturbation scale. Boundary effects without high-margin comparison could reflect a generic sensitivity to perturbation. IOD treats a coordinate as practically supported only when the descriptive, null, and causal checks agree.

4 Experimental Design

The benchmark contains 10 controlled structured task families and 8 independent seeds per task, for 80 task-seed units. Each task has a fixed representation topology, making local, global, and relational hypotheses meaningful. Content lookup, global sum, long-range XOR, and frequency parity stress global information transfer. Shifted motif and semi-real motif tasks stress local or local-global

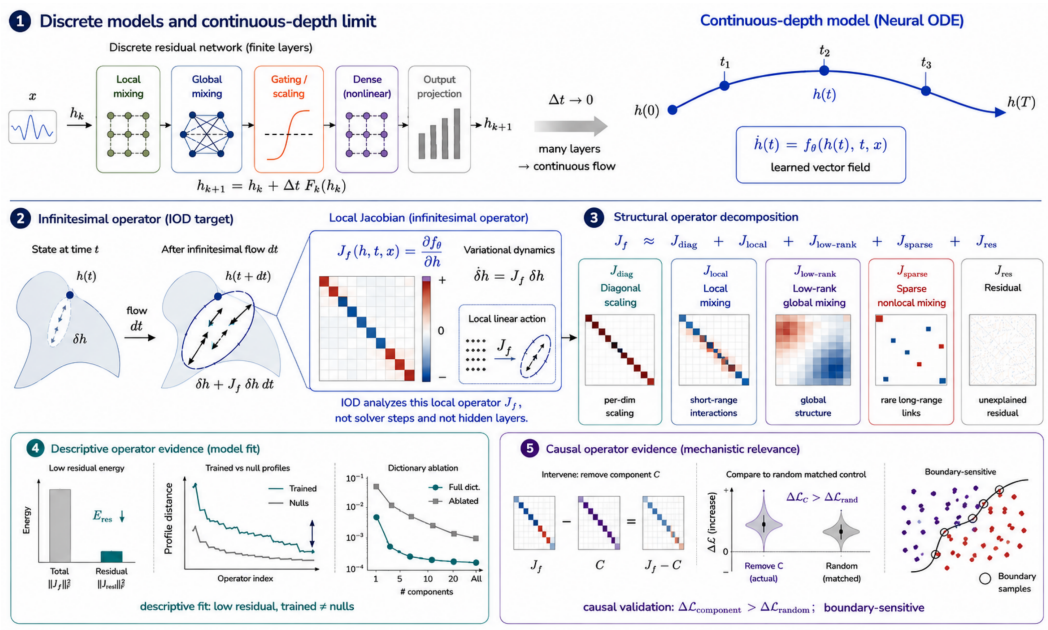

Figure 1: Overview of IOD. The method analyzes the infinitesimal operator governing local perturbation dynamics, decomposes it into constrained structural coordinates, and evaluates those coordinates through descriptive and causal audits.

Table 1: Task-level summary. Accuracy and residual are means across 8 seeds. Causal columns report mean loss increase beyond repeated random matched controls.

Task	Acc. ↑	Residual ↓	Low-rank ↑	Local ↑	Diag ↑
Content lookup	0.994	0.013	1.384	0.180	0.230
Decoy local/global	0.997	0.016	1.887	0.220	0.172
Frequency parity	0.999	0.015	0.889	0.381	0.076
Gated local switch	1.000	0.011	1.069	0.059	0.011
Global sum	0.973	0.010	0.499	0.170	0.258
Long-range XOR	0.998	0.013	0.844	0.013	-0.002
Semi-real motif	0.979	0.015	1.734	1.047	0.624
Shifted motif	0.975	0.023	1.234	0.990	1.033
Sparse edge XOR	0.978	0.008	0.572	0.171	0.124
Sparse pair pointer	0.996	0.013	1.694	0.609	0.449

structure. Sparse fixed-edge XOR and sparse pair pointer test sparse or relational behavior. Gated local switch tests conditional feature selection, and decoy local/global tests conflict between local evidence and global decision rules.

For each task and seed, the audit trains a continuous-depth model, trains a label-shuffled null, evaluates untrained and random-matched controls, computes operator decompositions over sampled states and times, runs component interventions, performs dictionary ablations, and records solver and rank/radius robustness. The supplementary material gives exact protocol details, null definitions, and additional tables. The main paper is self-contained in the sense that all central quantities, support criteria, and conclusions are defined here.

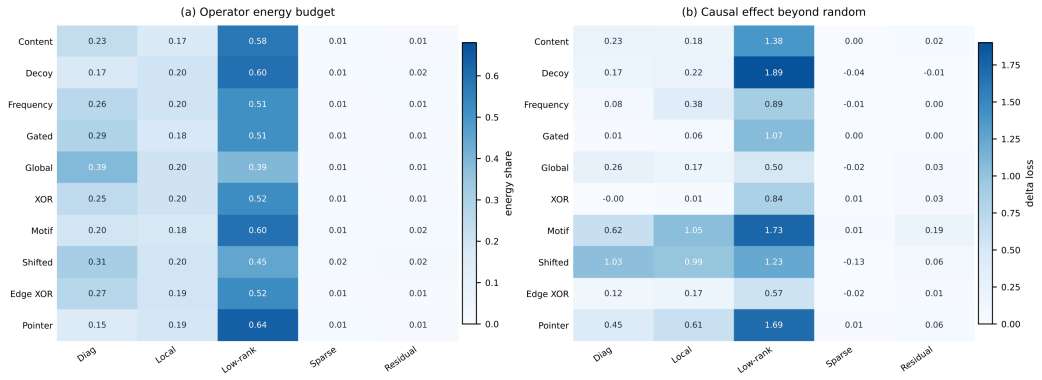

Figure 2: Structural decomposition and causal validation. Left: mean trained operator budgets. Right: mean loss increase beyond repeated random controls at the maximum stable intervention strength. Numerical annotations give the plotted values.

Table 2: Component-level summary. Task support counts tasks whose seed-level confidence interval exceeds $\tau = 0.015$. Boundary gap is the boundary minus high-margin causal effect.

Component	Budget	Effect	Support	Gap
Diag	0.253	0.298	8/10	0.102
Local	0.191	0.384	8/10	0.137
Low-rank	0.533	1.181	10/10	0.372
Sparse	0.010	-0.019	0/10	-0.005
Residual	0.014	0.042	1/10	0.018

5 Results

5.1 Models Solve the Tasks

The trained models achieve mean test accuracy 0.9887, with the lowest task-level mean accuracy 0.9727 on global sum. Label-shuffled models remain near chance, so the operator diagnostics are applied to learned task-aligned functions rather than failed training runs. Table 1 summarizes the main task-level quantities.

5.2 Structural Compression and Causal Effects

Figure 2 reports the two central measurements. The fitted trained dynamics are dominated by low-rank global, diagonal, and local components, while sparse nonlocal and residual energies are small. The mean residual energy is 0.0138, indicating that the constrained coordinates explain nearly all measured local variational energy under the tested projection.

The causal audit gives a stronger conclusion. Low-rank global is supported on all 10 tasks, with the largest effects on decoy local/global, semi-real motif, sparse pair pointer, and content lookup. Local coordinates are supported on 8 of 10 tasks and are especially strong on semi-real and shifted-motif tasks. Diagonal coordinates also pass support on 8 of 10 tasks, but their role is less consistent. Sparse nonlocal and residual coordinates are not supported as main mechanisms in this benchmark.

5.3 Null Baselines and Dictionary Ablations

Trained operator profiles are far from untrained, label-shuffled, and random-matched controls. Mean L2 budget distances are 0.521 ± 0.048 for untrained, 0.381 ± 0.066 for label-shuffled, and 0.531 ± 0.018 for random-matched controls. This makes it unlikely that the trained budgets arise from architecture, initialization, or perturbation scale alone. Coordinate permutations are closer (0.101 ± 0.025), which is expected because diagonal and local coordinates depend on the representation basis.

Table 3: Dictionary ablation summary. Smaller residual increase is better; the full dictionary is the reference.

Variant	Residual increase ↓
No local	0.003 ± 0.010
No sparse	0.012 ± 0.005
Diag+low-rank only	0.013 ± 0.012
No diag	0.024 ± 0.022
Low-rank only	0.035 ± 0.030
Local+low-rank only	0.039 ± 0.025
No low-rank	0.112 ± 0.036
Local only	0.792 ± 0.066

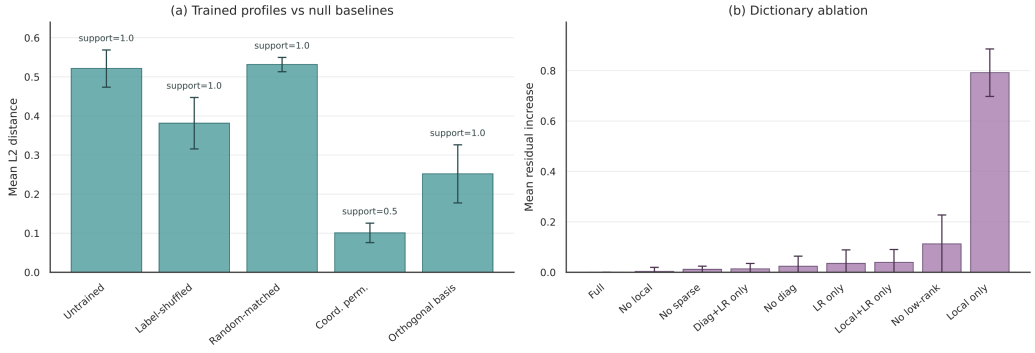

Figure 3: Null and dictionary audits. Left: mean distance between trained operator profiles and null controls. Right: residual increase under dictionary ablations relative to the full dictionary. Error bars show variability across tasks or task-seed samples.

Dictionary ablations separate descriptive reconstruction from causal relevance. Removing low-rank global produces the largest single-removal residual penalty (0.112 ± 0.036), whereas removing local produces a smaller residual penalty (0.003 ± 0.010) despite local having clear causal effects on many tasks. Sparse removal has a small penalty (0.012 ± 0.005). Thus low-rank global is important on both descriptive and causal axes, local is often causally important, and sparse nonlocal is weak on both axes in the tested suite.

5.4 Synthetic Recovery and Energy-Causal Alignment

The synthetic experiments test estimator behavior when the ground-truth operator structure is known. In the matched setting, component recovery is accurate: coefficient correlations are at least 0.9997 for all nonconstant synthetic components, with mean absolute errors near 10^{-3}. When the true low-rank coordinate is removed from the dictionary, the relative Jacobian error becomes large. This misspecification check shows that the estimator can expose a missing coordinate rather than silently absorbing it into unrelated terms.

Figure 4 also shows that energy budgets are useful screening signals but not causal proof. The positive association between energy and intervention effect supports using budgets to select candidate coordinates. The scatter has substantial spread, so the main claims are based on random-control interventions rather than on energy alone.

5.5 Boundary, Solver, and Hyperparameter Robustness

Boundary-focused interventions are more damaging than high-margin interventions for the meaningful coordinates. The largest boundary-high-margin gap is observed for low-rank global, followed by local and diagonal coordinates; sparse and residual interventions do not show the same pattern. This supports the use of IOD as a diagnostic for decision-sensitive examples, not only as an average-profile summary.

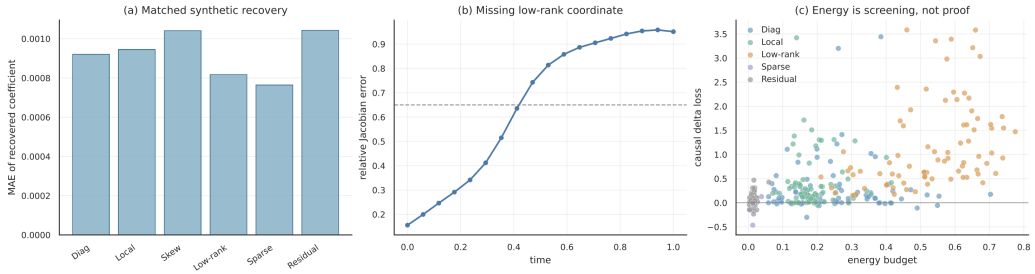

Figure 4: Synthetic and energy-causal audits. Matched synthetic recovery gives small coefficient errors; removing the true low-rank coordinate causes large relative Jacobian error; energy budgets are useful screening signals but are not equivalent to causal evidence.

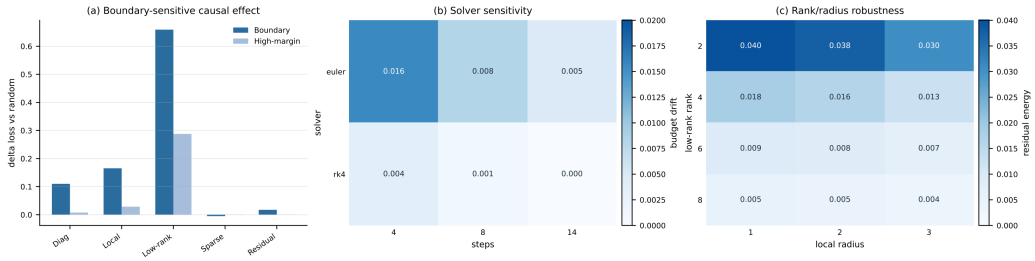

Figure 5: Practical diagnostic and robustness audits. Boundary examples are more sensitive to meaningful operator interventions than high-margin examples; solver sensitivity is small; rank/radius sweeps show a smooth complexity-residual trade-off.

Solver and dictionary checks address two numerical alternatives. RK4 variants have negligible budget drift and prediction changes, while Euler variants produce larger but still modest deviations. Rank/radius sweeps show a smooth complexity-residual trade-off: increasing low-rank rank lowers residual energy, and increasing local radius gives a smaller improvement. The conclusions are therefore not tied to a brittle single solver or dictionary setting.

5.6 Interpretation Safeguards

Several safeguards delimit the claims. The decomposition is basis-sensitive because diagonal and local coordinates are defined in a chosen representation convention; coordinate-permutation and orthogonal-basis controls are therefore reported as caveats rather than dismissed. The low-rank coordinate describes dense interaction through a low-dimensional subspace, but it is not an identified neural module. The local coordinate describes neighborhood-constrained Jacobian action only when the representation topology is fixed. The residual is a diagnostic failure channel, not a catch-all mechanism. These restrictions are why the paper reports operator-coordinate diagnostics instead of module identities.

The support statistic also has a narrow meaning. It counts tasks where the seed-level confidence interval for the random-control-adjusted loss increase exceeds $\tau = 0.015$. It does not estimate universal component importance, and it should not be compared directly with accuracy. Its purpose is to make claims such as "low-rank global is supported on 10/10 tasks" auditable and falsifiable within the benchmark.

6 Discussion, Limitations, and Conclusion

IOD supports a practical workflow for continuous-depth diagnostics. Structural budgets identify candidate operator coordinates, null baselines remove architecture and initialization artifacts, random-control interventions test task relevance, and boundary examples reveal coordinates that

stabilize fragile decisions. The current evidence is strongest for low-rank global coordinates and secondarily for local coordinates. Diagonal coordinates are useful but less central; sparse nonlocal and residual coordinates remain negligible in the present benchmark.

The benchmark is controlled and structured. Broader external validity requires additional domains such as images, graphs, and irregular time series. The decomposition is a diagnostic projection, not an identifiable ground-truth factorization of the learned vector field. Causal interventions approximate component removal through local Jacobian actions rather than through retraining-free ablation of explicitly factorized vector-field components. Alternative coordinates may improve causal alignment or reduce redundancy.

The main conclusion is conservative. Continuous-depth models can be studied as learned operator flows, and their decision-relevant infinitesimal transformations can be measured, stress-tested, and compared. This makes continuous-depth modeling a promising setting for mechanistic diagnostics of learned computation, provided that descriptive reconstruction and causal evidence remain separate.

Reproducibility Statement

The submission includes an anonymized code/data supplement and deterministic audit pipeline. The audit records all 80 task-seed units, model summaries, operator budgets, causal interventions, dictionary ablations, solver sensitivity, rank/radius sweeps, and figure-generation tables. Randomness is controlled at Python, NumPy, and PyTorch levels; the original paper-grade run used CPU execution with six independent workers.

AI Use Statement

AI tools were used for language editing, formatting assistance, and consistency checks. The scientific claims, experimental design, reported results, and final submission decisions remain the responsibility of the authors.

References

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, 2019.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Benjamin P. Chamberlain, James Rowbottom, Davide Eynard, Francesco Di Giovanni, Xiaowen Dong, and Michael M. Bronstein. Beltrami flow and neural diffusion on graphs. In *Advances in Neural Information Processing Systems*, 2021a.

Benjamin P. Chamberlain, James Rowbottom, Maria Gorinova, Stefan Webb, Emanuele Rossi, and Michael M. Bronstein. GRAND: Graph neural diffusion. In *International Conference on Machine Learning*, 2021b.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. In *Advances in Neural Information Processing Systems*, 2023.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In *Advances in Neural Information Processing Systems*, 2019.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5:1–11, 2017.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, 2020.

Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. PDE-Net: Learning pdes from data. In *International Conference on Machine Learning*, 2018.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, 2018.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, 2019.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62:352–364, 2020.

Anh Tong, Thanh Nguyen-Tang, Dongeun Lee, Duc Nguyen, Toan Tran, David Hall, Cheongwoong Kang, and Jaesik Choi. Neural ode transformers: Analyzing internal dynamics and adaptive fine-tuning. In *International Conference on Learning Representations*, 2025.

A Supplementary Material

B Detailed Experimental Protocol

The controlled benchmark contains 10 task families with 8 random seeds per task, for 80 task-seed units. Each task is generated with a fixed representation topology so that local, global, and relational operator hypotheses have well-defined meaning. The suite covers global information transfer, local and local-global motifs, sparse or relational behavior, conditional feature selection, global aggregation, and decoy local/global conflicts.

For each task and seed, the audit trains a continuous-depth model and a label-shuffled null model, then evaluates untrained, random-matched, coordinate-permuted, and orthogonal-basis controls. Each task-seed unit stores model summaries, operator budgets, interventions, ablations, solver checks, rank/radius sweeps, and basis sensitivity. The final paper-grade run completed all 80 units.

C Dictionary and Support Criterion

The dictionary is a restricted structural coordinate system. The diagonal component captures coordinate-wise scaling. The local component captures neighborhood-constrained support with radius ρ. The low-rank global component captures dense interactions whose rank is at most r. The sparse nonlocal component captures a limited number of off-band long-range entries. The residual is the part of the sampled Jacobian left unexplained by these projections.

For causal support, a component m is tested against repeated random matched controls. The reported effect is

$$\Delta_m = \Delta L_m - \mathbb{E}[\Delta L_{\mathrm{rand}}].$$

A task supports the component when the seed-level confidence interval exceeds the practical margin $\tau = 0.015$. This is an audit rule rather than a generic performance metric. The raw means and intervals are retained in the result tables.

Table 4: Main compact result table preserved from the paper-grade result bundle. Accuracy and residual are means across seeds. The top diagnostic component is selected by audited budget and causal-support metrics.

Task	Acc.	Residual	Top causal pass	Expected primary structure
Content lookup	0.9943	0.0131	yes	low-rank global
Shifted motif	0.9748	0.0234	yes	local
Sparse edge XOR	0.9775	0.0079	yes	sparse nonlocal
Sparse pair pointer	0.9956	0.0128	yes	sparse or low-rank global
Gated local switch	0.9995	0.0111	yes	diagonal or local
Decoy local/global	0.9969	0.0159	yes	low-rank global
Semi-real motif	0.9790	0.0153	yes	local and low-rank global
Global sum	0.9727	0.0102	yes	low-rank global
Long-range XOR	0.9976	0.0132	yes	low-rank global
Frequency parity	0.9993	0.0147	yes	low-rank global

Table 5: Component-level summary with standard deviations. Budget share is descriptive; task support is causal.

Component	Budget share	Causal effect	Support
Diag	0.253 ± 0.137	0.298 ± 0.324	8/10
Local	0.191 ± 0.065	0.384 ± 0.374	8/10
Low-rank	0.533 ± 0.138	1.181 ± 0.490	10/10
Sparse	0.010 ± 0.005	-0.019 ± 0.041	0/10
Residual	0.014 ± 0.007	0.042 ± 0.058	1/10

D Null Baselines

The null suite serves distinct purposes. Untrained models test architecture and initialization bias. Label-shuffled models test whether input distribution and optimization without task-aligned labels can create similar profiles. Random-matched controls test whether causal effects arise simply from subtracting a perturbation of comparable scale. Coordinate permutation and orthogonal-basis controls test representation sensitivity. The strongest supported claim is distinctness from untrained, label-shuffled, and random-matched controls; basis-sensitive controls are retained as caveats.

E Dictionary Ablation Details

Table 7 reports residual increases for dictionary variants. The full dictionary is the reference. Removing low-rank global produces the largest single-removal penalty, whereas removing sparse nonlocal has a small effect. This supports the role of low-rank global as a central descriptive coordinate and indicates that the current sparse coordinate is not a main explanatory component for the tested tasks.

F Synthetic Recovery and Misspecification

The synthetic experiments test estimator behavior when ground-truth structure is known. In the matched setting, IOD recovers component coefficients with small error. When the true low-rank coordinate is removed from the dictionary, the relative Jacobian error becomes large. This shows that the method can detect a missing structural coordinate rather than silently absorbing it into unrelated components.

G Boundary, Solver, and Rank/Radius Audits

Boundary interventions compare correctly classified examples near the decision boundary with high-margin correctly classified examples. Low-rank global has the largest boundary-high-margin gap,

Table 6: Null-baseline distances. Larger distance indicates stronger separation from the trained operator profile. Values are mean ± standard deviation over tasks.

Control	L2 distance ↑
Untrained	0.521 ± 0.048
Label-shuffled	0.381 ± 0.066
Random matched	0.531 ± 0.018
Coordinate permutation	0.101 ± 0.025
Orthogonal basis	0.252 ± 0.074

Table 7: Dictionary ablation summary. Smaller residual increase is better.

Variant	Residual increase ↓
Full	0.000 ± 0.000
No local	0.003 ± 0.010
No sparse	0.012 ± 0.005
Diag+low-rank only	0.013 ± 0.012
No diag	0.024 ± 0.022
Low-rank only	0.035 ± 0.030
Local+low-rank only	0.039 ± 0.025
No low-rank	0.112 ± 0.036
Local only	0.792 ± 0.066

followed by local and diagonal coordinates. Sparse and residual interventions do not show the same pattern.

Solver sensitivity checks compare RK4 and Euler variants. RK4 variants have negligible prediction changes and budget drift, while Euler variants produce larger but modest deviations. Rank/radius sweeps show a smooth complexity-residual trade-off: increasing low-rank rank lowers residual energy, and increasing local radius gives a smaller improvement. These results make the main conclusions less likely to be artifacts of a single solver or dictionary hyperparameter.

H Overview Figure

Figure 8 shows the conceptual flow of the diagnostic. It is included here to avoid spending main-paper space on a schematic while preserving the visual map of the method.

I Result Provenance

The final paper-grade run completed all 80 task-seed units used in the paper. The included code/data supplement contains the compact result tables and verification script used to check the reported values.

Table 8: Matched synthetic recovery. Correlations and errors are computed between known synthetic coefficients and recovered estimates.

Component	Corr.	MAE
Diag	0.9999	0.0009
Local diffusion	1.0000	0.0009
Skew transport	0.9999	0.0010
Low-rank global	1.0000	0.0008
Sparse nonlocal	0.9997	0.0008
Dense residual	–	0.0010

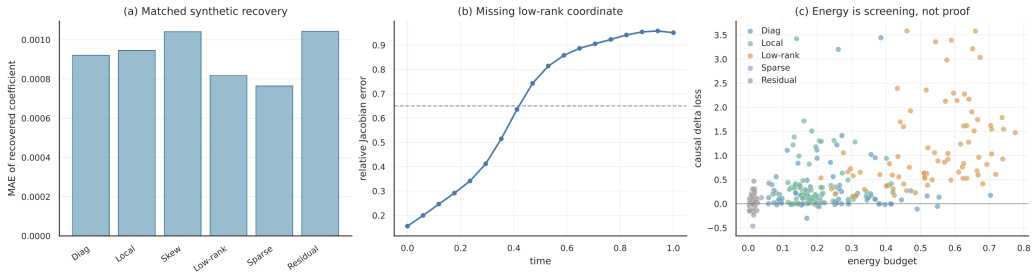

Figure 6: Synthetic and energy-causal audits. Matched synthetic recovery gives small coefficient errors; removing the true low-rank coordinate causes large relative Jacobian error; energy budgets are useful screening signals but are not equivalent to causal evidence.

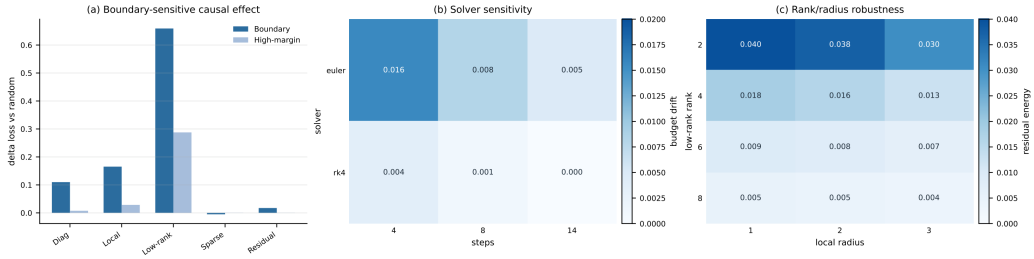

Figure 7: Practical diagnostic and robustness audits. Boundary examples are more sensitive to meaningful operator interventions than high-margin examples; solver sensitivity is small; rank/radius sweeps show a smooth complexity-residual trade-off.

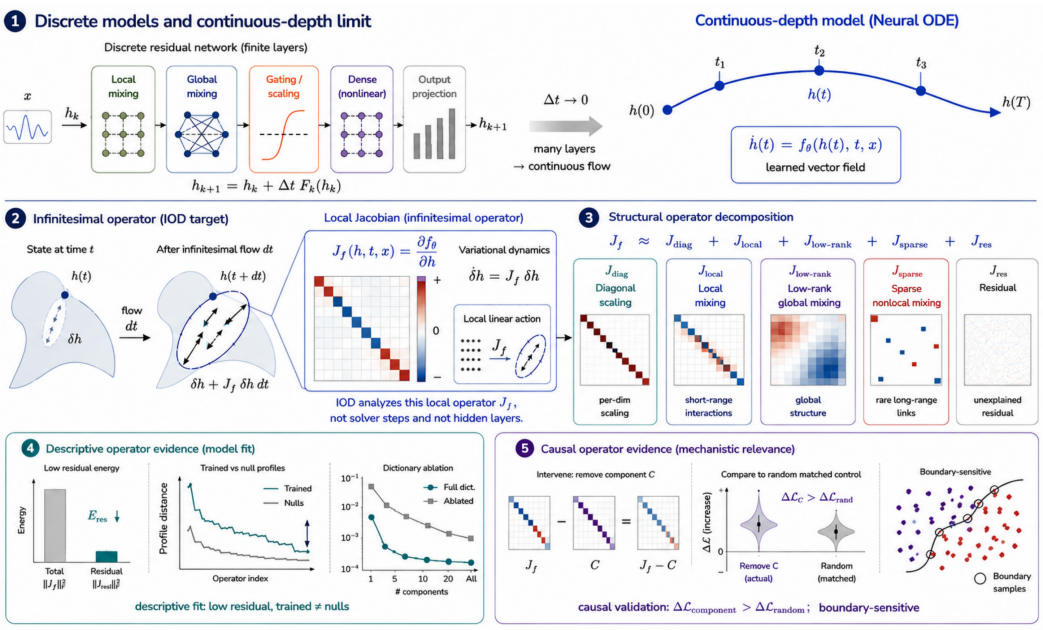

Figure 8: Overview of IOD. The method analyzes the infinitesimal operator governing local perturbation dynamics, decomposes it into constrained structural coordinates, and evaluates those coordinates through descriptive and causal audits.